Танатар Н. В.

Информационно-аналитическое обслуживание высших органов государственной власти с использованием автоматизированных систем ключевых фрагментов публикаций

Отличительная особенность информационно-аналитического обеспечения высшего аппарата государственной власти состоит прежде всего в обслуживании немногочисленных потребителей со специальным кругом задач, требующих их четкой постановки и оперативного решения, а также тесного контакта с заказчиками информации. К числу наиболее сложных проблем в этой связи следует отнести определение специфических информационных интересов, которых довольно часто до конца не осознают и сами потребители аналитических материалов. Поэтому уточнение информационных потребностей в большинстве случаев происходит уже в процессе работы. Сама подготовка информационно-аналитических материалов состоит из ряда последовательных процедур, начиная с составления списка первоисточников для просмотра, разработки методик отбора и классификации материалов, их автоматической обработки и заканчивая анализом занесенной в базы данных (БД) информации.

Организацию автоматизированных систем условно можно разделить на три этапа: информационно-технологический, информационный и аналитический. В задачи первого этапа входит прежде всего создание оптимальной программной среды для накопления, обработки и сохранения информации в системе, а также обеспечение необходимой электронно-вычислительной техникой, компьютерными сетями и средствами их связи.

Не менее ответственным является этап формирования информационных потоков. Он предполагает продуманный и тщательный отбор информации, исключающий всякого рода информационный шум. Более того, вводимая в БД информация, сохраняя ключевые позиции исходного текста, должна быть максимально «отжата» и наилучшим образом структурирована для последующей эффективной работы с ней. Поэтому в данном случае предпочтение оказывается собственным БД, создаваемым для решения конкретных проблем. Избыточная информация, не говоря уже о нерелевантной, лишь усложняет работу с информационными массивами.

На этапе анализа БД осуществляется формирование разнообразных выходных файлов в соответствии с индивидуальными запросами заказчиков информации.

Отдельно следует остановиться на процедурах постоянного наблюдения за теми или иными характеристиками социально-политических процессов или явлений. При выполнении такого рода работ возникают трудности как с определением изучаемых параметров и их пороговых значений, так и с установлением частоты оценки этих параметров. Процесс мониторинга усложняет необходимость разнопланового воспроизведения результатов наблюдений в текстовых файлах и в статистической форме. Для решения поставленных задач требуется в полном объеме предусмотреть набор параметров и их количественные характеристики, выявить причинно-следственные связи между ними и включить методологические принципы исследования в описание функционирования технических систем.

Окончательный анализ информации и ее экспертные оценки, сопряженные с составлением прогностических записок и выработкой конкретных управленческих решений, производятся в соответствующих структурных подразделениях высших органов государственной власти.

С учетом перечисленных особенностей такого рода деятельности и требований конкретных заказчиков в Лаборатории прогнозно-аналитической библиометрии (с 1998 г. Отдел организации и использования документального фонда ФПУ) Национальной библиотеки Украины им. В. И. Вернадского в 1993 г. был основан новый класс автоматизированных систем, базирующихся на ключевых цитатах публикаций. Впервые такая система была создана для анализа процесса обсуждения проекта новой Конституции Украины в прессе [1]. Обсуждение в газетах проекта в целом и его отдельных статей, конкретные предложения по их изменению, дополнению, исключению и др. требовали точной передачи текста оригинала. Принципы формирования БД стандартных информационных систем, за исключением полнотекстовых, не соответствовали этому условию. Вместе с тем полнотекстовые БД оказались чрезвычайно громоздкими для проведения таких исследований. Поэтому в нашем случае для наполнения БД был использован не весь документ, а его отдельные фрагменты, которые точно и сжато передавали содержание каждого конкретного предложения. Для индексирования отобранных цитат был разработан специальный двухфасетный классификатор. Были созданы программные средства конвертирования входной информации в

структурированные согласно статьям проекта новой Конституции выходные текстовые файлы в виде постатейно сгруппированных фрагментов публикаций с их полным библиографическим описанием. Отдельно формировались статистические таблицы, которые также передавали структуру вынесенного на всенародное обсуждение проекта.

