Aspectos semânticos em um sistema de integração de informações na Web

Rosalie Barreto Belian[1], Ana Carolina Salgado[1]
[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
e-mail: {rbb, acs}@cin.ufpe.br

Resumo

Sistemas de integração de informações têm sido propostos com o objetivo de integrar informações de fontes de dados múltiplas e heterogêneas. Estes sistemas têm como desafio resolver a heterogeneidade da informação a fim de disponibilizar uma visão concisa e uniforme de dados distribuídos, abstraindo suas diferenças sintáticas, estruturais e semânticas. A integração de dados constitui uma das etapas necessárias para conseguir a completa interoperabilidade entre aplicações pretendida no cenário da Web Semântica. Por este motivo, conceitos e técnicas da Web Semântica têm sido assimilados no desenvolvimento de sistemas de integração de dados na Web que consideram a natureza semântica da informação. Este artigo propõe a adoção de conceitos tais como metadados, contextos e ontologias, em um sistema de integração de informações baseado em mediação. Estes conceitos serão utilizados para identificar correspondências e resolver conflitos semânticos entre informações de diversas fontes de dados heterogêneas na Web. É também proposta a inclusão de um processo que incorpora tratamento semântico em um processo de integração de informações estrutural e sintático existente.

Abstract

Information integration systems have been proposed with the goal of integrating information from multiple and heterogeneous data sources. These systems have the main challenge of resolving information heterogeneity in order to offer a concise and uniform view of the distributed data, abstracting out their syntactic, structural and semantic diversities. Data integration fits in the Semantic Web scenario which demands for complete application interoperability. Thus, Web Semantic issues are been used in the development of data integration systems based on the information semantic nature. This paper proposes the inclusion of semantic concepts, such as metadata, contexts and ontologies, in a mediator-based information integration system. Such issues will be used to identify correspondences and to solve semantic conflicts among information from diverse heterogeneous web data sources. We also propose an information integration process, which incorporates semantic issues into an existent structural and syntactic information integration process.

1. Introdução

A construção de sistemas com o objetivo de integrar dados provenientes de múltiplas fontes distribuídas na Web tem como principais desafios resolver a heterogeneidade da informação e apresentá-la aos usuários e aplicações de forma concisa e uniforme, abstraindo suas diferenças sintáticas, estruturais e semânticas. Um dos maiores problemas enfrentados no desenvolvimento destes sistemas consiste em resolver a heterogeneidade semântica de objetos encontrados nas fontes de dados e uniformizá-los, possibilitando então a utilização de mecanismos para tratamento de suas diversidades estruturais e sintáticas [1].

Sistemas de integração de informações na Web compõem o cenário da Web semântica [2] constituindo um dos pré-requisitos para a completa interoperabilidade entre aplicações desta área. Neste sentido, conceitos da Web semântica têm sido assimilados no desenvolvimento de sistemas de integração de informações na Web. Conceitos como ontologias, metadados e contextos têm sido

empregados com o objetivo de tratar semanticamente a informação em sistemas de integração de dados [3,4,5,6] com os papéis apresentados a seguir.

- Uma ontologia, como definida em [7], "é uma especificação explícita de uma conceitualização." Uma ontologia de um dado domínio de conhecimento oferece um vocabulário terminológico de referência que pode ser utilizado na resolução de conflitos semânticos entre conceitos e termos utilizados nas fontes de dados distribuídas [1,7].
- Metadados [1], geralmente definidos como "dados sobre dados", podem ser utilizados para descrever significado, conteúdo, organização ou objetivo dos dados. Em um sistema de integração de informações, metadados têm o papel fundamental de fornecer, por exemplo, informação relevante sobre as fontes de dados para a integração de esquemas e conteúdo.
- Um contexto "contém metadados relacionados ao seu significado, propriedades (tais como fonte, qualidade, e precisão), e organização" [4,5]. Em um sistema de integração de informações, contextos podem conter descrições sobre a natureza estrutural, organizacional e semântica das fontes de dados distribuídas. Contextos são considerados ferramentas eficazes no tratamento da heterogeneidade da informação [5].

