Pesquisa semântica aplicada a bases de dados geográficos

Mário L. O. Flecha[1]

Gerente de Apoio ao Desenvolvimento de Sistemas da Prodemge
Mestre em Administração Pública e Ciência da Computação pelo Departamento de Ciência da Computação da UFMG e Escola de Governo da Fundação João Pinheiro. Áreas de interesse em Processamento Semântico, *Software* de mediação de acesso a recursos de dados.

José Luis Braga[2]

Professor de Informática da Universidade Federal de Viçosa
Doutor em Informática pela PUC/RJ em 1990, e Mestre em Ciência da Computação pela UFMG em 1981. Até julho de 99 está como pesquisador visitante na *University of Flórida, Gainsville*, trabalhando com Tecnologias da Informação no *Institute of Food and Agricultural Systems*. Suas áreas de interesse são Sistemas de Suporte a Decisões e Sistemas de Informação Cooperativos.

Palavras-chave

Cooperatividade no acesso a bases de dados_ Mediador_ Base de dados_ Processamento semântico_ Sistemas especialistas_ Recuperação de informação_ Sistemas baseados em conhecimento semântico_ Sistemas de informação cooperativos.

Resumo

Este artigo descreve um software voltado para pesquisa semântica de informações o qual pode ser utilizado na pesquisa a bases de dados georreferenciados. São exemplificados casos de utilização em sistemas de banco de dados com dados

alfanuméricos e é mostrado um caso de implementação de consulta a dados alfanuméricos de uma base de localidades de todo o Brasil cuja fonte são os Correios. Em boa medida foi obtido um software independente de idiossincrasias impostas por interfaces de aplicações de usuário e especificidades de bases de dados, sendo por isto uma solução para múltiplos casos.

O elemento básico da solução foi desenvolvido como um conjunto de visões na linguagem SQL como forma de implementar e ter acesso a bancos de dados semânticos em tabelas relacionais, os quais em seu conjunto compõem um Motor de Inferência Relacional - MIR. No plano conceitual desenvolvemos uma técnica para representação da semântica de conceitos do senso comum denominada Semântica de Conteúdo - SC -, a qual permite modelar conjuntos de significados tacitamente aceitos no uso cotidiano da linguagem humana e que são expressos pela semântica cabível nos elementos léxicos de um dado domínio de uso da linguagem natural.

A semântica é derivada dos termos e da leve estrutura e organização lexical da linguagem, ao invés de uma abordagem sintática, de maior complexidade, ou da excessiva particularidade das semânticas derivadas de esquemas de bases de dados.

A indexação indireta permite a criação de bases de conhecimento semântico por meio da ligação de termos contextualizados aos identificadores de instâncias de bases de dados de fatos relacionados a um universo de discurso. O software proporciona um nível de abstração que favorece o usuário e os serviços de acesso a recursos de dados geográficos, entre outros.

Abstract

This paper reports an experienced and evaluated three layer architectural solution for co-operative database access in real use environment. It looks for mediator software

independence from idiosyncrasies of user application interfaces and specific database schemata. The cornerstone to implement this architecture is a set of relational databases and views manipulated by a Relational Inference Machine - RIM. It can be used in a variety of applications, one of them is to make it possible to get the semantics provided in the terms delivered by the users when accessing geographic alphanumerical data that supports a GIS.

Conceptually RIM is supported by a semantic model implemented as a modelling technique we have named Content Semantics – CS – which, by its turn, supports pragmatic concept ontology to be expressed as contextualized terms in well-defined semantic domains. Such concepts are derived from the lexicon and from its lightweight structure in comparison to a more formal and complex syntactical approach or to a particular semantics of schemata.

By means of "onto-components" an indirect indexation could be established whose implementation would result in semantic knowledge bases used as links between contextualized terms of ontology and instances of fact databases over a domain. Mediator software encapsulates the ontology and semantic knowledge bases in order to provide beyond itself a right level of abstraction for user applications and database service layers.