Авторы проекта новой Конституции проявили большой интерес к полученным аналитическим материалам, и в дальнейшем основные принципы построения системы были использованы при создании других систем социально-политической тематики. Наибольшие из них предназначались для отслеживания имиджа политической элиты и политических партий Украины в прессе и пресс-мониторинга выборов в Верховный Совет Украины в 1994, 1998 гг. и Президента Украины в 1994 г. с соответствующими текстовыми и статистическими выкладками [2].

В сформированных на базе ключевых фрагментов публикаций системах, с одной стороны, были использованы принципы построения хорошо известных информационных систем, с другой – предложены новые оригинальные подходы. Для наполнения БД в нашем случае, как и в информационных системах документального типа, используется документ. С информационными системами фактографического типа их сближает то, что входной документ при поиске информации теряет свойства единого и неделимого целого. Размещением электронных записей в виде иерархического дерева или сети информационных модулей они напоминают гипертекст. В гипертексте это могут быть страницы первоисточника, параграф или раздел, а в нашем случае - это ключевые цитаты публикации, несущие определенную смысловую нагрузку. Особенность данных систем состоит в том, что занесенный в БД документ функционирует не как единое целое, а как совокупность отдельных, независимых одна от другой цитат. При формировании таких систем понятие «ключевая цитата» относится не только и не столько к тексту, из которого она взята, сколько к теме, которую отображает информационная система. Поэтому неудивительно, что «ключевой» по отношению к исследуемой теме фрагмент в контексте документа, из которого выделяется, может играть второстепенную роль или же выполнять вспомогательную функцию. Например, уточнять или разъяснять отдельные положения основного текста.

В последние годы повысились требования заказчиков и усложнились задачи к такому классу систем. Они коснулись как процедур обработки исходных текстов, так и способов конвертирования

информационных массивов. Возникла необходимость использовать монотематические по наполнению системы для многоаспектного поиска информации при подготовке выходных материалов. Сортировка электронных документов как по каждому из заданных признаков, так и по заданным их комбинациям требует специального программного инструментария. Кроме того, понадобилось расширение спектра количественных показателей и библиометрических параметров исследуемой проблемы. С учетом новых задач были пересмотрены и усовершенствованы некоторые принципы построения таких систем, модифицирован ряд технологических процессов формирования и функционирования систем. Прежде всего, существенно изменен этап подготовки материалов для наполнения БД. Ключевую роль в модернизации системы сыграло использование новых подходов при создании классификаторов.

В каждом конкретном случае процессу построения классификатора предшествовал структурный анализ исследуемой проблемы, суть которого сводилась к выявлению ее составных элементов и связей между ними. После такой предварительной процедуры строился граф проблемы, воссоздающий в схематическом виде ее модель от общих понятий (параметров) типа: объект исследования; субъект исследования; исследуемая – тема до элементов проблемы в их конкретных проявлениях. В рамках данной модели каждому значению выделенных элементов проблемы присваивается определенный код (индекс). Это и является первым вариантом рабочего классификатора проблемы, организованного по фасетно-иерархическому типу, в котором совокупность значений в пределах каждого элемента представлена соответствующими фасетными классами [3].

