УДК 004.652

М.А. Бутакова, С.М. Ковалев, Е.В. Климанская

МОДЕЛЬ РЕЛЕВАНТНОСТИ СЛАБОСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ В ТЕМПОРАЛЬНЫХ БАЗАХ ДАННЫХ*

Предлагается новый подход к моделированию и оценке релевантного поиска слабоструктурированной информации в темпоральных базах данных. Установлена связь между информационной слабой структурированностью данных, сохраняемых в распределенных сетевых информационных системах и информационной темпоральностью, возникающей в связи с потребностями обращения к информации, принадлежащей к определенному периоду. Рассмотрены основные модели описания слабоструктурированной информации в существующих базах данных. Представлены модели времени, на основе которых осуществляется построение темпоральных баз данных. Рассмотрены модели темпорального хранения данных, а также существующие подходы и реализация темпоральных баз данных. Представлена существующая вероятностная модель информационного ранжирования на основе многократной бернуллиевской языковой модели. Основным теоретическим результатом работы является развитие многократной бернуллиевской языковой модели для оценки информационной релевантности слабоструктурированной информации при её хранении в темпоральных базах данных.

Слабоструктурированные данные; темпоральные данные; базы данных; модель информационной релевантности.

M.A. Butakova, C.M. Kovalev, E.V. Klimanskaya

RELEVANCE MODEL OF SEMISTRUCTURED INFORMATION IN TEMPORAL DATABASES

In the article new approach to modeling and an assessment of relevant search of semistructured information in temporal databases is offered. Connection between information with weak structure of the data kept in distributed network information systems and information temporal, arising in connection with requirements of the appeal to information belonging by the certain period is established. The main models of the description of semistructured information in existing databases are considered. The models of time on the basis of which creation of temporal databases are presented. The models of temporal data storage, and also existing approaches and realization of temporal databases are considered. The existing probabilistic model of information ranging on the basis of repeated Bernoulli language model is presented. The main theoretical result of work is development of repeated Bernoulli language model for information relevance evaluation of semistructured information at its storage in temporal databases.

Semistructured data; temporal data; databases; information relevance model.

Введение. Представление и хранение информации в современных сетевых распределенных информационных системах и базах данных (БД) связано с рядом факторов, требующих доработки принципов поиска и отбора значимой информации. К таким факторам относятся недостаточная выраженность категории и тематики информации, проявляющаяся в информационной слабой структурированности. Другим фактором является потребность получения информации, связанной с фактами из прошлого или за некоторые, не всегда четко определенные интервалы времени. Данный аспект является информационной темпоральностью. Наложение данных факторов приводит к идеям усовершенствования подходов в хранению

* Работа выполнена при финансовой поддержке РФФИ, проекты: 12-08-00798-а, 13-08-12151-а, 13-07-13159-офи-РЖД.

темпоральной информации, обладающей в добавок признаками слабой структурированности. Интуитивно понятно, что имеющиеся реляционный и объектно-ориентированный подходы к хранению информации либо не достаточны в перечисленных подходах для адекватного отражения свойств слабоструктурированности и темпоральности информации, либо эти свойства могут храниться в избыточном виде с регистрацией всех информационных связей и временных событий, которые происходят с информационными сущностями и их атрибутами. Несмотря на то, что происходит существенное удешевление носителей и появляется принципиальная возможность хранения сверхбольших объемов информации, например, проект *CommonCrawl* (www.commoncrawl.org, «моментальный снимок всех web-страниц Интернета»), пользующийся сервисами *AMAZON Advanced Web Services*, разумнее всего подходить к решению задачи, с точки зрения повышения эффективности хранения темпоральных данных и интеллектуализации отбора слабоструктурированной информации, актуальной на некоторую временную дату.

