Неструктурированные данные с временно́й составляющей в библиотеках хранения данных

В. Абрамович, П. Кальчинский, К. Вецель

В статье вводится понятие документов с нечеткой временно́й составляющей в электронных библиотеках хранения данных (Data Warehouse Library – DWL). В противоположность временны́м последовательностям в хранилищах данных понятие неструктурированного содержимого временно́й размерности в библиотеках хранения данных довольно нечетко. Более того, временна́я компонента в документах и данных зависит от проводимого анализа (то есть прогнозирования). Следовательно, концепция соответствующих документов с временно́й составляющей в DWL может стать решающей во всей концепции поддержания баз данных с неструктурированным содержанием.

1. Введение

В первой части настоящей статьи мы собираемся кратко представить проект автоматического обеспечения корпоративных баз данных, содержащих документы, отобранные из подписных веб-источников. Это ведет к созданию библиотеки, которая содержит неструктурированную информацию в дополнение к структурированному содержимому баз данных. Далее мы кратко опишем существующие техники обработки документов временно́й размерности до их включения в DWL. В последней части статьи мы представим преимущества документов временно́й размерности, которые, как мы полагаем, являются решающими в идее организации гиперхранилищ информации для обоснования и принятия решений.

2. Библиотеки хранения данных

Поддержка хранения данных с неструктурированным содержимым ни коим образом не состыковывалась с парадигмой хранения, пока не появились первые серьезные предложения по этому вопросу (напр. [6, 8, 12]). Мы начали исследования в этой области в 1997 году. Начиная с этого момента, мы развиваем концепцию отбора документов, собранных из предварительно определенных веб-источников в корпоративные хранилища данных. В противоположность традиционной задаче отбора, отбор профилей в нашем решении производится автоматически на основании содержимого хранилища данных. Мы полагаем, что дополнение структурированного содержимого неструктурированными данными, выбираемыми автоматически из внешних источников, положительно повлияет на решения, принимаемые компанией. Хранилище данных, расширенное таким образом, иногда называют информационным гиперхранилищем (superstore). Высший уровень иерархии метаданных в области бизнеса обычно состоит из набора разделов хранилища. Мы будем рассматривать их в качестве объектов, нуждающихся во внешней информационной поддержке. Информация, необходимая для каждого конкретного раздела хранилища, может быть указана в профиле отбора информации, который в дальнейшем будем называть профилем раздела [3].

Наше решение основано на понятии фильтра для выборки часто запрашиваемой информации [3], который на основании информации, хранящейся среди метаданных компании, строит библиотеку подходящих документов для хранилища данных. В то время как пользователи работают со структурированным содержимым хранилища данных, система запрашивает DWL и выбирает те документы, которые относятся к данным хранилища, просматриваемым в настоящий момент. Затем документы предоставляются пользователю в форме страниц, похожих на результат поиска с помощью поисковой системы в интернет. Мы допускаем, что хранилище данных имеет организацию, схожую с Web [10].

Отличительной чертой нашего подхода является (1) предварительная фильтрация, (2) автоматическое разделение хранилища дан-

ных на профили и (3) создание неизменяющегося набора документов, отобранных из Web, по профилям разделов хранилища. Под предварительной фильтрацией мы понимаем ручной выбор внешних информационных провайдеров, которые и будут подвергнуты фильтрации. В этом случае автоматический фильтр отбирает документы только из проверенных источников, на которые компания решила подписаться. Подписка очень важна для всего процесса отбора. Когда организация подписывается на определенный источник информации, он заранее должен быть проверенным в первую очередь на соответствие нуждам компании и надежность. Предполагается, что профиль раздела отражает потребности в информации для данного конкретного раздела, поэтому он незначительно меняется со временем. Усложненные механизмы создания и модернизации профилей разделов хранилища являются главной темой другой программы исследований. Более подробную информацию по данной теме можно найти в [2, 3].

Должным образом организованная DWL будет содержать документы, отобранные по профилям разделов из надежных веб-источников. Эти документы будут описаны при помощи метаинформации, хранящейся в базе данных управления DWL, похожей на информацию, хранящуюся в метаданных компании. Следовательно, документы в DWL будут готовы к просмотру через структурированное содержимое хранилища данных.