На базе таких классификаторов осуществлялось формальное описание документов с использованием фасетных формул. Для индексирования входных файлов в виде ключевых цитат публикаций были предложены многофасетные формулы с теоретически неограниченным количеством фасет и значений в каждом из них. Такой подход позволяет формировать для каждого аспекта (параметра) проблемы свой фасет с соответствующим ему набором элементов и создавать информационные системы для сложных и многоплановых тем. Кроме того, предусмотрены фасеты для фиксирования «оттенков» передачи проблемы в публикации и, в зависимости от запросов пользователей, других индикаторов оценки проблемы. Представляя конкретный параметр исследуемой проблемы,

каждый из введенных фасетов занимает строго фиксированное место в фасетной формуле. Таким образом, фасетную формулу составляют взятые в строго фиксированной последовательности индексы фасетных классов соответствующих элементов проблемы и избранных индикаторов ее оценки с разделяющими их синтаксическими знаками. Индексы каждого предыдущего фасета отделяются от индексов последующего с помощью точки с запятой (;). В процедуре индексирования входных материалов также учтены случаи, когда один элемент может одновременно иметь несколько значений. Все выявленные в тексте значения такого элемента фиксируются с помощью соответствующих им индексов в отведенном для раскрываемого ими параметра месте фасетной формулы (фасете), отделяясь друг от друга запятой (,). Кроме того, технологически предусмотрена возможность структурного и содержательного усовершенствования классификатора по мере развития проблемы. В него на уровне структуры добавляются фасеты, соответствующие новым параметрам, а на уровне содержания в классификатор заносятся значения элементов этих параметров и каждому из них присваивается индивидуальный индекс. В результате такой процедуры фасетная формула пополняется новыми фасетами, а сам классификатор – их фасетными классами. Что же касается конкретных значений ранее зафиксированных элементов, то некоторые из них со временем «отмирают» или же модифицируются, параллельно появляются новые значения существующих элементов. Все эти изменения также вносятся в классификатор проблемы на уровне индексов и соответствующих им значений. В настоящее время для индексирования заносимых в БД цитат используется пятифасетная формула, в которой первый фасет соответствует параметру «объект исследования», второй – параметру «субъект исследования», третий – параметру «тема исследования», четвертый – параметру «модальность публикации», пятый – параметру «источник информации». В зависимости от задач исследования одни и те же элементы проблемы и соответствующие им классы значений могут использоваться для раскрытия разных параметров. Например, элементы «политические лидеры», «политические партии» через их конкретные значения в одних системах («Имидж политической элиты Украины в прессе», «Имидж политических партий Украины в прессе») отражают объект исследования, в других («Выборы в ВС Украины») – выступают в качестве субъектов исследования и т. д.

Новые классификационные подходы внесли существенные коррективы в процесс обработки первоисточников [4]. Процедура контент-

анализа публикаций с использованием новых классификационных схем направлена на выделение из текста фрагментов, соответствующих наименьшему целостному модулю информации в пределах исследуемой проблемы. В рамках такого модуля вычленяются элементы проблемы, адекватные конкретным значениям классификатора и между ними устанавливаются связи для последующей формальной передачи содержания фрагмента публикации посредством фасетной формулы. При этом семантические элементы текста и классификатора должны либо, совпадать, либо нести идентичную смысловую нагрузку. Каждой из выделенных цитат после ее тщательного анализа присваивается совокупность определенных индексов (кодов), соответствующих конкретным значениям классификатора и располагающихся в строго фиксированной последовательности в фасетной формуле. По мере необходимости фиксируются индикаторы оценки проблемы в специально предусмотренных для них фасетах. По существу на основе разработанного классификатора с помощью фасетной формулы производится формальное описание выделенного фрагмента текста, а совокупность фасетных формул всех выделенных фрагментов обеспечивает формальное описание документа в целом в контексте исследуемой проблемы. Не относящаяся к данной проблеме информация не выделяется из текста и не заносится в БД. Введенный в информационную систему документ представлен полным библиографическим описанием первоисточника и совокупностью ключевых фрагментов текста, каждый из которых заиндексирован согласно его содержанию. Предложенная процедура обработки публикаций является своеобразным «информационным ситом», пропускающим лишь релевантную теме информацию в виде экстрагированных из текста фрагментов. Среди преимуществ данной технологии, во-первых, следует отметить сведение информационного шума к минимуму. Во-вторых, сформированные БД, сохраняя в виде ключевых цитат текст оригинала, являются достаточно компактными и удобными в работе. Среди недостатков технологии – большие затраты ручного труда, как при обработке первоисточников, 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 настоящее время не получили распространения. Более того, чрезвычайно низкое качество печати таких изданий делает невозможным использование процедуры электронного узнавания текста со сканера. По мере решения издательских проблем

технологический процесс формирования БД будет развиваться в направлении его поэтапной автоматизации.