Целью данной работы является разработка вероятностной модели релевантности слабоструктурированной информации на некоторый момент и интервал времени. Данная модель позволяет оценить насколько изменяется релевантность для запрашиваемой информации, имеющей признаки слабой структурированности, в зависимости от учета темпоральности её хранения. Следующий раздел предназначен для установления связи между свойствами слабоструктурированности и темпоральности информации.

Слабоструктурированная информация, БД и темпоральность. Термин «*semi-structured data*», слабоструктурированные данные впервые был обозначен два десятилетия назад в работе [1]. К слабоструктурированной информации относят информацию по некоторой сходной тематике, получаемую из разнородных источников и характеризующуюся различием видов ее представления. Наиболее удачным является представление информации, релевантной некоторой теме и получаемой в виде коллекции объектов с текстовым, графическим, видео и другим представлением. Мера релевантности, т.е. степени соответствия информации выбранной тематике, естественным образом может быть оценена по-разному, и, вследствие этого структура БД для хранения слабоструктурированной информации и схема данных может изменяться с течением времени. Данное обстоятельство вообще не означает, что схема данных не может быть определена четко, а означает, что задаваемая схема данных (например, количество и типы полей в таблицах, связи между таблицами) является действительной лишь некоторое (пусть даже весьма продолжительное) время.

Приведем пример слабоструктурированного документа. Особый интерес в отношении хранения и поиска представляют слабоструктурированные данные, представленные в виде цельных документов, относящихся к какой-либо категории или классу. Типичным примером слабоструктурированного документа, однако, четко попадающего во вполне определенный класс, является «счет-фактура». Он является наиболее проверяемым документом при аудите, но наибольшее количество судебных разбирательств связано именно с ним, потому, что Налоговый кодекс РФ в статье 169 «Счет-фактура» лишь общие требования по их заполнению. Таким образом, организации могут различным образом формировать нумерацию, адрес, подписи и другие реквизиты документа в процессе своей деятельности, реформирования организации и принципов учета. Счета-фактуры компаний с иностранным капиталом (инвойсы) могут содержать поля, практически не используемые в отечественных документах, например, поле *Value Added Tax* (налог на потребление, являющийся некоторым аналогом отечественного налога на добавленную стоимость) Следовательно, самым гибким путем адаптации авто-

матизированных систем обработки таких документов является применение документо-ориентированных БД. Одна из моделей слабоструктурированных данных *OEM* (*Object Exchange Model*) была реализована в системе *TSIMMIS* [2], а другая, построенная на принципах бисимуляционной эквивалентности, представлена в работе [3].

Элементы реализации идеи слабоструктурированной обработки и хранения данных имеются в бессхемных БД, относящихся к типу *NoSQL* [4] систем. Их особенностью, в частности является горизонтальное масштабирование хранилища данных и поддержка поиска и индексирования по произвольным полям, а в некоторых БД имеется возможность составления произвольных запросов выборки данных. Наиболее простым способом реализации слабоструктурированного хранения данных является динамическое хранилище ключей и значений, как реализовано в БД *Redis* и *Riak*. Другим подходом к обеспечению возможности динамического изменения структуры БД является столбцовая реализация хранения данных (противоположно строковой в реляционных базах данных), при которой есть возможность определения разного количества столбцов для различных строк. По такому принципу устроены БД *HBase, Cassandra, HyperTable*.

К таковым, в частности, относятся БД *MongoDB, CouchDB*. В БД *MongoDB* документ будет относиться и храниться в какой-либо коллекции, а форматом хранения служит структура на языке *JSON*. Схема БД *MongoDB,* полностью изменяемая, использующая технологию *Google MapReduce*, допускающая построения широкого спектра индексов документов: уникальных, составных, геопространственных и вложенных. Для устойчивости и надежности хранения данных применяется атомарность операций, журналирование, технология асинхронной репликации с сегментацией по нескольким наборам реплик. Бессхемная БД *CouchDB* также использует массово-списочные функции *map*/*reduce*, интерфейс *REST API* для непосредственной обработки удаленных данных через *HTTP* протокол, а также формат описания данных *JSON*.