Sistemas de integração de dados baseados em mediação apresentam um esquema integrado com o propósito de compatibilizar características e informações relevantes para seus usuários com a capacidade de resposta de fontes distribuídas de dados. A resolução da heterogeneidade semântica, neste cenário, considera o emprego de ontologias de domínio como ferramentas que possibilitam automatizar o processo de esclarecimento terminológico entre as fontes de dados. Em um sistema baseado em mediação, a organização e coleta de metadados precisam ser consistentes o bastante para subsidiar o processo de integração tornando-o independente de características estruturais e de representação das fontes de dados. Finalmente, o processamento de informação sensitiva ao contexto em um sistema de mediação, respeitando características individuais de fontes de dados heterogêneas e autônomas na Web, permitem a formulação de consultas mais expressivas na geração do mediador, bem como a produção de resultados mais precisos na execução das consultas do usuário.

Este trabalho tem como objetivo apresentar aspectos semânticos pertinentes à integração de informações na WEB identificados no processo de especificação do sistema Integra. O Integra é um sistema para integração de informações distribuídas em fontes de dados na WEB [8]. O Integra possui uma arquitetura baseada em mediação que adota a abordagem GAV (*Global as View*) [9] na definição de mapeamentos entre o esquema de mediação e os esquemas das fontes de dados.

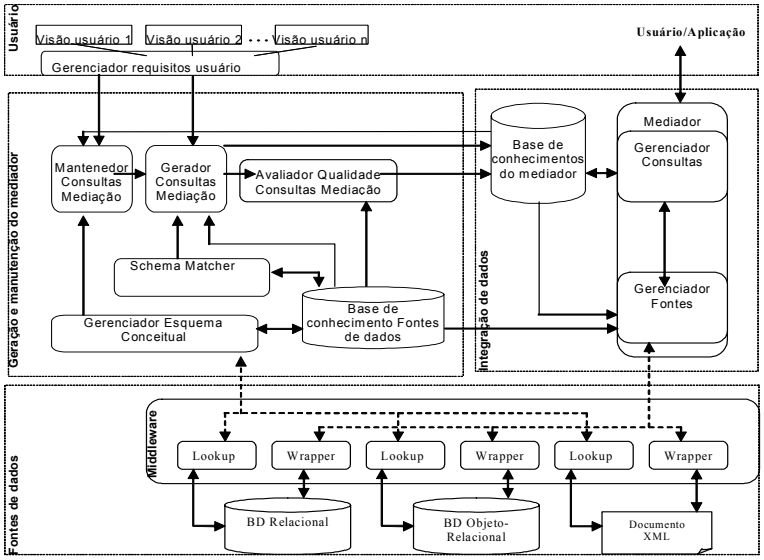

Figura 1. Arquitetura do Sistema Integra [8]

O sistema Integra utiliza XML como modelo comum para intercâmbio de dados e XML Schema como representação padrão para o esquema de mediação e esquemas das fontes de dados [10]. A arquitetura original do sistema Integra pode ser visualizada na Figura 1.

Nesta arquitetura, o módulo em que se encontram as **fontes de dados** produz as informações necessárias para a geração e manutenção do mediador, como os esquemas destas fontes que são coletados pelos módulos *Lookup*. Os *Wrappers* por sua vez, são responsáveis por traduzir as sub-consultas para o formato particular de cada fonte de dados e devolver seus resultados para o módulo de integração de dados. O módulo de **integração de dados** é responsável pela reestruturação e integração dos dados provenientes das fontes de dados autônomas apresentando para usuários e aplicações uma visão XML integrada dos dados distribuídos. Outros componentes deste módulo são utilizados na otimização do tempo de resposta das consultas do usuário.

O módulo de **geração e manutenção do mediador** processa as informações dos esquemas das fontes de dados gerando e atualizando as consultas de mediação, mantendo também a consistência dos mapeamentos entre elementos de mediação e elementos das fontes de dados que serão utilizados na execução das consultas do usuário no módulo de integração de dados. O módulo do **usuário** compreende os componentes responsáveis pela configuração e gerenciamento dos requisitos do usuário.

