УДК 681.82:519.81

О.В. Канищева, Сайед Мохаммад Таухид Сиддики, Н.В. Шаронова

ИСПОЛЬЗОВАНИЕ МЕТОДОВ DATA MINING И TEXT MINING ДЛЯ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

1. Введение, актуальность работы

В начале XXI столетия информация становится одним из наиболее значимых стратегических ресурсов, оказывающих решающее воздействие на развитие общества. Электронная информация играет все большую роль во всех сферах жизни современного общества. В информационных хранилищах, распределенных по всему миру, собрано огромное количество текстовых данных. Накопление информационных ресурсов Интернет многократно усугубило проблему информационной перегрузки. Сырые неструктурированные данные составляют не менее 90% информации, с которой имеют дело пользователи. Найти в таких данных нечто ценное можно лишь посредством специализированных технологий.

Перспективным в настоящее время является использование моделей и методов новой информационной технологии, базирующейся на результатах, полученных при решении проблем искусственного интеллекта. Несмотря на достигнутые в последние десятилетия результаты в области моделирования интеллекта, вопросы семантической обработки текстовой информации все еще не достаточно изучены. Доступность методов записи и хранения данных привели к бурному росту объемов хранимых данных. Объемы данных настолько внушительны, что человеку просто не по силам проанализировать их. Хотя необходимость проведения такого анализа вполне очевидна, ведь в этих «сырых данных» заключены знания, которые могут быть использованы при принятии решений.

Для обработки этих «сырых данных» в последнее время появились решения, известные под общим названием «Data mining». Эти решения позволяют обнаруживать при помощи математических методов (моделирование, прогнозирование, кластеризация, классификация и т.д.) ранее неизвестные нетривиальные практически полезные и доступные для интерпретаций новые знания. С появлением Интернета, систем электронного документооборота и проблем с выработкой стандартов для документации все большее количество информации стало храниться в текстовом виде. Это привело к появлению решений для обработки текстовой информации – Text mining. Такие системы осуществляют при помощи лингвистических методов следующие действия: тематический поиск в текстах, классификация документов, ответ на запросы, тематическое индексирование документов, поиск по ключевым словам, выявление объектов и связей между ними, реферирование и т.д. Технологии Text Mining как раз и предназначены для проведения смыслового анализа, обеспечения навигации и поиска в неструктурированных текстах. Применяя построенные на их основе системы, пользователи смогут получить новую ценную информацию – знания.

Технология глубинного анализа текста – Text Mining – это тот самый инструментарий, который позволяет анализировать большие объемы информации в поисках тенденций, шаблонов и взаимосвязей, способных помочь в принятии стратегических решений. Кроме того, Text Mining – это новый вид поиска, который в отличие традиционных подходов не только находит списки документов, формально релевантных запросам, но и обеспечивает достаточно высокий уровень анализа с целью принятия эффективного решения.

Процесс осмысленного поиска является далеко не тривиальным, часто в коллекции документов присутствует только намек на необходимую информацию. Необходимы мощные интеллектуальные возможности, чтобы найти то, что требуется. В названии технологии слово «mining» (добыча руды) выступает как метафора нахождения глубоко «зарытой» информации [1].

Следует заметить, что технологии глубинного анализа текста исторически предшествовала технология добычи данных (Data Mining), методология и подходы которой широко используются и в методах Text Mining. Технологии Text Mining, кроме того, присуща объективность — отсутствует субъективизм, свойственный человеку-аналитику. Важный компонент технологии Text Mining связан с извлечением из текста его характерных элементов или свойств, которые могут использоваться в качестве метаданных документа, ключевых слов, аннотаций, для оформления документации. Другая важная задача состоит в отнесении документа к некоторым категориям из заданной схемы их систематизации. Text Mining также обеспечивает новый уровень семантического поиска документов.

2. Постановка задачи

В данной статье рассматриваются современные подходы к методам обработки информационных потоков с целью извлечения знаний по технологии Data Mining в отрасли Text Mining. Ставится задача исследования методологий в сфере применения Text

Mining, выделяется специфика современных требований к эффективной интеллектуальной переработке данных. Анализируются проблемы, которые неудовлетворительно решаются существующими методами предварительной обработки и доступа к большим объемам информации, рассматриваются особенности различных информационно-поисковых систем и средств извлечения знаний. Большое внимание уделено новому направлению обработки текстовой информации – «глубинному анализу текстов» (Text Mining), объединяющему в себе технологические и методологические подходы компьютерной лингвистики и искусственного интеллекта.

