

UM SISTEMA PARA AUXILIO AO PROJETO E INTEGRAÇÃO DE ESQUEMAS CONCEITUAIS SEMANTICOS

Jano Moreira de Souza *1

SUMARIO

Este trabalho apresenta as características gerais de um sistema para auxílio ao projeto e integração heterogênea de Bases de Dados descritas usando modelos semânticos de dados. Detalhes da implementação, conceitos e algoritmos utilizados podem ser encontrados em [Souz86]. Uma taxonomia de tipos de integração de banco de dados é fornecida graficamente. A importância da tarefa de comparação semântica dos esquemas conceituais que descrevem os banco de dados, a serem integrados, é salientada e conceitos desenvolvidos em seis outras áreas da Ciência da Computação são sugeridos para uso nessa comparação.

ABSTRACT

This paper outlines a Computer Aided Design and Integration System for databases described using different Semantic Data Models. Details of the implementation, underlying concepts and algorithms used can be found in [Souz86]. A taxonomy of integration types is presented graphically. The paper stresses the importance of the semantic comparison of the conceptual schemata describing the database to be integrated. Concepts borrowed from six others fields of the Computer Science are suggested for use in the comparison task.

1. Professor Adjunto - COPPE/UFRJ

Caixa Postal 68.511 - CEP 21945 - Rio de Janeiro - RJ - Brasil

1 - INTRODUÇÃO

Como consequência do intenso debate ocorrido na área de Banco de Dados, da década de 70 até o começo da década de 80, sobre a conveniência de um modelo de dados sobre os outros [Date74, Sib176], emergiu um consenso em torno da necessidade de suportar a coexistência de diversos modelos de dados [Smit81, Mann83, Mcle80, Melo83]. Os requisitos das aplicações e/ou a preferência dos usuários devem ser os determinantes da escolha do modelo de dados. O enfoque adotado no presente trabalho baseia-se nas seguintes premissas:

- 1) Existe uma necessidade real de utilizarem-se diferentes modelos semânticos de dados e de integrar (física e/ou logicamente) bases de dados descritas por seu intermédio;
- ii) a completa equivalência (semântica ou sintática) de construções de ambos esquemas a ser integrados são muito raras em banco de dados reais;
- iii) A integração (e principalmente a comparação) de grandes esquemas, ou de esquemas ricos em descrição semântica, manualmente é virtualmente impossível.

No estudo da integração de esquemas conceituais semânticos três aspectos foram considerados:

- 1) Uma taxonomia dos tipos de integração (ou das arquiteturas dos banco de dados após a integração);
- ii) algoritmos e conceitos úteis para a comparação e mapeamentos dos esquemas conceituais;
- iii) a arquitetura e funcionalidade necessárias a um sistema especialista que auxilie na tarefa de integração.

Os três aspectos acima foram estudados em profundidade em [Souz86]. No presente artigo seus tópicos mais relevantes serão apresentados nas seções 2, 3 e 4.

2 - UMA TAXONOMIA DOS TIPOS DE INTEGRAÇÃO

A integração de banco de dados pode tomar várias formas dependendo de diversos fatores. Embora cada tipo de integração tenha suas especificidades, há suficiente em comum entre elas para justificar um tratamento geral. Os dois fatores mais importantes do tipo de integração são: a) o fato de existir ou não uma integração física dos dados, além da integração lógica dos meta-dados (esquemas), e, b) a ocasião em que se dá a integração. Esta pode ser:

- 1) Depois que os bancos de dados já se encontram em uso. Isto é, integração de banco de dados já populados;
- ii) durante o desenvolvimento de um banco de dados de grande porte e/ou complexidade, situação em que várias pessoas ou equipes trabalham em paralelo e necessitam consolidar ('integrar'), periódicamente suas porções do projeto;
- iii) no momento da consulta. O usuário executa uma integração rápida, dinâmica, válida somente para aquela consulta.

Os dois aspectos determinantes citados acima são por sua vez função das características das aplicações que utilizam esses bancos de dados ('integrado' e a integrar). As características que consideramos mais importantes são:

- a) o padrão de utilização previsto para o banco de dados integrado - (ex. somente consultas, atualizações frequentes, atualizações locais x globais, etc...). O padrão de utilização previsto pode tornar proibitiva algumas das arquiteturas (que serão apresentadas a seguir), seja por impedir certas funcionalidades (por exemplo, atualizações em vistas) ou por tornar proibitivo o desempenho de certas transações;

- b) considerações geográficas, dado que algumas arquiteturas podem ser mais apropriadas para um banco de dados distribuído do que outras;
- c) custos da integração física, isto é, a integração física das bases de dados pode não ser compensadora por envolver custos de cópia e/ou transmissão de dados elevados;
- d) custos da perda da produção, pois certos tipos de integração podem exigir a completa paralisação do sistema existente, durante a integração, por um período inaceitável;
- e) custos ou tempo para re-escrever transações/aplicações pode ser excessivos;
- f) dificuldade em definir os mapeamentos entre os esquemas antigos e o integrado (incluindo mapeamentos inversos, para atualizações).
- g) 'propriedade dos dados', pois devido a razões de confidencialidade ou segurança, pode ser importante não efetuar a integração física dos dados;
- h) heterogeneidade do banco de dados integrado.

