Técnicas de Banco de Dados para a World Wide Web - Resumo

FÁBIO FAGUNDES SILVEIRA

Universidade da Região da Campanha
CCEI - Centro de Ciências da Economia e Informática
Av. Tupy Silveira, 2099, 96400-030 Bagé, RS, Brasil
e-mail: ffs@urcamp.tche.br

Resumo. Este artigo apresenta um resumo da pesquisa [FLO98], que tem por objetivo a discussão referente a aplicação de conceitos de banco de dados para pesquisas e gerenciamento de informações na World Wide Web.

Palavras-chave: Banco de Dados, WWW, pesquisa, dados semi-estruturados

Abstract. This paper presents a summary of the article [FLO98], whose aims are to discuss database concepts aplications for query and management of informations on the WWW.

Keywords: Databases, WWW, query, semistructured data.

1. Introdução

Atualmente, a World Wide Web, ou simplesmente web, tem se caracterizado como um dos maiores mecanismos de disseminação de informações. Paralelamente à crescente utilização deste veículo e conseqüente crescimento de informações disponíveis, é necessário a aplicação de conceitos de Banco de Dados para tornar possível os processos de gerenciamento e utilização das mesmas. O enfoque central dá-se ao fato de que os dados contidos na rede são formados por textos, figuras, arquivos de áudio e vídeo, sendo estes dados considerados não estruturados [HDN98]. Desta forma, torna-se mais complexo o processo de gerenciamento destes dados do que em bancos de dados convencionais. Juntamente com as técnicas de BD, outras tecnologias visam igualmente ajudar na tarefa de administração, entre as quais pode-se citar a inteligência artificial e a hipermídia. Entretanto, com relação à aplicação de conceitos de BD, é possível destacar três classes de problemas relacionadas à administração da informação: modelagem e pesquisa, extração e integração de dados e, por último, construção e reestruturação de web sites.

2. Representação de dados para tarefas de Banco de Dados/Web

Para realizar as tarefas supra citadas, deve-se utilizar um método de modelagem do domínio (WWW). É necessário modelar a própria Web, estrutura dos sites, estrutura interna das páginas e o conteúdo dos sites em granulações menores que uma página.

a) Modelos de dados gráficos: neste modelo, os nodos representam as páginas na web, e os arcos, suas ligações. A modelagem destes dados está baseada em um gráfico rotulado, no qual juntamente foram desenvolvidas várias linguagens de consulta que permitem formular queries regulares sobre estes gráficos.

b) Modelos de dados semi-estruturados: em muitos casos, a estrutura dos dados na web é irregular, dificultando a modelagem, inclusive de sites, que não possuem um esquema fixo. Dados semi-estruturados referem-se aqueles cujo esquema não é fornecido de antemão podendo inclusive estar implícito nos próprios dados, além do esquema ser relativamente grande e podendo freqüentemente mudar. Além disso, o esquema é descritivo e não prescritivo e, os dados não são fortemente tipados. Neste modelo, não existem restrições para arcos que saem de um nodo do gráfico ou para tipos de valores dos atributos.

Devido às características dos dados semi-estruturados, é importante a capacidade de pesquisa no esquema, sendo esta suportada por linguagens de consultas de dados semi-estruturados por meio de variáveis de arcos, presas aos rótulos.

Existem outras características dos modelos de dados da web, como construções específicas na sua representação. Deve-se representar as páginas, bem como suas conexões, que podem ser para a mesma página ou então para outros sites. Estes modelos diferem de outros, na habilidade de modelar a ordem dos elementos do banco de dados, na modelagem de estruturas de dados aninhados e apoios para tipos de coleções de elementos, como por exemplo, matrizes. Todos os modelos aqui mencionados representam somente estruturas estáticas. A modelagem de estruturas de sites, não considera páginas dinâmicas criadas por interações de usuários. Um aspecto importante das linguagens de pesquisa na web é a necessidade de estruturas complexas como resultado de uma pesquisa. Por isso, suas expressões de consultas contém componentes estruturados além do tradicional componente de filtragem dos dados.

3. Modelando e pesquisando na Web

Sendo a web vista como um grande banco de dados, semelhante a um gráfico, é natural que as pesquisas realizadas vão além do paradigma básico de busca das informações atualmente suportadas pelos mecanismos de pesquisas hoje existentes. Deve-se, entretanto, levar em consideração a estrutura interna das páginas, bem como a estrutura externa das ligações que as interconectam.