RIM was developed using SQL views built over the "onto-components" to provide semantic consultation and easy knowledge maintenance services. By means of a concise, fast and comprehensive interface the user application can request, receive and keep, as near as possible, the right level of relevant information required for the user.

There are a variety of uses in real situations for RIM and CS, some of them over increasingly large and intensively accessed databases, especially in legacy and federated environments over a network. Several interactive public database systems are daily using these mediation services and tools over millions of data records stored in large databases. Of course, lots of unexplored issues and possibilities still remain to be searched and solved.

Key-words

Co-operation in database accesses_ Mediator_ Database_ Semantic processing_ Expert systems_ Information retrieval_ Semantic knowledge-based system_

Co-operative information system.

1.Introdução

Na cooperação com usuários consulentes de bases de dados geográficos, especialmente na Internet, a pesquisa semântica pode favorecer grandemente a localização de informação relevante e útil. Dois importantes objetivos são o suporte semântico a consultas e a abstração da heterogeneidade de ambientes, facilitando assim a busca por informação relevante em grandes massas de dados organizados de formas várias e com tecnologias de armazenamento de diferentes épocas.

O uso de conhecimento por parte de softwares mediadores permite cooperar com o usuário devido à redução do nível de incerteza e irrelevância e a diminuição do fluxo de bytes inúteis na rede, incrementando a eficiência do conjunto. Softwares de mediação podem promover a integração inteligente de informação proporcionando

abstração que oculte fontes de dados heterogêneas, incluindo as de tecnologias menos recentes, originalmente não concebidas para uso em redes [WG97; Wie96; Wie92], ou pouco convencionais.

O crescimento do acesso do grande público aos recursos de dados geográficos foi acelerado pelas tecnologias de rede, comunicação e processadores, demandando serviços de integração inteligente de informação. Há necessidade de interfaces cooperativas para acesso a bases de dados geográficas, de forma a popularizar e permitir a fácil localização e manipulação das informações por usuários leigos.

Ao invés da escassez de informação o excesso e irrelevância trouxeram questões relacionadas aos ambientes de recursos de dados legados, heterogêneos e pouco convencionais.

Um suporte conceitual vem do campo de pesquisa sobre processamento semântico de informação. Foi tomado como base o trabalho de [Qui68] sobre um modelo representativo computacional da memória associativa humana e seu uso lingüístico na criação e obtenção de significado a partir de um contexto.

Buscando uma arquitetura de mediação genérica, independente de esquemas específicos de bases de dados ou de peculiaridades de aplicações de usuários, desenvolvemos uma técnica denominada Semântica de Conteúdo [Fle94; Fle97]. Ela foi projetada para ser um método simples e conciso de: extração da semântica do conteúdo dos bytes e de utilização de recursos de dados heterogêneos.

Exceto o compartilhamento da crença no papel desempenhado pela semântica lexical e alguns conceitos fundamentais, a idéia de SC difere em propósito e escopo do trabalho de [Qui68] acerca de um modelo representativo da memória semântica associativa humana e como ela funciona.

[Qui68] desenvolveu e demonstrou um modelo semântico representativo de como trabalha a memória humana de longa duração. Ele também formulou uma rica base conceitual e produziu questões relevantes para o campo de pesquisa do processamento semântico de informação. [Qui68] focou a construção e validação de uma hipótese.

Por outro lado, a principal preocupação de SC é fornecer uma forma prática para modelagem de domínios semânticos baseados na estrutura léxica e contextual da informação transportada nas palavras ou termos da linguagem natural. Os termos são essencialmente apontadores semânticos para objetos do mundo real referenciados pelas bases de dados.

A leveza estrutural do léxico, tal como definido pela Teoria Semântica de [Ull77], pode dar sustentação e tornar mais rápido o tratamento automático do significado, favorecendo a mediação cooperativa para acesso a dados semi-estruturados ou geográficos, como por exemplo os arquivos de mapas ou HTML.

O software utiliza mecanismos lingüísticos baseados no léxico para uma abordagem prática e contextual do tratamento da ambigüidade, polissemia, sinonímia e homonímia [Ull77].