É importante ressaltar que o sistema Integra foi originalmente proposto para resolver apenas aspectos sintáticos e estruturais na integração de informações. Neste trabalho, sua arquitetura foi estendida de forma a incluir aspectos para tratamento semântico da informação. A arquitetura estendida utiliza os conceitos de ontologias, metadados e contextos na resolução da heterogeneidade da informação.

Este artigo está organizado como descrito a seguir. Na seção 2 é discutido o emprego dos conceitos para tratamento semântico da informação no Integra. Na seção 3 é apresentada a visão estendida do Integra descrevendo o processo de integração de informações resultante, e na seção 4 são apresentadas algumas conclusões e trabalhos futuros.

2. Integração semântica de informações

Em um processo de integração de informações a resolução de conflitos estruturais e sintáticos entre objetos deve se dar apenas após o estabelecimento da sua similaridade semântica. O estabelecimento da similaridade entre objetos baseada em princípios puramente esquemáticos e estruturais foi discutida na literatura e considerada ineficiente para determinar a integração destes objetos [6]. Nos sistemas que utilizaram esta abordagem o processo de integração se baseava no conhecimento prévio da semântica dos objetos para integração. Neste caso, a integração dos dados ocorria com base nos rótulos ou identificadores de entidades e atributos das fontes de dados distribuídas. Estes sistemas contribuíram fortemente para o desenvolvimento e maturação dos aspectos envolvidos no tratamento estrutural e sintático, permitindo o desenvolvimento posterior de sistemas com base em processos de integração mais realistas, que incorporam o tratamento semântico da informação.

No sistema Integra uma ontologia de domínio está sendo utilizada com o objetivo de permitir a interpretação semântica dos conceitos encontrados nas fontes de dados. Conceitos semanticamente similares devem ser identificados nas fontes de dados e posteriormente integrados. A informação semântica neste processo é necessária para identificar o significado correto do termo e então proceder à integração de esquemas e conteúdo a que o Integra se propõe. Neste contexto, a ontologia de domínio estabelece o vocabulário de referência que descreve conceitos, termos e relacionamentos do domínio de conhecimento no qual atua o sistema. Um processo para estabelecimento de correspondências semânticas entre termos das fontes de dados e a ontologia deve então ser executado sendo suportado por um conjunto de metadados rico o suficiente para descrever informações estruturais, sintáticas e semânticas. No Integra, metadados estão sendo utilizados para descrever: características das fontes de dados, termos da ontologia, conceitos das

fontes de dados, conceitos de mediação, mapeamentos entre conceitos de mediação e conceitos das fontes de dados, e consultas de mediação.

Uma outra característica do Integra que merece ser destacada é a forma de organização de metadados do sistema agrupados através de contextos. Contextos são utilizados no Integra agrupando metadados relacionados às fontes de dados e seus elementos, e no nível de mediação, associados aos conceitos do usuário. Desta maneira, os metadados utilizados no processo de integração no sistema Integra são contextualizados de acordo com o elemento tratado: fonte de dados ou mediador. Metadados relacionados às fontes de dados são utilizados na geração de seus esquemas XML para o sistema, bem como na captura da semântica mais adequada para seus elementos no processo de estabelecimento de correspondências semânticas com a ontologia de domínio. Da mesma forma, o conteúdo das respostas das sub-consultas processadas nas fontes de dados devem estar de acordo com o contexto de cada fonte em particular. Neste caso, uma transformação ou adequação do contexto da fonte para o contexto de mediação deve ser realizada possibilitando a integração destes dados resolvendo conflitos existentes no seu conteúdo. No Integra, um contexto pode manter informações tais como: grau de similaridade e relacionamentos semânticos, mapeamentos entre conceitos de mediação e suas entidades relacionadas nas fontes de dados remotas, tipos de dados e restrições, precisão, entre outros.