Наряду с привлекательными возможностями документо-ориентированного хранения данных есть и особенности, которые не особенно приветствуются разработчиками автоматизированных систем на основе БД. Среди таковых: отсутствие транзакций, невозможность автоматического приведения типов данных, затрудненная работа с массивами данных в отношении их сортировки и фильтрации, требование приведения данных к определенному формату и отсутствие других привычных функций БД.

Таким образом, прослеживается связь с темпоральностью хранения и обработки рассматриваемого класса информации. Однако, темпоральность структур, применяемых для хранения слабоструктурированных данных, не должна нарушать общих принципов хранения данных в БД: во-первых, должны быть определены временные рамки актуальности и истинности данных; во-вторых, должны быть заданы временные интервалы взаимодействия пользователя с БД, в течение которых не изменяется структура и содержание запрашиваемых данных. Так же, как и для «обычных» баз данных, темпоральные базы данных должны поддерживать целостность, непротиворечивость и актуальность хранимых данных, а также темпоральную однородность (гомогенность) доменов хранимых значений.

Приведенные рассуждения лишь немного иллюстрируют «сложные взаимоотношения» между слабоструктурированностью и темпоральностью и её реализацией в БД, которая более подробно обсуждается в следующем разделе статьи.

Принципы темпоральности, реализуемые в базах данных. Исследования в области темпоральных баз данных ведутся более трех десятилетий [5, 6, 7]. Идеи темпоральной обработки данных как дополнения реляционной модели в дальней-

шем были реализованы в расширении языка запросов *SQL*-92 в языке *TQuel* [8], а затем в спецификации языка *TSQL2* [9]. Направление исследований темпоральности в базах данных основывается либо на фиксации фактов в некоторые моменты времени, либо на проверке действительности (*valid* – валидности) фактов в некоторых временных интервалах.

Моделями представления времени в темпоральных базах данных являются: непрерывная модель времени, в которой отсчеты времени изоморфны вещественным числам; уплотненная модель времени, в которой отсчеты времени изоморфны рациональным числам и дискретная модель времени, в которой временные отсчеты сопоставлены с целыми числами. Для логической связи модельного времени базы данных и реального времени сегменты времени гранулируются на «минуты», «часы», «дни» и др., а наименьшая из гранул (иногда называемая «хронон») может, в принципе, иметь различный размер (от наносекунд до столетий). Во всех моделях используются теоретико-множественные операции, определенные над единичными хрононами, их последовательностью и произвольными множествами хрононов. Множество хрононов, дополненное множеством смысловых предикатов, образует множество темпоров хранения данных, которое при полной упорядоченности по временному домену можно подвергать операциям отношения.

Различия в моделях представления времени привели к появлению различных моделей хранения данных в базах. Модели темпорального хранения данных рассматриваются на трех уровнях абстракции: 1) на концептуальном уровне, где являются расширениями для *ER* (*Entity-Relationship*) моделей; 2) на логическом уровне, как дополнение реляционного и объектно-ориентированного подходов; 3) на физическом уровне, где применяются специализированные типы данных для хранения временных значений, например, *TIMESTAMP*, *CALENDAR*. Среди наиболее известных моделей хранения (которых насчитывается более двух десятков) в темпоральных базах данных следует отметить [9], что они различаются по способу фиксации и отслеживания темпоральных событий на: 1) модели нетранзакционной темпоральности, в которых устанавливается связь между истинностью некоторых фактов и данных с некоторыми промежутками времени, объединяемых с помощью теоретико-множественных операций, в течение которых может изменяться валидность данных; 2) модели транзакционной темпоральности, в которых регистрируются только моменты времени, когда имели место некоторые факты или существовали некоторые данные; 3) битемпоральные модели, являющиеся комбинацией двух первых. Перечисленные модели хранения дают возможность реализации темпоральных систем управления базами данных (СУБД) в следующих вариантах: 1) СУБД с дампом таблиц; 2) СУБД с историей изменений; 3) СУБД с возвратом к предыдущему времени; 4) СУБД битемпоральная.