3. Информационный поиск в автоматизированных информационных библиотечных системах

Современное общество переживает этап информационной революции. Научные библиотеки сосредотачивают в себе сконцентрированные и систематизированные знания и обеспечивают доступ к ним.

Наличие знаний, их накопление при отсутствии современной системы классификации, обработки, поиска и предоставления информационных услуг приводят к информационному кризису.

Можно выделить следующие направления создания систем доступа к информации:

– разработка дружественных интерфейсов интерактивного и почтового взаимодействия, ориентированных на конечных пользователей;

– представление стратегии навигации и поиска в форме, адекватно отражающей предметно-смысловой образ знаний;

– создание систем навигации и поиска, адаптируемых под конечного пользователя с применением систем классификации и систематизации;

– представление результатов поиска за «комфортное» время и в форме, обеспечивающей наглядную визуализацию информации;

– организация информационно-аналитической обработки знаний.

В настоящее время используются две основные стратегии информационного поиска:

- поиск по предварительно классифицированной информации;
- контекстно-свободный поиск.

Первая стратегия является базовой и ориентирована на применение различных систем классификации и систематизации, обеспечивающих направленно-детерминированный поиск.

Вторая предлагает свободный поиск на основе информационно-поискового образа. Создание современных аппаратно-программных систем поиска, а также бурное развитие технологии информационных хранилищ предполагает ее широкое применение в дальнейшем [2].

Решение проблемы доступа к информации, обеспечит формирование интеллектуального, демократического и свободного в информационном отношении общества.

Назначение библиотеки как социального института, обеспечивающего обществу возможность эффективного использования накопленного им документального потенциала, реализуется в предоставлении читателям разнообразного ассортимента информационных услуг в оптимальные сроки и желаемого качества.

Отличительной чертой современного общественного развития является стремительное возрастание объемов и значимости информации, ее усиливающееся влияние на все области деятельности. Информация, традиционно воспринимается как средство общения, на глазах нынешнего поколения людей все больше проявляет себя как универсальное средство организации всех форм общественной жизни, в решающей степени определяющее эффективность работы технических систем, отраслей, предприятий и народного хозяйства в целом. Понимаемая как процесс создания информационных систем, основанных на новейших средствах вычислительной техники и технологиях обработки данных, и применяемая для обеспечения доступа к накопленным в мире знаниям, информатизация охватила все стороны современного общества.

Автоматизация поиска открыла новые возможности для библиографической практической деятельности. Вместе с тем явно ошибочно представление о ней как о средстве легкого преодоления всех сложностей поиска информации. Проблему автоматизации библиотечного поиска нельзя считать решенной в наши дни. Всегда есть и будут вопросы использования технических средств, требующие теоретической разработки.

Таким образом, для современного этапа характерно повышенное внимание к вопросам справочно-библиографического обслуживания, связанным с автоматизированным библиографическим поиском, организация которого является одной из наиболее сложных проблем автоматизации.

Автоматизированный поиск можно рассматривать как процесс, направленный на обнаружение релевантной библиографической информации, как действие, возникающее в ответ на поступивший запрос. Поэтому его можно определить как совокупность процессов поиска вторично-документальной информации, релевантной содержательным или формальным признакам документа, указанным в поступившем запросе.

Автоматизированный библиографический поиск является более эффективным, чем традиционный.

С помощью высокоэффективной библиографической системы могут быть частично преодолены барьеры, возникающие на пути потребителя к до-

кументу, которые образуют противоречия между постоянно возрастающими потребностями в документальных источниках информации и существующими возможностями их удовлетворения. Эту систему можно считать эффективной лишь в том случае, если она обеспечивает реализацию поисковой, коммуникативной и оценочной функции библиографической информации, т. е. содействует быстрому и результативному поиску нужных документов, своевременному ознакомлению потребителей с быстрорастущим потоком документов, позволяет выбрать из них наиболее ценные и т.д. [3].

Автоматизация процессов справочно-библиографического обслуживания приводит не только к более оперативному, полному, качественному удовлетворению запросов, но и повышает информационную культуру специалиста. В современных условиях информационная культура личности является необходимой основой успешной деятельности в любой сфере, социально престижным качеством специалиста.

Автоматизация библиографического поиска – одно из важных направлений дальнейшего повышения эффективности и качества справочно-библиографического обслуживания.