3. Web-документы с временной составляющей

В этой статье мы собираемся уделить должное внимание аспекту DWL документов с временной составляющей. В соответствии с парадигмой хранения данных все элементы хранилища должны соответствовать определенному моменту времени [9, 11]. Поскольку неструктурированное содержимое документов бывает нечетким и двусмысленным, их упорядочивание по времени бывает невозможным. Наиболее удобно пользоваться DWL, представленной в виде множества документов, соотносимых с данными, просматриваемыми в настоящий момент (например, годовым отчетом или диаграм-

мой), при этом нет необходимости вручную формировать соответствующие запросы. Поддержание временн**о**го соответствия между документами и данными может стать решающим для их дальнейшего эффективного использования.

Перед тем, как каждый конкретный документ будет включен в DWL, он должен быть описан при помощи метаинформации. Эта метаинформация позволяет использовать данный документ, ассоциируя его со структурированным содержимым хранилища. В настоящее время семантика отбираемых документов выявляется с помощью существующих техник индексации [5, 7], при этом извлечение метаинформации о временн**о**й составляющей крайне затруднено.

Более того, оценка соответствия по времени DWL-документа содержимому хранилища, просматриваемому в текущий момент, зависит от насущных целей. Например, документы, описывающие будущее (со ссылкой на набор данных, анализируемых в настоящий момент), могут быть полезны в задаче прогнозирования. В задаче анализа падения уровня продаж, напротив, временн**у**ю составляющую, определенную в запросах, необходимо отодвинуть назад (см. рис. 1).

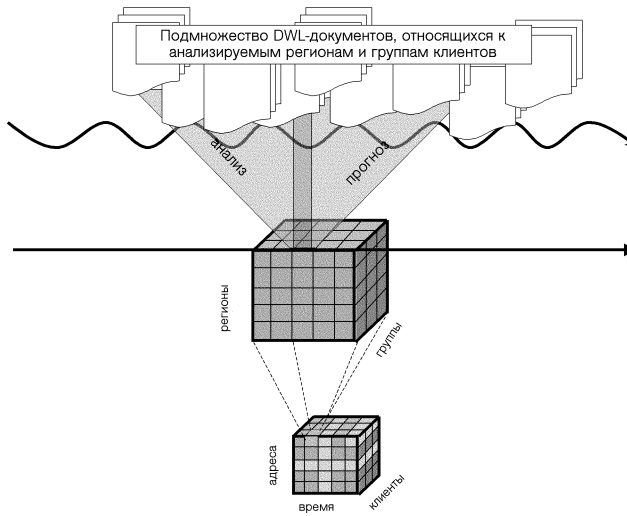

Рис. 1. Временн**о**й аспект DWL-документов

3.1. Нечеткая временная составляющая

Обычно упорядочивание по времени транзакционных данных не составляет большого труда и производится во время процесса «очистки» данных [4]. В сущности, даже простейший факт в транзакционных системах четко описывается временными переменными, в то время как содержание документа (по отношению к моменту его создания) может:

- относиться к ситуации в настоящем, например, текущие курсы валют или рыночные показатели,
- относиться к ситуации в прошлом, например, процент инфляции за прошедший период и комментарии,
- относиться к ситуации в будущем, например, прогнозирование процентной ставки,
- быть комбинацией вышеперечисленного.

По указанным причинам точное упорядочивание по времени внешних документов в принципе невыполнимо. И все же, на наш взгляд, можно указать промежуток времени, который покрывается содержанием каждого конкретного документа. Следовательно, временную компоненту документов, собранных в DWL, можно описать нечетким числом. Такое указание нечеткого числа для каждого документа в DWL коренным образом изменит подход к использованию информационных гиперхранилищ.

Промежуток времени, который покрывается документом, отобранным из некоторого веб-источника, может быть закодирован нечетким числом t_0t_1, где $[t_0; t_1]$ обозначает отрезок временной оси, к которому относится содержание данного документа с наибольшей вероятностью (π). В данном случае удобно использовать многоугольное нечеткое число (с многоугольным шаблоном принадлежности), как это показано на рис. 2.

Информацию, необходимую для упорядочивания по времени некоторого документа можно хранить либо в мета-секции этого документа, либо в базе данных управления DWL. Пусть новый документ, полученный в октябре 2000 г., содержит анализ уровня инфляции за предыдущие 3 месяца и прогноз на следующие два месяца. Описание

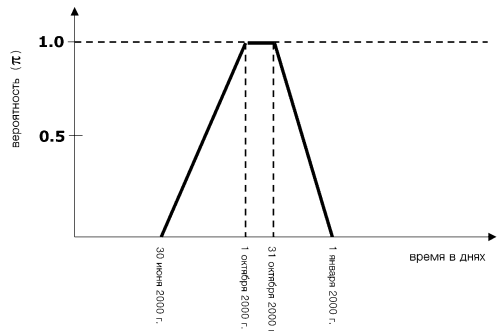

Рис. 2. Нечеткое упорядочивание по времени DWL-документа

упорядочивания, соответствующее рис. 2, может быть представлено в следующей форме:

<META NAME="time dimension" CONTENT="30Jun2000 1Oct2000 31Oct2000 1Jan2000">,

что обозначает, что документ вероятнее всего относится к октябрю 2000 г.