Um dos problemas encontrados no estudo da integração de banco de dados é a confusão existente na literatura quanto ao problema que cada pesquisador tenta resolver. Mais adiante será apresentada, em forma gráfica, uma classificação dos tipos de integração encontrados na literatura ou em sistemas comerciais. A partir dos dois bancos de dados apresentados na figura 1, seis arquiteturas para o banco de dados integrado podem ser obtidas. Elas são apresentadas nas figuras 2 até 7. O leitor estudioso da área não terá dificuldades em identificar alguns sistemas a partir dessa representação (ex. MULTIBASE [Daya82] e Multidatabase [Litw82] são exemplos de integração tipo 'Super-Vista' (fig.4)). A classificação foi desenvolvida levando em consideração oito características do banco de dados integrado, resultante, que são listadas a seguir. Uma caracterização mais precisa de cada tipo, bem como seu estudo comparativo pode ser encontrado em [Souz86].

Características Usadas Para Classificar Tipos de Integração:

- i) Esquema antigo torna-se vista;
- ii) esquema integrado é uma vista;
- iii) problemas com atualizações de transações locais;
- iv) problemas com atualizações de transações globais;
- v) integração física;
- vi) transações locais são afetadas;
- vii) transações globais são re-escritas;
- viii) presta-se a integração heterogênea.