4. Метод компарации при работе с информационными объектами

Систематизация документов и организация предметно – тематического поиска в базах данных информационной системы определяется, прежде всего, инструментом (методикой) системы классификации, используемой в конкретном каталоге. Соответствие найденной информации запросу пользователя зависит от выбранных правил структурного объединения баз данных, в основе которых лежит методология классификации объектов. Для построения математической модели классификации документов в полнотекстовой базе данных и ключевых слов в словаре предлагается использовать один из основных методов теории интеллекта – метод сравнения, или метод компараторной идентификации информационных объектов [4, 6]. Классическая задача идентификации состоит в том, что по входному x и выходному y сигналам объекта определить закон $y = F(x)$ преобразования сигнала этим объектом. Такую идентификацию называют прямой, поскольку она осуществляется при непосредственном доступе к выходному сигналу объекта. Однако в ряде случаев возникает необходимость в косвенной идентификации объекта, когда у исследователя нет прямого доступа к выходному сигналу. Многие задачи этого типа можно решать методом компараторной идентификации объекта. Данный метод позволяет излагать основные положения теории интеллекта дедуктивным способом, исходя исключительно из физически наблюдаемых фактов, он хорошо зарекомендовал себя при обработке лингвистических объектов различных уровней языка.

Обрабатываемые информационными системами объекты являются дискретными, конечными и детерминированными, что позволяет использовать при обработке этих объектов метод компараторной идентификации. Сущность метода представлена на рис. 1.

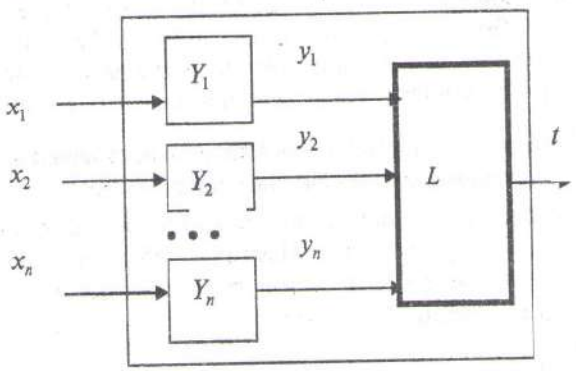

Рис 1. Сущность метода компараторной идентификации

На вход системы подается множество сигналов $x_1, x_2, \ldots, x_n$. Под сигналом мы понимаем условные знаки, служащие для передачи информации (тексты документов, ключевые понятия и т.д.). Входные сигналы берутся из конечных множеств $X_1, X_2, \ldots, X_n$, причем $x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n$. В результате работы системы, осуществляющей обработку информации, на выход поступает определенное множество элементов $y_1, y_2, \ldots, y_n$. В нашем случае под $y_1, y_2, \ldots, y_n$ можно понимать ключевые понятия, дескрипторы УДК, ББК и т.д. Причем $y_1 \in Y_1, y_2 \in Y_2, \ldots, y_n \in Y_n$.

Элементы $y_1, y_2, \ldots, y_n$ однозначно зависят от сигнала $x_1, x_2, \ldots, x_n$. Это означает, что существуют функции $y_1 = f_1(x_1), y_2 = f(x_2), \ldots, y_n = f(x_n)$, которые ставят в соответствие каждому

$$x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n$$

элемент

$$y_1 \in Y_1, y_2 \in Y_2, \ldots, y_n \in Y_n.$$

Причем каждому из сигналов $x_i \in X_i$ соответствует вполне определенный элемент $y_i \in Y_i$. Таким образом, каждая из функций f_i представляет собой сюръекцию, отображающую множество X_i на множество $Y_i, i \in \{1, 2, \ldots, n\}$. Функции f_i характеризуют способность системы соотносить информацию, передаваемую условным знаком – сигналом с элементом, отражающим его смысл и соответствующим той или иной классификации.

В ходе работы системы проверяется существование некоторого отношения Q, связывающего эле-

менты $y_1, y_2, \ldots, y_n$, которые появляются на выходе системы после сигналов $x_1, x_2, \ldots, x_n$, поступающих на вход системы. Таким образом, в процессе работы компаратор реализует предикат

$$q = Q(y_1, y_2, \ldots, y_n),$$

соответствующий отношению Q. Предикат q характеризует механизм сравнения элементов $y_1, y_2, \ldots, y_n$. Именно эта операция сравнения позволяет назвать данный метод методом компарации. Предикат

$$P(x_1, x_2, \ldots, x_n) = Q(f_1(x_1), f_2(x_2), \ldots, f_n(x_n))$$

характеризует работу системы, осуществляющую интеллектуальную обработку документальной информации, которая на сигналы $x_1, x_2, \ldots, x_n$ реагирует ответом $q = P(x_1, x_2, \ldots, x_n)$. Моделирование любой из задач аналитико-синтаксической обработки информации заключается в том, чтобы из свойств предиката P, осуществляющего компараторную идентификацию информационных объектов, извлечь внутреннюю структуру сигналов $x_1, x_2, \ldots, x_n$, элементов $y_1, y_2, \ldots, y_n$, вид функций $f_1, f_2, \ldots, f_n$ и вид предиката Q.