3. O Sistema Integra estendido

O processo de integração de informações está baseado em duas dimensões principais: um processo de aquisição de informações estruturais e semânticas sobre as fontes de dados para geração do mediador, e uma segunda fase considerando o processamento de consultas do usuário e a produção de seus resultados integrados no sistema, como pode ser visualizado na arquitetura (Figura 1). O processo para geração do mediador semanticamente alinhado considera as seguintes etapas [11]:

- **Extração e tradução dos esquemas das fontes de dados**: nesta etapa os esquemas das fontes de dados são coletados e convertidos para o modelo comum de dados.
- **Comparação de elementos das fontes de dados e conceitos da ontologia**: esta tarefa consiste em identificar a similaridade semântica entre elementos dos esquemas das fontes de dados e conceitos e termos da ontologia de domínio. Esta etapa produz um conjunto de correspondências semânticas que esclarecem o significado de cada elemento encontrado nas fontes de dados. Entidades, relacionamentos e atributos das fontes de dados são submetidos ao processo de comparação com a ontologia de domínio.
- **Agrupamento de conceitos semanticamente similares**: esta tarefa unifica conceitos semanticamente similares produzindo um conjunto de "clusters", que serão utilizados pelo usuário na definição de seus requisitos. Nesta etapa é iniciada a formatação dos contextos do mediador associados a cada conceito "cluster", no entanto sua configuração é apenas concluída na fase de geração do esquema de mediação.
- **Definição de requisitos do usuário**: esta atividade considera o universo de conceitos gerados na etapa anterior e que foram semanticamente interpretados através da ontologia. O usuário deve selecionar neste universo os conceitos que são relevantes para sua aplicação.
- **Geração do esquema de mediação**: nesta etapa os metadados mantidos nos contextos de mediação e das fontes de dados são utilizados na geração do esquema de mediação.

Para suportar o processo descrito anteriormente, na arquitetura estendida foi criado o módulo da ontologia, que contém os elementos ontológicos do domínio de conhecimento de atuação do sistema e que fornece informação semântica aos outros módulos da arquitetura. Além deste módulo foram criados os módulos para estabelecimento da **correspondência semântica de entidades** das fontes de dados e conceitos da ontologia (*semantic entity matcher*), **unificação semântica** de conceitos similares (*semantic entity unifier*) e **geração do esquema de mediação** (*mediator schema generator*) [11] detalhados a seguir.

Correspondência semântica de entidades: este módulo compara entidades, atributos e relacionamentos obtidos dos esquemas das fontes de dados buscando o seu esclarecimento semântico através de conceitos e termos da ontologia. Um processo sintático de comparação (com base na grafia) entre elementos dos esquemas das fontes e a ontologia é realizado produzindo correspondências semânticas entre estes. Antes da realização do processo de comparação um pré-tratamento é realizado com o objetivo de realizar uma normalização nos nomes de elementos dos esquemas das fontes (hífens, gênero, número, grau, etc.).

Unificação semântica: este módulo busca unificar entidades e atributos similares utilizando os relacionamentos semânticos associados a conceitos das fontes de dados gerados pelo módulo anterior. Os conceitos similares são agrupados produzindo um "cluster". Um conceito é considerado similar a outro conceito se eles possuem um grau significativo de similaridade semântica. O processo de unificação produz uma coleção de conceitos ("clusters") que serão utilizados pelo usuário na configuração de seus requisitos relevantes. A informação de mapeamento entre os conceitos do mediador e das fontes de dados é gerada neste momento, mas a configuração do contexto do mediador será completada com a definição dos requisitos do usuário.

Geração do esquema de mediação: de acordo com os requisitos do usuário, este módulo completa a informação necessária para definir o esquema do mediador. Os mapeamentos entre conceitos das fontes de dados e do mediador serão utilizados na geração das assertivas de correspondência utilizadas na geração das consultas do mediador.