As primeiras ferramentas de pesquisa na internet baseavam-se na busca por palavras ou frases presentes em documentos descobertos pelos "*web crawlers"*. Esforços tem sido realizados para superar este paradigma, utilizando largamente as estruturas de ligações da web nas consultas, a fim de melhorar o desempenho de busca na rede.

Devido à presença destes dados, chamados de semi-estruturados, existe a necessidade de modificar ou remodelar as linguagens tradicionais, pois estas não são dirigidas a dados deste tipo. Entretando, várias linguagens estão sendo propostas para consultas sobre estes dados. Algumas são extensões da linguagem OQL (versão da SQL para utilizar em bases de dados OO), como a LOREL [AQM97], a UnQL [BDHS96] e a StruQL [FFLS97], que também utilizam gráficos rotulados como um modelo de dados flexíveis, enfatizando a característica de consultar o esquema dos dados, tratando suas irregularidades, como por exemplo, a falta ou repetição de campos e registros heterogêneos. Convém ressaltar que estas linguagens não foram desenvolvidas especificamente para a web.

Outras propostas, abordam linguagens de consulta para WWW, como a W3QL [KS95] e a WebSQL [MMM97], que podem combinar condições de padrões de texto que aparecem no conteúdo das páginas com os padrões de gráficos que descrevem as estruturas de ligações das páginas. A WebSQL tem como proposta modelar a web como sendo um banco de dados relacional, composto por sua vez de duas relações virtuais: *document* e *anchor*. A primeira relação apresenta uma tupla para cada documento na web, e a segunda, apresenta uma tupla para cada âncora em cada documento da web.

4. Integração de Informações

A web possui um número crescente de fontes de informações que podem ser vistas como *containers* de conjuntos de tuplas, que estão contidas em páginas HTML ou "escondidas" atrás de interfaces de formulários. Programas chamados "*wrappers"* podem ser escritos, dando a impressão que o web site está servindo a conjuntos de tuplas. Chama-se *web source* a combinação entre o site e o wrapper a ele associado. Sistemas de integração de dados na web tem por objetivo responder consultas que requeiram extração e combinação de dados a

partir de múltiplas fontes dispersas na web. Vários são os problemas detectados na construção destes sistemas, sendo semelhantes aos encontrados em banco de dados heterogêneos, como grande número de web sources em constantes evoluções, escassez de metadados sobre aspectos das fontes e um maior grau de autonomias destas fontes. Na construção destes sistemas, deve-se assumir duas abordagens distintas: A primeira refere-se ao fato de que os dados são movidos para um único local e as consultas realizadas sobre eles, sendo que estes devem ser constantemente atualizados, obtendo, entretanto, consultas mais velozes. Na segunda, os dados permanecem em seus locais originais, sendo as consultas realizadas localmente, garantindo a atualização dos dados, embora diminuindo a velocidade. Esta abordagem é mais apropriada devido ao número de fontes que são numerosas e autônomas, exigindo com isso métodos sofisticados para otimização e execução das consultas.

Há duas características que diferem estes sistemas de banco de dados tradicionais: Primeiro, os dados não são obtidos diretamente, mas sim através do uso de wrappers. Estes têm a tarefa de traduzir os dados do web site para uma forma que possa ser processada pelo sistema externo que realiza a pesquisa. Segundo, o usuário não especifica pesquisas de acordo com o esquema, segundo o qual os dados estão armazenados. Ao invés disso, é utilizado um esquema mediador que são específicos para cada aplicação de integração, evitando assim, que o usuário tenha que conhecer o esquema de cada uma das fontes. O sistema deve possuir informações sobre as fontes, uma vez que tem que traduzir as pesquisas do usuário para os esquemas das fontes, precisando com isso de dados como por exemplo, conteúdos, atributos, restrições, capacidades de processamento de pesquisas, etc. Para resolver o problema da especificação das fontes, bem como a tradução das queries, duas abordagens são consideradas: *Global as view* (GAV) [GMPQ97, PAGM96, ACPS96, HKWY97, FRV96] e *Local as view* (LAV) [LRO96, KW96, DG97, FW97]. Outro problema encontrado é que uma fonte não aborda de forma completa o assunto em questão, constituindo-se, tal conhecimento, de grande importância ao sistema de integração. Com relação à capacidade de processamento de pesquisas, o sistema integrador deve conhecer as limitações de cada site, pois as restrições devem-se ao fato de como os dados estão estruturados, pois podem estar em sistemas tradicionais de banco de dados, ou simplesmente em arquivos estruturados, por exemplo.