В общем случае система получает k заданий, которые выполняет поочередно для различных наборов входных сигналов. Закономерности обработки сигналов записываются в виде системы логических условий:

$$\begin{cases} K_1(L_1, L_2, \ldots, L_k) = 1 \\ K_2(L_1, L_2, \ldots, L_k) = 1 \\ \ldots \\ K_j(L_1, L_2, \ldots, L_k) = 1, \end{cases} \quad (1)$$

связывающих между собой предикатные переменные $L_1, L_2, \ldots, L_k$. Здесь $K_1, K_2, \ldots, K_j$ — предикаты от предикатов $L_1, L_2, \ldots, L_k$. Предикат $L_i(x_1, x_2, \ldots, x_n)$, $i \in \{1, 2, \ldots, k\}$ задан на декартовом произведении $X_{1i} \times X_{2i} \times \ldots \times X_{ni}$. Решение $L_1 = P_1, L_2 = P_2, \ldots, L_k = P_k$ удовлетворяет системе уравнений (1).

Под универсумом элементов U мы будем понимать всевозможные тексты документов полнотекстовой базы данных, вторичные документы (реферат, аннотация, библиография), ключевые понятия, дескрипторы, рубрики и подрубрики, входящие в рубрикатор информационной системы и т.д. Из элементов универсума образуются подмножества $X_{1i}, X_{2i}, \ldots, X_{ni}$, сообразуясь с конкретной задачей обработки информации. На декартовых произведениях $X_{1i} \times X_{2i} \times \ldots X_{ni}$, определенных предикатами P_i, характеризуют работу системы, выполняющую ту или иную аналитико-синтаксическую обработку документально-информационных потоков.

Вводятся предикатные переменные $L_1, L_2, \ldots, L_k$, которые связываются логическими уравнениями. Эти уравнения выступают исходными постулатами метода компараторной идентификации. Из них, как из аксиом, дедуктивно выводятся зависимости, характеризующие внутреннюю структуру элементов универсума U и предикатов $P_1, P_2, \ldots, P_k$.

Для решения этой задачи необходим единый универсальный математический аппарат. Желательно, чтобы этот математический аппарат был ориентирован также и на моделирование всех уровней лингвистической обработки текстов документов. Опыт исследования закономерностей передачи информации на естественном языке, а именно с такой информацией мы имеем дело в автоматизированных информационных системах, показывает, что рационально пользоваться одним формальным аппаратом описания закономерностей передачи и интеллектуального преобразования информации. Необходим формальный аппарат для описания предикатов, которые реализуются при любом виде интеллектуальной обработки текстовой информации, для формирования уравнений, описывающих свойства этих предикатов. Таким наиболее универсальным аппаратом, служащим для описания закономерностей обработки информации на естественном языке, и является алгебра конечных предикатов и метод компараторной идентификации [5, 6].

5. Метод компарации при работе с информационными объектами

Центральной задачей аналитико-синтаксической обработки документов является разбиение всех рассматриваемых текстов на классы эквивалентности с тождественным или почти тождественным смыслом, т. е. та или иная классификация документов. Для дальнейшей работы введем несколько рабочих определений. Определимся с понятиями документа, дескриптора, ключевого термина, дескрипторно-текстового предиката.

Под документом обычно понимают любой материальный носитель семантической информации, которая может быть выражена знаковой формой и зафиксирована любым образом. Множество документов, которые мы будем рассматривать, представляют собой некоторую, достаточно четко очерченную совокупность текстов полнотекстовой базы данных. Под полнотекстовой базой данных понимается именованная совокупность данных, отражающая состояния объектов и их отношений в определенной предметной области, в которой в качестве данных выступают любые целостные тексты.

Под *дескриптором* будем понимать ключевое слово, выбранное из группы условной эквивалентности ключевых слов. Предикат $P(t, r)$, заданный на декартовом произведении множества текстов и ключевых слов, называется дескрипторно-текстовым, поскольку он задает отношение между текста-

ми документов и ключевыми терминами, отражающими смысл этих документов. Введенный дескрипторно – текстовый предикат $P(t,r)$; полученные предикаты эквивалентности E_1 и E_2 позволяют провести разбиение документов и ключевых терминов на классы эквивалентности, представляющие собой определенные подтемы и подрубрики. При этом для каждого класса можно ввести обозначение подтемы, объединяющей данный класс. Ясно, что тексты, входящие в полученные нами классы эквивалентностей, не тождественны по смыслу – они являются эквивалентными относительно выражаемой ими подтемы.