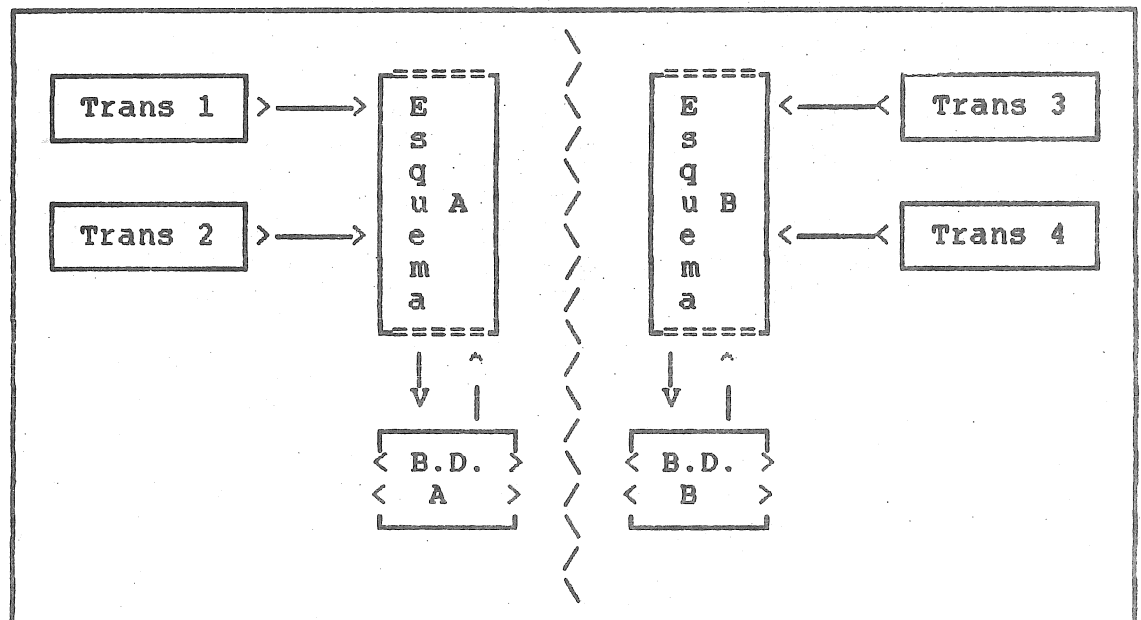


Figura 1 - Bancos de Dados Antes da Integração

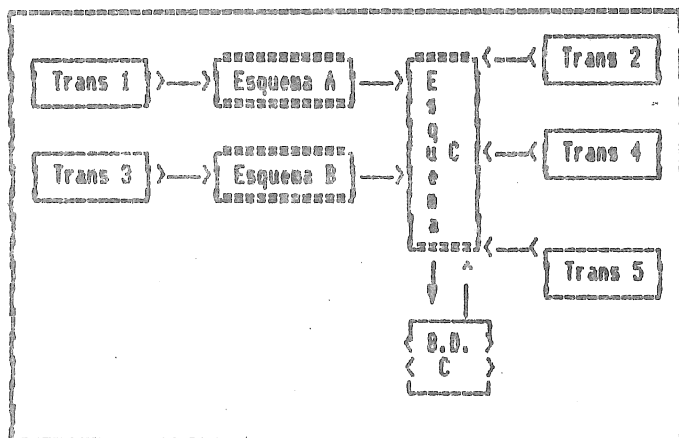


Figura 2 - Integração Total Com Sub-Esquemas

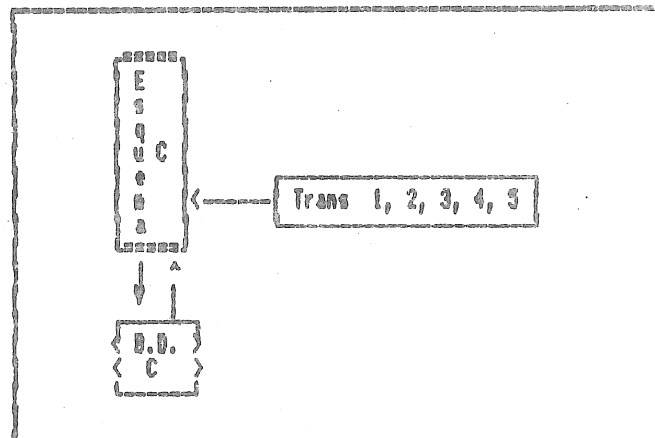


Figura 3 - Integração Total

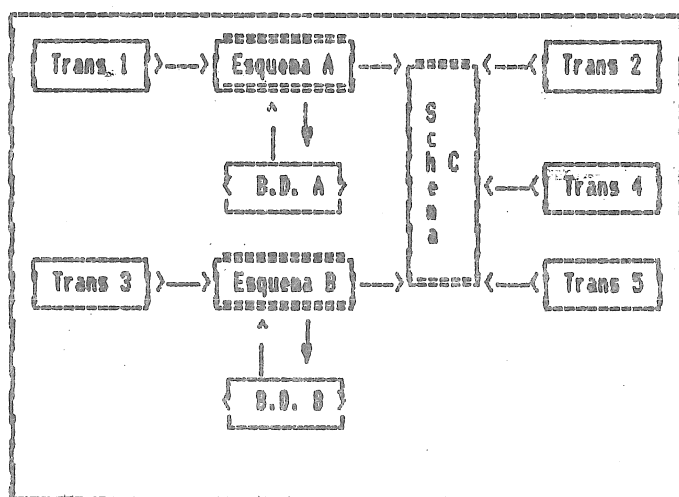


Figura 4 - Super-Vista

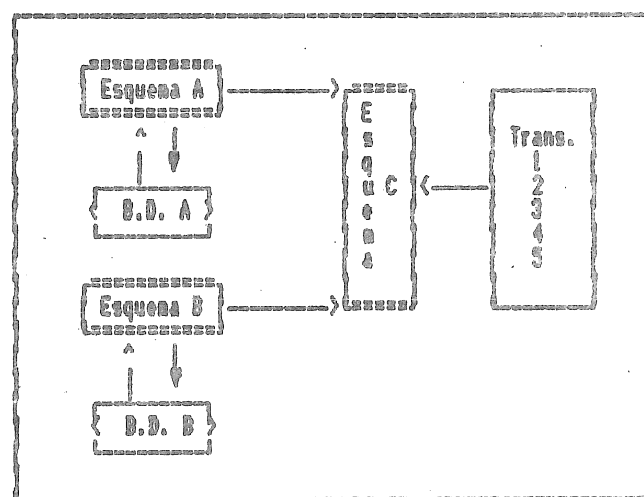


Figura 5 - Banco de Dados Distribuido com Esquema Global

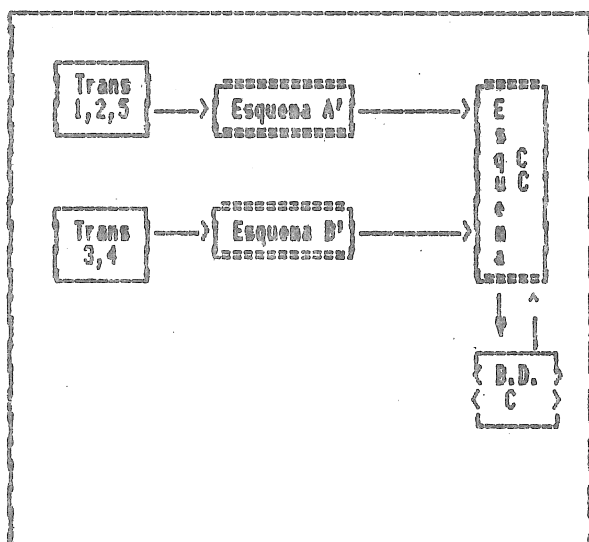


Figura 6 - Integração Total com Sub-Esquemas Modificados

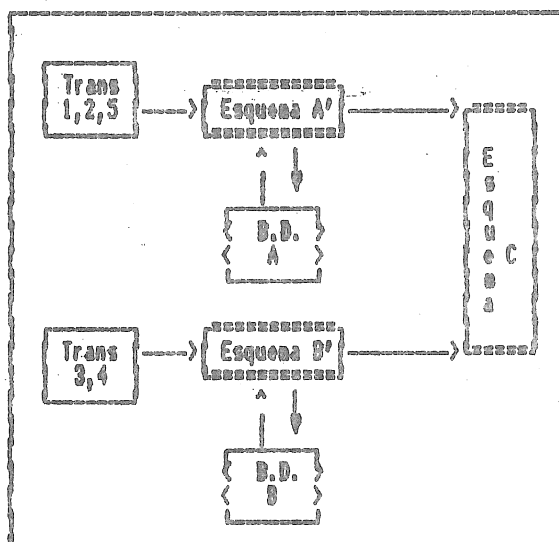


Figura 7 - Banco de Dados Distribuido com Esquema de Integração

3 - COMPARAÇÃO DE ESQUEMAS

Para grandes esquemas, ou para aqueles ricos em descrição semântica, a comparação manual torna-se praticamente impossível. O problema é agravado pelo fato de que, na prática, dificilmente se encontrará um par de construções com exatamente a mesma descrição, tanto do ponto de vista sintático (nomes por exemplo) como semântico. A solução encontrada foi utilizar o conceito de equivalência ao invés do de semelhança. O objetivo da comparação é encontrar subconjuntos de construções em um esquema *A* que pareçam o máximo com um outro subconjunto de construções do esquema *B*.