В отличие от значительного числа моделей темпорального хранения данных практическая реализация рассматриваемых идей в существующих СУБД не столь распространена [10]. Некоторые возможности темпорального хранения и обработки данных есть в СУБД *Oracle* (технология *FLASHBACK DATABASE*). Включение данной технологии дает возможность возврата базы данных (всей, а не отдельных таблиц) к некоторой предыдущей временной точке, что не в полной мере соответствует рассмотренным выше принципам. СУБД *Teradata* имеет возможности временного анализа данных *Teradata Temporal.* Данная технология предназначена для регистрации и отслеживания изменений данных на некотором интервале времени с помощью дополнительных полей в таблицах и запросах. В описании к СУБД *Teradata* [11] указывается, что в ней осуществлены битемпоральное хранение данных в таблицах и язык запросов *TSQL2*. Сравнительно недавно, в 2011 г., язык *SQL* в стандарте *SQL:2011* пополнился расширениями, которые позволяют создавать

темпоральные таблицы и оперировать с ними [12]. Язык *SQL:2011* частично поддерживает модель транзакционной темпоральности, запросы позволяют обновлять или удалять изменения в таблицах, произошедшие, начиная с некоторой стартовой до некоторой финальной даты. Поддерживается возможность обращения к предыдущим версиям. Битемпоральность поддерживается несколько в ином, чем в битемпоральной модели хранения данных представлении. Она существует в виде возможности сохранения версионности БД при обновлении изменений, привязанных к некоторой дате транзакции.

В следующем разделе рассмотрим основной результат: модель релевантности слабоструктурированной информации в темпоральных БД.

Модель релевантности информации. Предлагаемая модель основывается на принципе вероятностного ранжирования, предложенного в работе [13]. Для этого рассмотрим некоторый класс D слабоструктурированных документов как принадлежащий с некоторой вероятностью $P(D|R)/P(D|N)$ либо к классу R релевантных (соответствующих ожиданиям пользователей) документов или к классу N нерелевантных документов. Классы R и N могут рассматриваться с точки зрения языковых моделей, где $P(\omega|R)$ и $P(\omega|N)$ являются вероятностями нахождения слова ω в классе релевантных и нерелевантных документов. Оценка вероятности $P(D|R)$ может выполняться различными способами, а в терминах многократной Бернулиевской языковой модели:

$$P(D|R)=\prod_{\omega\in D}P(\omega|R)\prod_{\omega\notin D}(1-P(\omega|N)).$$

Ранг релевантности документа оценивается по соотношению:

$$\frac{P(D|R)}{P(D|N)}\sim\prod_{\omega\in D}\frac{P(\omega|R)}{P(\omega|N)},$$

для которого в [13] предложены методы приближенных вычислений. Эти вычисления основаны на предположении, что для слов $q_1,q_2,...,q_k$ в запросе и слов ω в релевантном документе существует одинаковое распределение M_R, а $M_R\in\boldsymbol{M}$ – семейство распределений. Выбирая распределение $M_{R_i}\in\boldsymbol{M}$ с вероятностью $P(M_{R_i})$ k раз, получаем, что вероятность наблюдения ω в $q_1,q_2,...,q_k$ будет следующей:

$$P(\omega,q_1,q_2,...,q_k)=\sum_{M_{R_i}\in\boldsymbol{M},i=1}^{k}P(M_{R_i})P(\omega,q_1,q_2,...,q_k|M_{R_i}). \quad (1)$$

Полагая, что ω и q_i распределены одинаково и независимы друг от друга, то для $M_{R_i}\in\boldsymbol{M}$

$$P(\omega,q_1,q_2,...,q_k|M)=P(\omega|M)\prod_{i=1}^{k}P(q_i|M). \quad (2)$$

Подставляя (2) в (1), получим

$$P(\omega,q_1,q_2,...,q_k)=\sum_{M\in\boldsymbol{M}}^{k}P(M)P(\omega|M_{R_i})\prod_{i=1}^{k}P(q_i|M). \quad (3)$$

Результат (3) известен из работы [14].

