УДК 519.767.6

Н.Ф. Хайрова, Н.В. Шаронова, Н.В. Борисова

Информационная технология создания *OLAP*-кубов для представления многомерного пространства знаний коллекции документов

Рассмотрена необходимость включения неструктурированной информации в процесс аналитического обеспечения корпоративного управления, объединяющая процедуры направлений *Data Mining* и *Text Mining*. Разработана технология создания атрибута информационных понятий *OLAP*-куба коллекции.

The necessity of including the nonstructured information into the process of analytical securing of corporative management which unites the procedures of Data mining and Text Mining is considered. The technology of creating the attribute of the information notions of the OLAR-cube collection is developed.

Розглянуто необхідність включення неструктурованої інформації в процес аналітичного забезпечення корпоративного керування, яка об'єднує процедури спрямування *Data Mining* та *Text Mining*. Розроблено технологію створення атрибута інформаційних понять *OLAP*-куба колекції.

Введение. Глобальная проблема, существующая при обработке информации, заключается не только в необходимости ее выбора как из традиционных приложений, так и из разнородных источников, но и в превращении ее в знания, используемые для эффективного управления бизнесом и обществом в целом. Задача усложняется еще и тем, что около 85 процентов информации хранится не в СУБД, а в текстовых документах и файлах: *Web*-страницах, электронных письмах и аналогичных документах. Стремительный рост объема полнотекстовой информации по-прежнему продолжается не только в глобальных сетях, где каждые 18 месяцев происходит удвоение объема неструктурированной информации, но и в корпоративных (ведомственных) информационных системах.

Высокие темпы роста объема информационного пространства при доминировании в его составе неструктурированных данных (как по абсолютным показателям, так и по более высоким темпам прироста) создают угрозу превращения хранилищ информации в «кладбища информации» [1].

Ранее в области интеллектуальной обработки информации существовали близкие, но не пересекающиеся области *Data Mining* и *Text Mining*. Средства направления *Data Mining* в основном обрабатывали структурированную информацию, т.е. данные, а направление *Text Mining* занималось обработкой слабоструктурированной (текстовой) информации. Но сегодня в связи с тем, что объединенная картина информационного пространства корпорации, ее «информационный ландшафт» включает в себя как структурированные данные, так и неструктурированную информацию, возникает необходимость в конвергенции этих двух направлений.

Ключевые слова: *Text Mining*, *OLAP*-кубы, извлечение знаний, метод компараторной идентификации.

Постановка задачи

Традиционно включение неструктурированной информации в процесс аналитического обеспечения корпоративного управления требует ее предварительного интеллектуального осмысления и смысловой разметки метаданными, которая преимущественно осуществляется вручную. В статье предложен метод динамического построения многомерного представления исследуемой коллекции документов. Под документом, в данном контексте, подразумеваются любые объекты, представляющие собой необработанные тексты, поступающие в корпорацию из различных источников и не имеющие установленной ценности. Группа документов, по кото-

рым осуществляется обработка, есть коллекция, или массив текстов.

Предлагаемое многомерное представление информации коллекции документов позволит использовать навигацию по *OLAP*-кубу для подборки необходимых документов или фактов, извлекаемых из данных документов. Аналитик может погружаться в элементы разных измерений (например, административно-географические области публикации документа), просматривать документы в ячейках с нужными значениями частот и др. Дополнительно могут быть использованы общие методы анализа и прогноза данных.

Независимые измерения гиперкуба в рассматриваемой модели представляют собой многомерное пространство объективных и достоверных знаний исследуемой коллекции документов. Атрибутами, характеризующими качественные значения измерений (т.е. осями *OLAP*-куба), служат метаданные документа (рубрики, авторы, дата публикации, источники и др.) и семантико-фактографическая информация, извлекаемая из документов. Количественным показателем, помещенным в ячейки куба, на пересечении измерений есть агрегатная функция *COUNT*().

Одно из главных преимуществ *OLAP*-технологии, как известно, заключается в использовании операции сжатия гиперкуба и построении его сечений, которая появляется благодаря использованию значений атрибутов-измерений более высоких уровней иерархии и соответствующего агрегирования значений показателей вычисляемых полей. В создаваемых кубах такая иерархия строится по датам публикации, географическому расположению мест издания, вхождению информационных понятий в документ.

Но более интересной представляется предлагаемая в данном исследовании классификация всех информационных понятий массива документов по некоторым смысловым признакам, осуществляемая аналогично интеллекту человека, с выделением семантических идентификаторов иерархии. Для чего предлагается моделировать одну из высших форм интеллектуальной деятельности человека – понимание и отнесение лексических единиц языка к одному классу по некоторым смысловым признакам.