Усовершенствованная система обработки БД представляет собой совокупность информационных файлов, аккумулирующих сведения о фасетах, значениях фасетных индексов и порядке сортировки информационных модулей, а также файл конфигуратора. Последнему отведена ведущая роль в технологическом процессе, поскольку он включает названия информационных файлов, используемых для индексирования и сортировки информационных модулей, названия БД, сохраняющих эти модули для дальнейшей обработки, названия выходных файлов и т. д. Специально написанные программные средства разбивают входной документ (информационный объект) на отдельные независимые фрагменты (информационные модули), автоматически снабжая каждый из них библиографическими данными, внесенными при описании документа в целом. В результате таких технологических преобразований формируется массив (сеть) ключевых фрагментов публикаций, представляющий собой информационный оттиск исследуемой проблемы. Кроме того, каждый информационный модуль такой сети при запуске соответствующих программ разбивается на три независимые составляющие, каждая из которых может функционировать самостоятельно. Одна из составляющих представляет собой цитату документа (содержательная информация), другая – полное библиографическое описание документа (библиометрический параметр) и третья – фасетную формулу (структурная характеристика выделенного фрагмента текста). Такая технологическая процедура преобразовывает ранее сформированный массив информации в полную публикационную матрицу, поддающуюся анализу по каждой из выделенных составляющих и по их произвольно заданной совокупности.

Предложенный способ представления и расчленения входной информации позволяет ею всячески манипулировать в процессе формирования выходных текстовых файлов и при получении количественных показателей.

Технологически предусмотрены всевозможные комбинации цитат в пределах параметров классификатора. В качестве доминантного может избираться любой из зафиксированных в фасетной формуле элементов или его отдельные значения. Проблему можно представлять как комплексно, то есть с учетом всех ее признаков, так и частично – в случае, когда пользователя интересуют только ее отдельные аспекты. В последнем случае полная публикационная матрица преобразуется в ее производные –

свернутые по одному, двум и т. д. или по всем, кроме одного, элементам. Эти элементы задаются потребителями информации и являются обязательными компонентами классификатора проблемы.

Таким образом, комплексно на всех этапах построения информационных систем, как в технологиях обработки первоисточников и формирования БД, так и в программных средствах конвертирования информационных объектов, было предусмотрено многовариантное и широкоаспектное использование тематически детерминированной входной информации. Анализ и синтез включенных в систему фрагментов по продуманным и заданным технологическим схемам позволяет получать нетривиальные информационные продукты. Выходные файлы могут быть представлены как в виде структурированных цитат публикаций с соблюдением определенной иерархии, так и статистических таблиц аналитико-синтетической обработки информационных модулей. Конечный продукт может иметь и комбинированный вид – структурированный текстовый материал дополняется количественными показателями. Реализован и нестандартный вариант выходного файла – структурированный текст в виде таблицы. По вертикали такой таблицы размещаются фрагменты публикаций согласно заданной потребителем иерархической схеме. Ступенями такой иерархии могут быть: объекты исследования, субъекты, темы исследования и т. д. В горизонтальной плоскости таблицы фиксируются модальность цитат, региональная принадлежность, источники информации и др. характеристики. Для удобства потребителей в пределах вертикальной иерархии по каждому конкретному значению элементов проблемы информационные модули группируются согласно признакам, зафиксированным в горизонтальном срезе таблицы. Очередность и значимость таких признаков задаются потребителями информации.