Классу L_a всех текстов $t \in T$, относящихся к одной подтеме, содержащему текст $a \in T$, соответствует предикат $L_a(t) = E_1(t,a)$. Таким образом, получаем

$$L_a(t) = \forall r \in R(P(t,r) \sim P(a,r)). \quad (2)$$

Классу Q_b всех ключевых понятий $r \in R$, относящихся к одной подрубрике с ключевым понятием $b \in R$, соответствует предикат $Q_b(r) = E_2(r,b)$. Получаем формулу:

$$Q_b(r) = \forall t \in T(P(t,r) \sim P(t,b)). \quad (3)$$

Формулы (2) и (3) выражают деление текстов на подтемы и ключевых понятий на подрубрики через предикат θ_2 ,объективно определяемый классификатором.

Таким образом, метод компараторной идентификации позволяет автоматически разделять тексты документов полнотекстовых баз данных на тождественные (по отношению к определенной подтеме) и разбивать ключевые термины, относящиеся к данной предметной области, на классы эквивалентностей, представляющие собой определенную подрубрику. Применение данного метода в автоматизированных информационных системах позволит пользователю вести узкотематический и многоаспектный поиск документов в полнотекстовых базах данных.

6. Основные результаты и выводы

Обзор современного состояния и перспектив развития математического и лингвистического обеспечения информационных систем, показал, что значительную пользу может принести система детализированных моделей, служащих основой для конструирования интегральной информационной системы. Ключевой проблемой этой области осталась автоматизация анализа содержания текстов документов, т. е. аналитико-синтаксическая обработка (классификация, предметизация, систематизация и т. д.). Существующие информационные системы в своем большинстве статичны и не поддерживают динамичность информационных процессов, заключающихся, прежде всего, в изменении условий классификации, возникающих в случае роста предметной области и развития знания.

В отличие от других ранее применяемых методов, метод компараторной идентификации, предложенный в [5], и используемый в данной статье для моделирования процессов обработки информационных объектов, позволяет моделировать процедуры интеллектуальной (в том числе семантической) обработки документов, учитывая изменение условий классификации путем корректировки словаря терминов. Классификация и систематизация документов в полнотекстовой базе данных основывается на тождественности текстов документов по отношению к определенной подтеме и тождественности ключевых терминов по отношению к определенной подрубрике. Использование данного метода позволило ввести и обосновать понятие дескрипторно – текстового предиката, формально представляющего отношения между текстом и соответствующим ему ключевым термином.

Поскольку Data Mining — это процесс обработки ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности, то предложенный в данной работе метод компараторной идентификации по отношению к «сырым» данным можно квалифицировать как один из логических методов, разрабатываемых в Data Mining. Таким образом, рассмотренные модели и методы обработки текстовой информации, использующие основные подходы Data Mining и Text Mining, представляют интерес для разработчиков информационно – поисковых, экспертных, аналитических средств информационных систем широкого назначения, в том числе для обучающих систем различных модификаций.

Список литературы: 1. *Пономарёв В.В.* Концептуальная модель комплекса средств лингвистического и программного обеспечения экспертно – поисковой системы с элементами социопсихолингвистической детерминации. – М.: «ДИАЛОГ-МИФИ», 2004. — 176 с. 2. *Пириев В.Ф.* Стратегии организации информационно-поисковой работы в электронных библиотечных фондах // Стратегія комплектування фондів наукової бібліотеки: Тез. доп. міжнар. наук. конф. (8–10 жовтня, 1996 р.). 1996. – С. 61–62. 3. *Капырина А.А.* Автоматизация библиографического поиска как фактор повышения эффективности справочно-библиографического обслуживания. Минск: Информационный центр по вопросам культуры и творчества, 1996. — 20 с. 4. *Хайрова Н.Ф., Шаронова Н.В.* Автоматизированные информационные системы: задачи обработки информации – Харьков:Нар. укр. акад., 2002. – 120 с. 5. *Шабанов-Кушнаренко Ю.П.* Теория интеллекта: Проблемы и перспективы. – Харків: Вища шк., 1987. – 158 с. 6. *Шабанов-Кушнаренко Ю.П., Шаронова Н.В.* Компараторная идентификация лингвистических объектов. – К., ИСИО, 1993. – 116 с.

Поступила в редколлегию 21.10.2005