Конечно, все составляющие в хранилище данных дискретны, но кодирование нечетких чисел выпуклыми многоугольниками позволяет легко осуществлять сравнения и интерполяцию, и, кроме того, такой подход является более гибким, чем фиксированное упорядочивание. К примеру, вероятности $\pi_{31\text{ июля }2000}$ и $\pi_{30\text{ ноября }2000}$ вышеуказанного документа можно приблизить следующим образом:

$\pi_{31\text{ июля }2000} = 0,67$

$\pi_{30\text{ ноября }2000} = 0,52$

Даже малые значения вероятности указывают, что данный документ связан с определенным временны**м** интервалом. Это замечание может оказаться полезным для пользователей, проводящих подробный анализ.

Одним из основных допущений нашего решения является предположение о том, что каждый документ, включенный в DWL, должен быть надлежащим образом упорядочен по времени. Это допущение вытекает из основной парадигмы хранения данных [9, 11] и существенным образом влияет на процесс отбора документов. При

этом документы, считающиеся подходящими в терминах профиля раздела, будут отброшены, если они не могут быть соответствующим образом упорядочены.

Выбор соответствующего интервала для упорядочивания DWL-документов определяет оценку того, насколько они подходят структурированному содержимому хранилища данных. Поскольку описание с помощью нечетких чисел может быть легко расширено на более крупные интервалы, мы можем дать каждому документу максимально точное из возможных описаний временной составляющей. Положим, что день – подходящая временная единица для описания документов, отобранных из Web-источников. Практически не существует документов, содержание которых значительно меняется в течение 24 часов с точки зрения принятия решения.

Чтобы соответствующим образом проиндексировать нечеткие временные интервалы, можно воспользоваться одной из методик, описанных ниже.

3.2. Методики упорядочивания документов по времени

Чтобы обозначить нечеткий промежуток времени, покрываемый содержанием документа, необходимо использовать несколько различных подходов. Для решения этой сложной задачи оказываются полезными техники информационного поиска, анализа источника документа, анализа структуры документа и индексации его содержания. Необходимо ввести некоторые правила для упорядочивания по времени.

Самым простым способом для определения временного интервала для DWL-документа является следующий: просмотреть его содержание и отметить момент времени, в который он появляется во внешнем источнике. Некоторая дополнительная информация об источнике тоже будет необходима. К примеру, если провайдер публикует обновленные прогнозы на 14 дней по одному и тому же URL-адресу, обновленные документы, извлекаемые ежедневно можно описать нечетким числом, определяемым следующей формулой: $(d_0 - 1, d_0, d_0 + 13, d_0 + 14)$, где d_0 обозначает день, когда документ был опубликован и извлечен поисковой системой. Значительным ми-

нусом этого метода является то, что его можно применять только к часто обновляемым источникам, таким, как прогнозы погоды или бизнес новости.

Усложненные методы упорядочивания по времени основаны на подробном анализе содержания документа. Некоторые техники, такие как извлечение информации (information extracting – IE) позволяют структурировать текстовые документы. В противоположность индексированию техника извлечения информации позволяет проанализировать информацию, которую надо выделить, по контексту и представить ее в структурированной форме в виде таблиц. Это значит, что средства IE смогут выявить не только точные ссылки на время, такие как «15 октября 2000», но и неточные, такие как, «месяцем ранее». Более того, эта методика позволяет отбросить временные соответствия, которые не повлияют на упорядочивание документа, например, такие как «самое значительное падение цен с 19 века». Техника извлечения информации (IE) основана на средствах обработки подтекста (shallow text processing). Эти механизмы намного сложнее механизмов обычной индексации текстов. Хорошим примером тому может послужить система, разработанная Нейманом [13]. Мы полагаем, что извлечение информации о времени из содержимого каждого документа облегчит упорядочивание по времени DWL-документов.

Конечно, применение этих методик не приведет к упорядочиванию документов по времени, сравнимому с проделанным вручную. И, тем не менее, придется довольствоваться существующими способами до тех пор, пока не будут разработаны более совершенные методы анализа текстов, и не только статистические. Иначе ни одна компания или организация не сможет обработать вручную столь большие потоки неструктурированной информации.