Цель исследования

Разработать метод построения многомерного представления информации коллекции документов в виде *OLAP*-кубов, включающих в себя атрибут информационных понятий коллекции документа, классифицированных по некоторым смысловым признакам. Для выделения семантических идентификаторов иерархии понятий коллекции документов моделировать интеллектуальное понимание и классификацию лексических единиц по некоторым смысловым признакам.

Предлагаемая информационная технология

Для формирования измерения информационных понятий коллекции документов с возможностью построения иерархических отношений предлагается использовать технологию, объединяющую анализ отдельных документов коллекции и компараторную идентификацию [2] документов и лексикона информационных понятий коллекции документов (рис. 1).

Рис. 1. Структурная схема формирования измерения информационных понятий коллекции документов *OLAP*-куба

На первом этапе осуществляется автоматический анализ отдельных документов коллекции, когда с некоторыми изменениями используется технология *Information Extraction* направления *Text Mining*, состоящая в выделении из документов, написанных на различных языках, информационных понятий, отражающих тематическую направленность документа с определенной степенью достоверности.

На этапе предпроцессорной обработки определяется язык документа и его формат, в соответствии с которым выбирается декодер. На

следующем этапе лингвистической обработки выделяются токены (лексемы) – экземпляры последовательности символов в документе, объединенные в семантическую единицу для обработки. Для этого при обработке русско- и украиноязычных материалов текст разбивается по пробелам и отбрасываются знаки пунктуации. При обработке образования притяжательных прилагательных и сокращений в английских текстах используется дополнительный алгоритм анализа использования апострофа.

Система выделения токенов включает в себя подсистему обработки лексем, имеющих графемное оформление. Для выявления семантически значимых смысловых понятий документа, имеющих графемное оформление, комбинируются два метода: словарный, куда включаются лексемы, графемное оформление которых трудно формализовать, и алгоритмический, использующий формальные правила универсальных случаев.

На морфологическом этапе обработки для приведения лексем к канонической форме используется стемминг, что предполагает приближенный эвристический процесс, в ходе которого от слов отбрасываются квазиокончания.

На этапе контекстного анализа для выделения терминологических словосочетаний используется следующий алгоритм: выделяются словосочетания по типу управления или согласования, затем осуществляется поиск подобных словосочетаний во всем документе. При наличии более трех подобных словосочетаний данное словосочетание будет считаться информационным понятием документа.

Для окончательного формирования словаря документа строится гиперболическая зависимость Ципфа ранга термина от частоты. В центральной зоне гиперболы находятся слова, максимально характеризующие данный текст и выражающие его специфичность, которые по определению есть информационными понятиями данного текста.

Таким образом, на первом этапе технологии в результате анализа всех документов коллекции формируется лексикон информационных понятий анализируемой текстовой коллекции хранилища документов, представляющий одно из измерений создаваемого гиперкуба *OLAP*.

Использование метода компараторной идентификации для определения иерархии семантических идентификаторов

Для объединения информационных понятий в классы эквивалентности с использованием смысловой классификации лексики предлагается использовать один из методов искусственного интеллекта – метод компараторной идентификации. Используя семантический треугольник знака Фреге [3] (рис. 2), на множестве лексикона информационных понятий $R = \{r_1, r_2, ..., r_n\}$ и множестве рассматриваемых в коллекции документов концептов $\Re = \{\rho_1, \rho_2, \ldots, \rho_m\}$, введем функцию понимания информационного понятия $\rho = f(r)$, где ρ – концепт информационного понятия. Под концептом подразумевается информация, которую несет r о возможных денотатах [4], т.е. совокупность суждений о каком-либо объекте, выражающая его сущность и относящая его по общим и специфическим признакам к предметам некоторого класса.

Рис. 2. Основные понятия модели смысловой классификации лексики

Понимая денотат, выражаемый именем понятия r, классификатор соотносит его с определенным концептом, смыслом, десигнатом ρ. Функция понимания лингвистического смыслового элемента описывает процесс установления классификатором тождества между именем информационного понятия и концептом, знаком которого он является. Если классификатор рассматривает множество информационных понятий лексикона коллекции документов, то функция f отображает множество понятий лексикона на множество всех значений функции, т.е. совокупность всех смыслов, порождаемых информационными понятиями из множества R.

Для выделения уровней иерархии измерения информационных понятий *OLAP*-куба необходимо выделить классы равнозначных или семантически близких информационных понятий данной предметной области. Такие информационные понятия соответствуют одним и тем же или близким по смыслу концептам, обычно относящимся к одному дескриптору, и, как показывают исследования, часто рассматриваемым в одном связном тексте деловой документации корпорации, соответствующей единой тематике.