Outra informação necessária diz respeito a otimização das pesquisas, onde há escassez de dados estatísticos referentes aos que se encontram nas fontes, resultando pouca informação para avaliação de custos e planos de execução das consultas. Há ainda a questão da construção dos wrappers. Conforme descrito, sua função é extrair dados de um site para permitir que estes possam ser manipulados pelo sistema integrador. O problema é que esta extração é normalmente realizada sobre páginas HTML, onde os dados estão em linguagem natural ou sob a forma de gráficos. Para contornar tal questão, pode-se utilizar uma técnica que prevê o desenvolvimento de gramáticas especializadas, para especificar como os dados devem ser colocados nas páginas. Outra técnica seria o desenvolvimento de mecanismos de aprendizagem indutiva para "wrappers" de aprendizagem automática. Um algoritmo seria usado para gerar uma gramática de extração de dados para as páginas subseqüentes, a partir de páginas cujos dados são rotulados e então, usados como exemplo. Quanto maior o número de páginas-exemplo, mais acurada será a gramática, onde o desafio é descobrir linguagens que possam ser aprendidas com um número menor de exemplos. Por último, existe a questão de encontrar-se objetos através das diversas fontes, pois estas, utilizando suas próprias convenções de nomes de objetos, acabam por dificultar o discernimento de objetos que se referem à mesma entidade no mundo. Vários sistemas tratam tal dificuldade através do uso de heurísticas de domínio específico e técnicas de recuperação de informações.

5. Construção e reestruturação de Web Sites

Do mesmo modo que os web sites provêm informações, pode-se também aplicar técnicas de Sistemas de Banco de Dados no processo de construção e manutenção de tais sites. Pode-se distinguir duas classes gerais de tarefas na execução destes processos: uma que os web sites são criados a partir de dados subjacentes e outra na qual são criados pela reestruturação de sites já existentes. Entre as tarefas que um construtor deve enfrentar destacam-se o processo de escolher os dados que serão apresentados no site, projetar a estrutura do site e projetar a apresentação gráfica das páginas. Sistemas foram desenvolvidos objetivando a utilização de técnicas de banco de dados ao problema da criação de sites, fornecendo com isso uma representação da estrutura deste, sendo a estrutura definida como uma visão dos dados existentes. Os sistemas diferem, entretanto, no modelo de dados, linguagem de pesquisa e na existência ou não de uma representação lógica do site ao invés de representação simples em HTML. A representação declarativa da estrutura do site permite criar versões múltiplas do mesmo, pois são definidas por uma query e não proceduralmente por um programa. A reorganização destas páginas requer simplesmente que seja reescrita a query de definição do site ao invés de reescrever um conjunto de programas para tal tarefa. É possível também reforçar as restrições de integridade e pesquisa, além de fornecer plataformas para o desenvolvimento de algoritmos de otimização para gerenciamento em tempo de execução de sites com grande quantidade de dados. Observa-se ainda que, a utilização deste paradigma para construção de páginas, facilita as tarefas de consulta, bem como permite melhor integração de dados a partir de fontes múltiplas na web. O sistema acessa um conjunto de dados servidos no site, podendo estes estarem armazenados em banco de dados, arquivos estruturados ou em sites existentes. Os dados são então representados em algum modelo de dados e o sistema fornece uma interface padrão para os mesmos. A criação é feita escrevendo-se uma expressão declarativa que represente a estrutura do mesmo, utilizando-se a linguagem específica do sistema. Para que o site torne-se navegável, o sistema deve conter um método para traduzir a estrutura lógica num conjunto de arquivos HTML.

6. Conclusões

Várias são as diferenças entre a World Wide Web e um Banco de Dados. Na web não há estruturas uniformes, modelos de dados, restrições de integridade, transações, nem mesmo uma linguagem padrão para consultas. Por estas razões, vários são os trabalhos desenvolvidos na área de banco de dados com o objetivo de possibilitar o tratamento de dados semi-estruturados. Entretanto, as abstrações desenvolvidas pela comunidade desta área, podem tornar-se a chave para resolver tais problemas com relação à complexidade da web, permitindo desta forma que a mesma forneça serviços mais valiosos.