4. Преимущества для пользователей

4.1. Дополнение к структурированным документам

Нечеткое упорядочивание по времени DWL-документов предусматривает дополнительные возможности сортировки документов,

семантически связанных со структурированным содержимым, просматриваемым в текущий момент (текущим состоянием хранилища). Этот критерий может быть частично использован для больших наборов документов. Во время того, как проводится анализ по многим размерностям, пользователю должно быть доступно некоторое подмножество DWL. Каждый документ будет считаться соответствующим структурированному содержимому хранилища, если он подходит по смыслу (семантике) и по промежутку времени, описываемому данными.

Выбор надлежащих документов по координатным осям, таким как «клиент», «регион» или «продукт» произвести довольно просто до тех пор, пока соответствие по времени зависит от целей проводимого анализа. Эта ситуация проиллюстрирована на рис. 3.

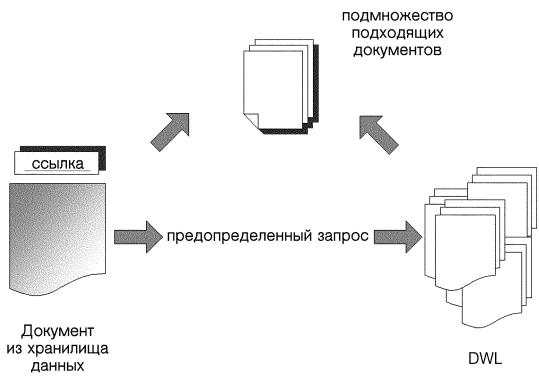

Рис. 3. использование DWL

С DWL можно проводить такие же операции, что и с обычным набором документов или поисковой системой, – формировать запросы, разбивать на категории или кластеры. Поэтому вместо использования предопределенных запросов, опытные пользователи могут также запрашивать нужную в данный момент специфическую информацию.

Временная компонента одинакова для всех существующих моделей хранилищ данных. Она является частью парадигмы хранения данных. Следовательно, дальнейшие исследования в этой области оправданы.

4.2. Преимущества нечеткого упорядочивания по времени в DWL

Основная цель создания DWL – предоставлять пользователю подходящие документы в процессе того, как он просматривает структурированное содержимое хранилища данных. К примеру, если пользователь пролистывает годовой отчет, ему будет предложено подмножество DWL-документов, соответствующих содержимому, просматриваемому в текущий момент. Эти документы будут извлекаться в режиме on-line путем формирования предопределенных запросов в DWL. Этот процесс, безусловно, требует наличия определенных моделей метаданных и определенных механизмов построения вышеупомянутых запросов. Эта область осталась за рамками настоящей статьи и отражена в [1, 2].

Предполагается, что каждое хранилище данных содержит набор документов, относящихся к разному времени. Это значит, что каждый элемент в нем должен соответствовать определенному моменту или периоду времени. Следовательно, анализ, проводимый пользователем, распространяется на определенный период, такой, как год, квартал или «1 октября 2000 г. – 25 октября 2000 г.»

Имея в информационной системе компании упорядоченную по времени электронную библиотеку хранения данных, мы могли бы выделить из нее документы, относящиеся к промежутку времени, покрываемому содержимым хранилища, которое просматривается в текущий момент, а после этого провести стандартную операцию выбора информации [5]. Выборку легко произвести, организовав запрос к базе данных управления DWL. Кроме того, указанный промежуток времени является нечетким и пользователь может указать свои предпочтения заданием вероятностного порога p_0. Документы, описывающие определенный момент на оси времени с вероятностью π большей p_0, будут считаться подходящими данному множеству объектов хранилища.

Благодаря нечеткому упорядочиванию по времени DWL-документов, пользователи могут указывать различные вероятностные пороги p_i для каждого этапа анализа. В задаче прогнозирования, например, можно установить большее значение порога для документов, относящихся к будущему, чем для документов, относящихся к

данным, просматриваемым в настоящий момент. Это приведет к более точному временному соответствию в документах, относящихся к будущим ситуациям.

4.3. Упорядочивание документов во множестве ответов

Указание порогов вероятности при нечетком упорядочивании по времени DWL-документов позволяет выделять нужные данные до того, как сформирован запрос для выбора информации (information retrieval). После этого, получившееся подмножество сравнивается с результатом предопределенного запроса, на основании текущего состояния хранилища данных. Документы, признанные соответствующими запросу, считаются подходящими и предоставляются пользователю.