Добавим в выражение (3) время так, чтобы оно участвовало в запросе на поиск релевантного документа среди слабоструктурированных документов. Пусть гранулой будет $d(date)$, тогда

$$P(\omega|Q) = \sum_{d} P(\omega|d,Q)P(d|Q), \tag{4}$$

где $q_1, q_2, ..., q_k \in Q$, а $P(\omega|d,Q)$ и $P(d|Q)$ означают вероятности получения документа, содержащего слово ω на темпоральную гранулу d, и получения релевантного документа из темпоральной БД на временную гранулу d, соответственно.

Для оценки $P(\omega|d,Q)$ обратим внимание, что $P(\omega|d,Q)$ может быть рассчитана как сумма произведений $P(\omega|d,Q)$ и $P(d|Q)$ (точнее $P(d|d_i,Q)$, но размер гранулы времени не меняется), поэтому для непрерывного времени $T \in R$:

$$P(\omega|d,Q) = \sum_{d\in(T\in R)} P(\omega|d)P(d|d_t,Q) = \sum_{d\in(T\in R)} P(\omega|d)P(d|,Q) \sim$$

$$\sim \sum_{d\in(T\in R)} P(\omega|d)P(d)\prod_{i=1}^{K} P(q_i|d), \tag{5}$$

где $P(\omega|d)$ – вероятность запроса слова ω в течение временной гранулы d.

Подставляя (5) в (4), получаем

$$P(\omega|Q) \sim \sum_{d} P(d|Q) \sum_{d} P(\omega|d)P(d)\prod_{k} P(q_k|,d). \tag{6}$$

В формуле (6) остается оценить $P(d|Q)$ – вероятность получения релевантной информации на время d по запросу Q. Для этого воспользуемся мерой информационного подобия $\bar{\omega}$ из работы [15]. Учитывая, что мера информационного подобия $\bar{\omega} \in [0...1]$ и несет смысл вероятности получения релевантных документов из числа N документов, то (6) записываем в виде

$$P(\omega|Q) \sim \bar{\omega} \sum_{d} P(\omega|d)P(d)\prod_{k} P(q_k|,d). \tag{7}$$

Формула (7) имеет смысл вероятностного получения некоторого слабоструктурированного документа, содержащего слово ω в запросе Q к темпоральной базе данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Luniewski A., Schwarz P., Shoens K., Stamos J., Thomas J.* Information Organization using Rufus // In Proc. ACM SIGMOD Int. Conf. Of Management of Data. 1993. – P. 560-561.
2. *Gaarsia-Molina H., Papakonstantiniu Y., Quass D., Rajaraman A., Sagiv Y., Ulman J., Windom J.* The THIMMIS project: integration of heterogeneous information sources // J. Intell. Inf. Syst. – 1997. – Vol. 8, № 2. – P. 117-132.
3. *Buneman P., Davidson S., Suciu D.* Programming Constructs for Unstructured Data // In Proc. Workshop on Database Programming Languages. – 1995.
4. *Редмонд Э., Уилсон Д.Р.* Семь баз данных за семь недель. Введение в современные базы данных и идеологию NoSQL. – М.: ДМК Пресс, 2013. – 384 с.
5. *Bubenko J.A. Jr.* The temporal dimension in information modeling // Technical Report RC 6187 №26479. – IBM Thomas J Watson Research Center. – 1976.
6. *Snodgrass R., Ilsoo Ahn*. Temporal Databases // Computer. – 1986. – Vol. 19, № 9. – P. 35-42.
7. *Snodgrass R.* A taxonomy of time databases // SIGMOD'85 Proc. of the 1985 ACM SIGMOD Int. Conf. on Management of Data. – 1985. – Vol. 14, № 4. – P. 236-246.