A visão da web como um grande site, é de vital importância, não sendo apenas um banco de dados, mas sim um sistema de informações construído em volta de vários bancos de dados. Neste contexto, um web site possui muitas semelhanças com sistemas de informações convencionais. A construção de sites, de acordo com estes princípios, acarretará mudanças na realização de pesquisas, bem como na integração dos dados oriundos de diversas fontes. Várias são as tendências que influenciarão o uso da tecnologia de banco de dados para a World Wide Web. A primeira é o XML [GMW98] que apesar de reduzir a necessidade do uso de wrappers, convertendo dados legíveis por seres humanos para dados legíveis por máquinas, ainda assim, permanecerão os desafios de integração semântica de dados vindos da web. A segunda tendência é o constante crescimento da chamada *web oculta*. Esta, caracteriza-se pela criação de páginas geradas por programas de interação com usuários e, portanto, não passíveis de indexação por *web crawlers*. Muitas são as pesquisas ainda a serem desenvolvidas nessa área, envolvendo a combinação de princípios de consulta para fontes de dados estruturados e não estruturados na World Wide Web.

7. Referências Bibliográficas

[ACPS96] S. Adali, K. Candan, Y. Papakonstantinou, and V.S. Subrahmanian. **Query caching and optimization in distributed mediator systems**. In Proc. of ACM SIGMOD Conf. on Management of Data, Montreal, Canada, 1996.

[AQM97] Serge Abiteboul, Dallan Quass, Jason McHugh, Jennifer Widom, and Janet Wiener. **The Lorel query language for semistructured data**. International Journal on Digital Libraries, 1(1):68-88, April 1997.

[BDHS96] P. Buneman, S. Davidson, G. Hillebrand, and D. Suciu. **A query language and optimization techniques for unstructured data**. In Proc. of ACM SIGMOD Conf. on Management of Data, pages 505-516, Montreal, Canada, 1996.

[DG97] Oliver M. Duschka and Michael R. Genesereth. **Answering recursive queries using views**. In Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), Tucson, Arizona., 1997.

[FFLS97] Mary Fernandez, Daniela Florescu, Alon Levy, and Dan Suciu. **A query language for a web-site management system**. SIGMOD Record, 26(3):4-11, September 1997.

[FLO98] Florescu, Daniela. Levy, Alon. Mendelzon, Alberto. **Database Techniques for the World Wide Web: A Survey**. SIGMOD Record, Setembro de 1998.

[FRV96] Daniela Florescu, Louiqa Raschid, and Patrick Valduriez. **A methodology for query reformulation in cis using semantic knowledge**. Int. Journal of Intelligent & Cooperative Information Systems, special issue on Formal Methods in Cooperative Information Systems, 5(4), 1996.

[FW97] M. Friedman and D. Weld. **Efficient execution of information gathering plans**. In Proceedings of the International Joint Conference on Artificial Intelligence, Nagoya, Japan, 1997.

[GMPQ97] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. **The TSIMMIS project: Integration of heterogenous information sources**, March 1997.

[GMW98] R. Goldman, J. McHugh, and J. Widom. **Lore: A database management system for XML**. Presentation, Stanford University Database Group, 1998.

[HDN98] Heuser, Carlos Alberto. Dorneles, Carina. Noronha, Marilene. **Ligando a Tecnologia de Bando de Dados com a Gestão de Documentos**. UFRGS, 1998.

[HKWY97] Laura Haas, Donald Kossmann, Edward Wimmers, and Jun Yang. **Optimizing queries across diverse data sources**. In Proc. of the Int. Conf. on Very Large Data Bases (VLDB), Athens, Greece, 1997.

[KS95] D. Konopnicki and O. Shmueli. **W3QS: A query system for the World Wide Web**. In Proc. of the Int. Conf. on Very Large DataBases (VLDB), pages 54-65, Zurich, Switzerland, 1995.

[KW96] Chung T. Kwok and Daniel S. Weld. **Planning to gather information**. In Proceedings of the AAAI Thirteenth National Conference on Artificial Intelligence, 1996.

[LRO96] Alon Y. Levy, Anand Rajaraman, and Joann J. Ordille. **Querying heterogeneous information sources using source descriptions**. In Proc. of the Int. Conf. on Very Large Data Bases (VLDB), Bombay, India, 1996.

[LSS96] Laks V. S. Lakshmanan, Fereidoon Sadri, and Iyer N. Subramanian. **A declarative language for querying and restructuring the Web**. In Proc. of 6th. International Workshop on Research Issues in Data Engineering, RIDE '96, New Orleans, February 1996.

[MMM97] Mendelzon, G. Mihaila, and T. Milo. **Querying the world wide web**. International Journal on Digital Libraries, 1(1):54-67, April 1997.

[PAGM96] Y. Papakonstantinou, S. Abiteboul, and H. Garcia-Molina. **Object fusion in mediator systems**. In Proc. of the Int. Conf. on Very Large DataBases (VLDB), Bombay, India, 1996.