Seis conjuntos de conceitos tradicionalmente utilizados para lidar com informações imprecisas foram estudados, a saber:

- i) fuzzy Sets;
- ii) reconhecimento de Padrões;
- iii) teoria da informação;
- iv) comparação sintática Uni-dimensional;
- v) comparação usando gramáticas Bi-dimensionais, tais como gramática de grafos;
- vi) teoria de Grafos, (isomorfismo de sub-grafos aliado à um algoritmo similar à programação dinâmica).

4 - UM SISTEMA PARA AUXILIO A INTEGRAÇÃO

O sistema desenvolvido objetivou auxiliar a integração permanente de banco de dados já existentes ou em desenvolvimento (conforme mencionado na seção 1). Entretanto muitas das técnicas e algoritmos empregados ou sugeridos podem ser utilizados no momento da consulta a banco de dados fracamente acoplados [Mac180]. O sistema é totalmente interativo e possui módulos para definir, comparar, modificar e visualizar construções de diversos esquemas, bem como para verificar sua consistência e definir mapeamentos intra e inter-esquemas. O módulo de comparação utiliza algoritmos heurísticos para lidar com similaridades ao

mapeamentos intra e inter-esquemas. O módulo de comparação utiliza algoritmos heurísticos para lidar com similaridades ao invés de equivalências.