Os módulos descritos foram inseridos na arquitetura original do Integra aproveitando o processo de tratamento sintático e estrutural existente. O processo de geração do mediador finalmente produz um mediador semanticamente alinhado que deverá ser utilizado pelo processo original no módulo de integração de dados (arquitetura).

4. Conclusões e trabalhos futuros

Este trabalho apresentou a adaptação do sistema Integra, que originalmente foi especificado com base em aspectos sintáticos e estruturais, para incorporar também o tratamento semântico da informação. Sua maior contribuição consiste na proposta de enriquecimento de um processo de integração de informações já amadurecido com mecanismos para tratamento semântico.

O processo de integração discutido neste trabalho está baseado nos conceitos de metadados, contextos e ontologias. Estes conceitos têm sido explorados em diversos sistemas de integração de informações devido ao seu grande potencial na resolução de conflitos semânticos da informação. O sistema Integra foi revisado para incluir estes conceitos tendo a sua arquitetura original sido redesenhada com o objetivo de produzir um esquema de mediação composto por informação semanticamente alinhada com a ontologia de domínio.

O Integra foi projetado para atuar na integração de dados distribuídos na Web, considerando fontes de dados autônomas, heterogêneas e semi-estruturadas. Sistemas como o Integra, que resolvem problemas de integração de dados considerando a natureza semântica da informação, pertencem ao cenário da Web Semântica, cujas aplicações demandam por uma completa interoperabilidade entre sistemas e dados [2]. No momento, as especificações dos sub-processos para identificação de similaridade entre termos da ontologia e elementos das fontes de dados, e o agrupamento de conceitos similares, estão sendo concluídos. Estão também sendo iniciados os estudos relacionados ao tratamento semântico na integração do conteúdo obtido das fontes de dados em resposta às consultas do usuário. A validação da arquitetura proposta será realizada através da implementação de um protótipo para aplicação na área de saúde.

Referências

1. Kashyap, V., Sheth, A.: Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies. Chapter in Cooperative Information Systems: Current Trends and Directions, M. Papazoglou and G. Schlageter Editors, 1996.
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, Scientific American, 284 (5), vol. 184, no. 5, pp. 34-43, 2001.
3. Reinoso-Castillo, J., Silvescu, A., Caragea, D., Pathak, J., Honavar, V.: Information Extraction and Integration from Heterogeneous, Distributed, Autonomous Information Sources – A Federated Ontology-Driven Query-Centric Approach. In: IEEE International Conference on Information Integration and Reuse. In press, 2003.
4. Goh, C., Madnik, S., Siegel, M.: Semantic Interoperability through Context Interchange: Representing and Reasoning about Data Conflicts in Heterogeneous and Autonomous Systems. Sloan School of Management, MIT, http://citeseer.ist.psu.edu/191060.html, 1997.
5. Wache, H., Stuckenschmidt, H.: Practical Context Transformation for Information System Interoperability. In Proceedings of the 3rd International Conference on Modeling and Using Context (CONTEXT'01), Lecture Notes in AI, Springer Verlag, 2001.
6. Ouksel, A., Sheth, A.: Semantic Interoperability in Global Information Systems. A brief introduction to the research area. SIGMOD Record, Vol. 28, No.1, March 1999.
7. Gruber, T.: A Translation Approach to Portable Ontologies. Knowledge Acquisition, V.5, n.2, p.199-200, 1993.
8. Lóscio, B.: Managing the Evolution of XML-based Mediation Queries. PHD Thesis, Federal University of Pernambuco, Brazil, 2003.
9. Levy, A.: Logic-Based Techniques in Data Integration. In: J. Minker, editor Logic-based Artificial Intelligence, Kluwer Publishers, 2000.
10. Lóscio, B., Salgado, A.C., Galvão, L.: Conceptual Modeling of XML Schemas. International Conference on Conceptual Modeling ER, WIDM, 2003.
11. Belian, R., Lóscio, B., Pires, C., Salgado, A.C.: Extending an Information Integration System with Semantics. Submitted to The 20th Annual ACM Symposium on Applied Computing, Santa Fe, New Mexico, 2005.