8. *Snodgrass R.* The temporal query language TQuel // ACM Trans. on Database System (TODS). – 1987. – Vol. 12, № 2. – P. 247-298.
9. *Snodgrass R.* The TSQL2 Temporal Query Language. – Springer Science Business Media, LLC. – 1995. – 647 p.
10. *Новиков Б.А. Горшкова Е.А.* Темпоральные базы данных: от теории к практическому использованию // Программирование. – 2008. – Т. 34, № 1. – С. 1-6.
11. *Al-Kateb M., Ghazal A., Crolotle A., Bhashyam R., Chimanchode J., Pakala S.P.* Temporal query processing in Teradata // EDBT'13 Proc. of the 16h Int. Conf. on Extending Database Technology. – 2013. – P. 573-578.
12. *Kulkarny K., Michels J.-E.* Temporal features in SQL: 2011 // ACM SIGMOD'12. – 2012. – Vol. 41, № 3. – P. 34-43.
13. *Robertson S.E.* The Probability Ranking Principle in IR // Readings in information retrieval. Morgan Kaufmann Publishers Inc. – San Francisco, 1997. – P. 281-286.
14. *Aurenko V., Croft W.B.* Relevance-based language models // In Proceeding of SIGIR 2001. – 2001. – P. 120-127.
15. *Бутакова М.А., Климанская Е.В., Янц В.И.* Мера информационного подобия для анализа слабоструктурированной информации // Современные проблемы науки и образования (электронный научный журнал). – 2013. – № 6.

Статью рекомендовал к опубликованию д.т.н., профессор Э.А. Мамаев.

Бутакова Мария Александровна – Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Ростовский государственный университет путей сообщения»; e-mail: butakova@rgups.ru; 344006, г. Ростов-на-Дону, пл. им. Ростовского Стрелкового Полка Народного ополчения, 2; тел.: 88632726543; кафедра информатики; профессор.

Ковалев Сергей Михайлович – e-mail: ksm@rfniias.ru; тел.: 88632726302; кафедра автоматики и телемеханики на железнодорожном транспорте; профессор.

Климанская Елена Владимировна – e-mail: elenaklimanskaja@mail.ru; тел.: 88632726543; кафедра информатики; аспирантка.

Butakova Maria Aleksandrovna – Federal State-Owned Budget Educational Establishment of Higher Vocational Education "Rostov State Transport University"; e-mail: butakova@rgups.ru; 2, square n.a. Rostovskogo Strelkovogo Polka Narodnogo Opolchenija, Rostov-on-Don, 344006, Russia; phone: +78632726543; the department of informatics; professor.

Kovalev Sergey Mikhailovich – e-mail: ksm@rfniias.ru; phone: +78632726543; the department of automatics and telemechanics on railway transport; professor.

Klimanskaya Elena Vladimirovna – e-mail: elenaklimanskaja@mail.ru; phone: +78632726543; the department of informatics; postgraduate student.

УДК 519.7

В.И. Финаев, Е.Н. Павленко, С.В. Кирильчик

РЕШЕНИЕ ЗАДАЧ УПРАВЛЕНИЯ С ПРИМЕНЕНИЕМ ИНТЕЛЛЕКТУАЛЬНЫХ ГИБРИДНЫХ СИСТЕМ*

При решении задач управления многими производственными процессами и объектами существует сложность в разработке эффективной системы управления, связанная с наличием неопределённости в исходных данных. Эти неопределённости могут касаться параметров объекта, условий его функционирования, возмущений от внешней среды, неопреде-

* Материалы статьи подготовлены в рамках выполнения работ по гранту Российского научного фонда № 14-19-01533