В данных системах предусмотрено, что статистические таблицы и графические изображения формируются на основе фасетных индексов цитат и библиографических данных. В качестве дополнительного критерия оценки материалов социально-политического характера был введен показатель модальности информации – отношение автора к описанным в публикации событиям, явлениям и т. п. Информационные технологии систем позволяют вводить новые элементы, необходимые в отдельных случаях конкретным потребителям для формальной передачи и анализа проблемы. 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, а в соответствии с его изменениями совершенствуется система и появляются новые возможности для анализа проблемы. Начиная с 1998 г., новый технологический режим используется для формирования и функционирования всех информационных систем, как вновь создаваемых, так и продолжающихся. Наибольший социальный интерес представил пресс-мониторинг последней предвыборной кампании – выборов Президента Украины в 1999 г., проведенный с использованием разнообразных возможностей системы [5, 6].

Как видим, системы, состоящие из ключевых фрагментов публикаций, обладают рядом преимуществ. Они позволяют охватывать большие массивы документов, поскольку включают лишь «экстракты» текстов первоисточников. Кроме того, предложенная технология контент-анализа текста предусматривает отбор фрагментов из непрофильных для данной проблемы публикаций, в которых они выполняют второстепенную роль ремарок, сносок и пр. Однако, выявление такой информации возможно лишь при очень тщательном анализе первоисточника. Для достижения максимальных результатов создателям входной в БД информации необходимо не только хорошо ориентироваться в обрабатываемом материале, но и знать аудиторию потребителей и способы использования ими формируемой информации. Программное обеспечение систем способствует эффективной обработке БД. Оно позволяет выбирать из всего массива занесенных документов лишь ту информацию и в той последовательности, в которой она необходима потребителю, естественно, исходя из перечня возможных вариантов, и оформлять ее в виде структурно организованных текстовых блоков и таблиц чисел.

Таким образом, создание таких систем направлено на максимально полное и концентрированное отображение любой темы или проблемы в информационном массиве с последующим его полным или избирательным анализом и синтезом. Сам процесс выделения интересующих потребителя блоков, как чего-то целостного, осуществляется на основании исключительно структурных критериев. В процессе генерирования сети системы используются механизмы моделирования проблемы, позволяющие выявлять в некотором массиве текстов содержательно близкие фрагменты. Существенно то, что такая сеть может дополняться новыми единицами информации, поскольку возможно практически неограниченное расширение классификатора с включением

в него как новых элементов, так и новых значений уже существующих элементов. Система принципиально открыта для новых информационных модулей. Присоединение их к уже существующим не может нарушить или испортить ее структуру. Добавление новых записей лишь расширяет сеть системы как единую и непрерывную. Поскольку каждая новая публикация может иметь не только общие с предыдущими материалами признаки, но и вносить новые элементы в содержание темы, то в процессе пополнения системы фрагментами таких публикаций возникает особый эффект, который можно условно назвать прогрессией проблемы или «дрейфом» ее содержания.

Оценивая в целом преимущества системы, следует выделить такие ее особенности, как избежание повторного библиографического описания при занесении фрагментов документа в БД, высокая степень структуризации информации, возможность формирования в рамках одной системы как локальной, так и глобальной сети, индивидуализация структуры и формы выходного информационного продукта.

Литература

1. Заєць П., Танатар Н. Зі шпальт газет – до проекту: За матеріалами обговорення проекту нової Конституції України у пресі // Голос України. – 1993. – 4 черв., № 103.
2. Танатар Н., Федорчук А. Усе – про політичну еліту // Вісн. НАН України. – 1997. – № 1– 2. – С. 91 – 92.
3. Танатар Н. Информационно-аналитические системы оценки состояния науки// Наука та науковознавство. – 1996. – № 3– 4. – С.124 – 132.
4. Сорока М., Танатар Н. Використання методу контент-аналізу при створенні автоматизованих інформаційних систем// Бібліотека. Наука. Культура. Інформація. Наук. праці НБУВ. – 1998. – Вип.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.-теоретичн. конф. 11-13 травня 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.