Анализируя содержание текста документа из рассматриваемой коллекции документов и понимая его, классификатор обычно формирует в своем сознании некий смысл – основное значение документа. Смысл документа однозначно определяется породившим его текстом (рис. 3) [5].

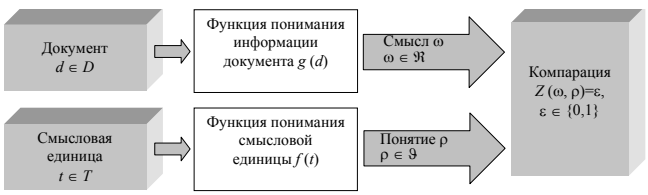

Рис. 3. Реализация метода компараторной идентификации для формирования классов эквивалентности лексики коллекции документов

Согласно определению толкового словаря [6], смысл – это идеальное содержание, идея, сущность, предназначение, конечная цель (ценность) чего-либо, целостное содержание какого-либо высказывания, не сводимое к значениям составляющих его частей и элементов, но само определяющее эти значения. Понимание классификатором текста документа обозначает компонент его мышления, психологическое состояние, определяющее корректное восприятие или интерпретацию данного документа, т.е. установление связи раскрываемых новых свойств объекта познания с уже известными.

Функция понимания текста $\omega = g(d)$ устанавливает зависимость смысла документа от имени (знака) документа, где $d \in D$, D – множество документов коллекции. Понимая смысл документа, классификатор однозначно устанавливает соответствие или несоответствие между определенным концептом, рассматриваемого в тексте денотата и смыслом текста, определяя предикат.

Используя метод компараторной идентификации, можно перейти от субъективного понимания смысла документов и концептов денотатов к объективному отношению между элементами лексикона информационных понятий коллекции документов и текстами документов коллекции, определяемым предикатом $P(d, r)$. Предикат аналитико-синтетической обработки документа $\varepsilon = Z(\omega, \rho)$, реализующий отношение предмета рассмотрения документа и концепта информационного понятия лексикона, связан с предикатом $P(d, r)$ отношением:

$$P(d, r) = Z((g(d), f(r)) = Z(\omega, \rho) \quad [7].$$

Используя предикат аналитико-синтетической обработки, можно получить предикаты эквивалентности, однозначно разбивающие лексикон информационных понятий исследуемой коллекции документов на слои разбиения, в которых все информационные понятия, относящиеся к одному слою, будут относиться к одному классу семантически близких смысловых единиц, объединяемых на оси измерений *OLAP*-куба:

$$\Psi_b(r) = \square d \in D(P(d, r) \sim P(d, b)).$$

Заключение. Таким образом, использование метода компараторной идентификации для объединения информационных понятий в классы эквивалентности с использованием смысловой классификации лексики позволяет выделить главный атрибут многомерного *OLAP*-куба тематической коллекции документов. Использование подобной технологии позволяет динамично строить многомерное представление массива слабоструктурированных документов, в которых в качестве измерений будут выступать как их метаданные, так и семантически значимые информационные понятия коллекции документов.

1. *Рубанов В.* Между стандартами управления и информационной стихией // Технологический прогноз. – 2010. – **3**. – С. 7–14.
2. *Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П.* Теория интеллекта: Учебник. – Харьков: СМИТ, 2007. – 576 с.

3. *Фреге Г.* Смысл и значение / Избран. работы – М.: Дом интеллектуальной книги, 1997. – 128 с.
4. *Поспелов Д.А.*, *Осипов Г.С.* Введение в прикладную семиотику // Новости искусственного интеллекта. – 2002. – № 6. – С. 28–35.
5. *Khairova N.* Multidimensional Representation of the Documents Collection // Computer Science and Information Technologies: Materials of the VIth Int. Scien. and Techn. Conf. CSIT'2011. – Lviv: Publ. House Vezha& Co, 2011. – P. 164–165.
6. *СЭС.* – М.: Сов. энциклопедия, 1988. – 1600 с.
7. *Хайрова Н.Ф.*, *Тарловский В.А.* Использование семантико-ориентированного лингвистического процессора для добывания новых знаний из потока документов / Вісн. НТУ «ХПІ»: Зб. наук. пр. – 2010. – № 67. – С. 132–138.

Поступила 01.12.2012
Тел. для справок: +38 052 707-6460, 714-6004, 707-6460, 337-5982, 707-6460, 99-1377, +38 050 966-3744, +38 066 795-4804, +38 098 423-1583 (Харьков)
E-mail: *nina_khajrova@yahoo.com*, *nvsharonova@mail.ru*, *borisova_nv@mail.ru*