Описанная ситуация похожа на традиционную задачу выбора информации. Следовательно, документы во множестве ответов могут быть упорядочены по значению каких-либо распространенных мер похожести, таких как, например, косинусная мера [5]:

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

Веса терминов $w_{i,j}$ документа $\mathbf{d}_i$ обычно вычисляются по известным формулам, а веса терминов для запроса $\mathbf{q}$ вычисляются по формуле Солтона–Бакли. Более подробную информацию по этому вопросу можно найти в [5, 14, 15].

В библиотеке данных упорядочивание документов можно модифицировать, введя для каждого документа максимальное значение функции принадлежности ($\pi_{\max}$), которое достигается в течение указанного периода времени. Например, косинусная мера для конкретного документа может быть модифицирована:

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \times \pi_{\max}^k$$

Значение экспоненты $k \geqslant 0$ определяет влияние временной составляющей на упорядочивание документа. Поэтому значение меры похожести не изменится для документов, относящихся к указанному промежутку времени ($\pi_{\max} = 1$), но будет уменьшено для $\pi_{\max} < 1$ соответственно.

Качество такого упорядочивания документов определяется качеством механизмов упорядочивания по времени, описанных в этой статье.

5. Выводы и направления дальнейших исследований

Концепция временной составляющей для документов в хранилище данных может стать решающей в общей концепции поддержки хранилищ с неструктурированным содержимым. Нечеткое упорядочивание по времени неструктурированной информации положительно влияет на возможности ее дальнейшего использования. Более того, нечеткое распределение по времени открывает новые пути увеличения качества поиска, производимого в DWL. Наиболее сложной задачей, без сомнения, является разработка механизмов извлечения информации о времени из документов. Указание общего значения k (см. формулу модифицированной меры похожести), определяющего воздействие временной компоненты на оценку соответствия между документом и запросом, требует дальнейших исследований. Документы, подходящим образом упорядоченные по времени в DWL, могут увеличить качество автоматического поиска и, следовательно, положительно повлиять на процесс выработки решений в организации.

Список литературы

[1] Abramowicz W., Kalczynski P.J. Intelligent Agents to Supply the HyperSDI System // Proceedings of the Sixth Annual Conference KSW 2000. Sep. 14–16 2000. Ciechocinek, Poland.

[2] Abramowicz W., Wecel K. Automatic Building HyperSDI Profiles on the Basis of Warehouse Metadata // Proceedings of the Sixth Annual Conference KSW 2000. Sep. 14–16 2000. Ciechocinek, Poland.

[3] Abramowicz W., Kalczynski P.J., Wecel K. Information Filters Supplying Data Warehouses with Benchmarking Information // Selected Aspects of Knowledge Discovery in Business information Systems / Abramowicz W., Zurada J. (eds.). Ch. 1. USA: Kluwer Academics, 2000.

[4] Adelman Sid. Data Quality // Data Warehouse – Practical Advice from the Experts / Bischoff J., Alexander T. (eds.) Ch. 10. USA: Prentice Hall Inc., 1997.

[5] Bayeza-Yates R., Berthier R.-N. Modern Information Retrieval. USA: ACM Press, Addison Wesley Longman Limited, 1999.

[6] Bischoff J., Yevich R. The SuperStore: Building More than a Data Warehouse // Database Programming and Design. Sep. 1996.

[7] Dumais S.T. Using LSI for information filtering // TREC-3 experiments / D. Harman (ed.) The Third Text REtrieval Conference (TREC3). National Institute of Standards and Technology. Special Publication. 1995.

[8] Hackathorn R.D. Web Farming for the Data Warehouse. USA: Morgan Kaufmann Publishers Inc., 1999.

[9] Inmon W.H., Hackathorn R.D. Using the Data Warehouse. New York: John Wiley & Sons, 1994.

[10] Kimball R., Merz R. The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse. USA: John Wiley & Sons, 2000.

[11] Kimball R. The Data Warehouse Toolkit. USA: John Wiley & Sons, 1996.

[12] Mattison R. Web Warehousing and Knowledge Management. McGraw-Hill Companies Inc., 1999.

[13] Neumann G., Schmeier S. Combining Shallow Text Processing and Machine Learning in Real World Applications // Machine Learning for Information Filtering. IJCAI 99. Sweden.

[14] van Rijsbergen C.J. Information Retrieval. London: Butterworths, 1979. (http://www.dcs.gla.ac.uk/Keith/Preface.html).

[15] Salton G., McGill M. Introduction to Modern Information Retrieval. McGraw-Hill Book Company, 1983.

[16] Text Retrieval Conference 1992-2000 resources and proceedings. (http://trec.nist.gov).