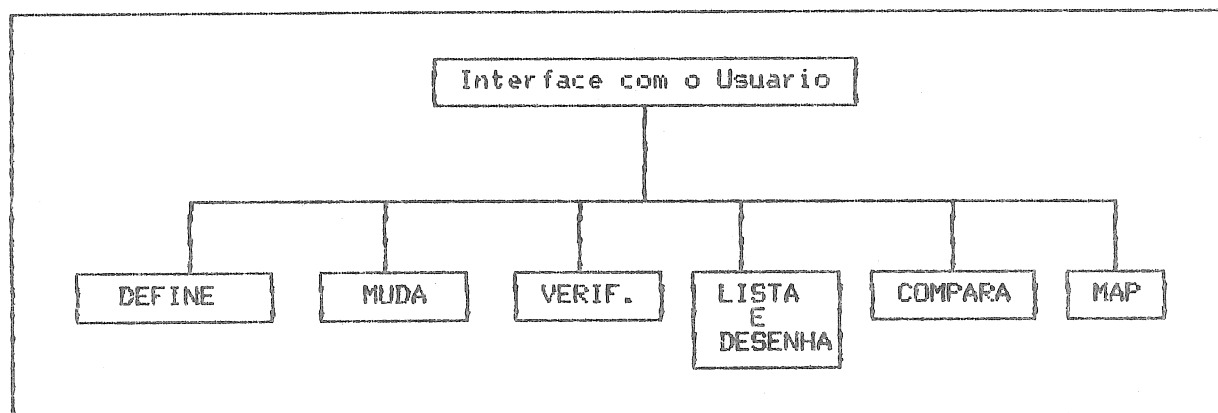


Figura 1 - Módulos do sistema

Embora o sistema possa ser utilizado para integrações heterogêneas, a adoção de uma arquitetura de quatro níveis (uma extensão direta da arquitetura de de três níveis proposta pela ANSI/SPARC [ANSI75]), juntamente com um modelo de dados canônico permite a integração elegante de banco de dados descritos por modelos de dados semânticos tais como NIAM [Verh82], CIM [Gust82] ou ACM/PCM [Brod82].

5 - CONCLUSÃO

O trabalho destaca três aspectos considerados importantes para a integração de esquemas conceituais semânticos: a) uma taxonomia dos tipos de integração, b) algoritmos e conceitos para a comparação e mapeamento e c) uma ferramenta automatizada para a comparação e integração. Os três aspectos são importantes para a integração, e tem sido muito pouco pesquisados até o momento. Devido à exigüidade de espaço para um aprofundamento dos três tópicos, a taxonomia mereceu maior destaque por ser pré-requisito para o entendimento dos outros, que serão objeto de futuros trabalhos.

REFERENCIAS

- [ANSI75] ANSI/SPARC, Study Group on Data Base Management Systems: Interim Report, ACM FDT Vol 7, No 2, New York, 1975.
- [Brod82] Brodie M.L. and Silva E., Active and Passive Component Modelling: ACM/PCM, in [29].
- [Cloc81] Clocksin, K.L. and Mellish, C., Programming in Prolog, Springer-Verlag, 1981.
- [Date74] Date, C.J., The Relational and Network Approaches: Comparison of The Application Programming Interfaces, IBM Technical Report RJ 1401(#21706), June 1974.
- [Daya82a] Dayal, U. and Hwang, H., View Definition and Generalization for Database Integration in Multibase: A System for Heterogeneous Distributed Databases. In Proc 6th Berkeley Workshop on Distr Data Managmt and Comm, Feb 1982, pp203-238
- [Gust82] Gustafson, M.R., Karlsson T. and Bubenko T.A. Jr, A Declarative Approach to Conceptual Information Modelling, in [Olle82].
- [Litw83] Litwin, W. e KabbaJ, K. Manipulation of Relational Multidatabases, INRIA Report, 1983.
- [Mann83] Mannino, M.V., A Methodology for Global Schema Design, PhD Thesis, University of Arizona, 1983.
- [McLe80] McLeod, D. and Heimbigner, D. A Federated Architecture for Data Base Systems. In AFIPS National Computer Conference, Vol 39, 1980.
- [Melo83] Melo, R.N., Pena, J.C.C. and Zakimi, B., Comparando Modelos Semanticos de Dados, Revista Brasileira de Computacao, Vol 3, No 1, pp 75-89, 1983.
- [Motr80] Motro, A. and Buneman, P., Automatically Merging Databases, Proc Compcom Conf Distributed Computing (IEEE), Fall 1980.
- [Motr81] Motro, A., Constructing Superviews, in Proc International Conference on Very Large Databases, Cannes, France, 1981.
- [Olle82] Olle, T.W. et al (eds.), Information Systems Design Methodologies: A Comparative Review, North-Holland Pub Comp, 1982.
- [Saha83] Saha, A.B., A Set Oriented Abstract Conceptual Schema Based on Logic, Internal Report, School of Information Systems, University of East Anglia, Dec 1983.
- [Schr75] Schreider, J.A., Equality, Resemblance and Order, Mir Publishers, Moscow 1975.
- [Smit81] Smith, J.M. et al MULTIBASE, Integrating Heterogeneous Distributed Databases Systems, in National Computer Conference, 1981.
- [Souz86] Souza, J. M., "Software Tools for Conceptual Schema Integration", PhD Thesis, University of East Anglia, 1986.