Lidar com a estrutura da linguagem ao nível sintático é mais complexo do que ao nível léxico, pois implica lidar com ordenação, função e toda complexidade e regras inerentes à linguagem natural. A atomicidade e maior indivisibilidade do elemento léxico tornam-se componentes básicos de mais fácil aproveitamento e utilização na recuperação cooperativa de informação.

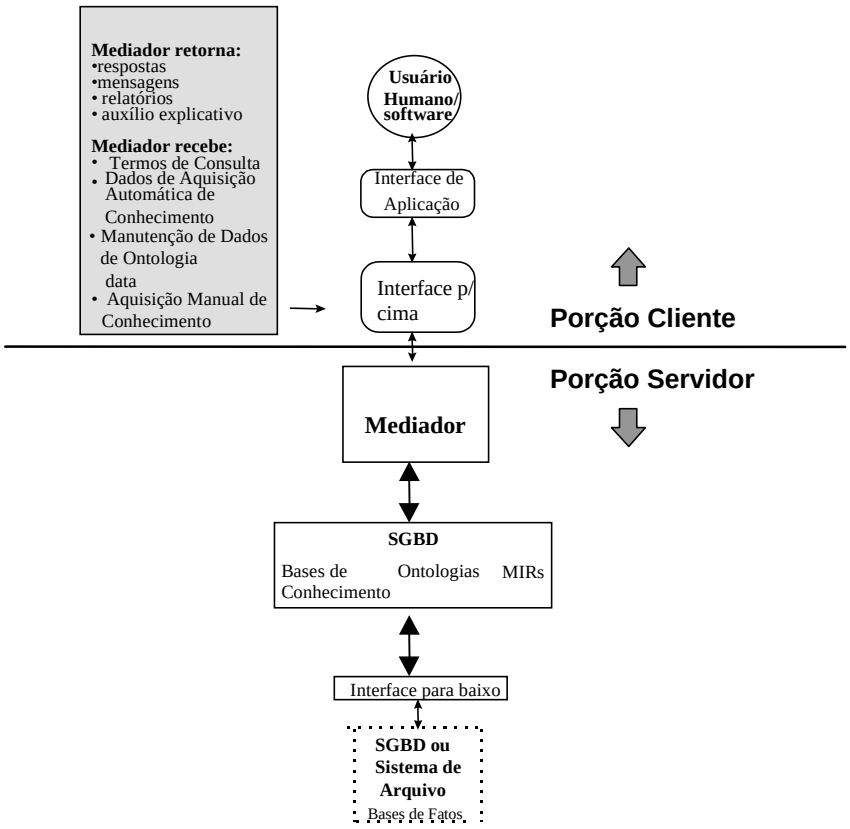
2.Uma arquitetura de software independente

Um software de consulta semântica independente não pode estar restrito à especificidade de uma aplicação ou a protocolos da camada de recursos de dados [Wie92; Wie96; WG97]. Esta dependência entretanto é ainda a situação dominante nas soluções atuais.

A camada de aplicação de usuário é quem deve desempenhar o papel de interface externa com o usuário e interna com uma camada de mediação encarregada de oferecer serviços de consulta semântica. Assim um mediador não se envolve com detalhes particulares de quaisquer aplicações, tornando-se amplamente reutilizável e útil para a camada de aplicação.

Um mediador apresenta uma interface “para cima” para receber as requisições de serviço por parte das aplicações de usuário. Certamente pode ser oferecida uma interface de usuário própria de um mediador, mas em geral ele trabalha como um componente de software utilizado internamente pelas aplicações de usuário [Fle97].

Um mediador também possui uma interface “para baixo” que o separa da camada de recursos de dados, provendo uma forma clara de requisitar serviços



[Fle97]. A figura 1 ilustra o funcionamento geral da arquitetura.

Figura 1- Arquitetura para mediadores independentes

A linguagem SQL ou qualquer outra mais específica de consulta a bases de dados geográficos provida por um SIG – Sistema de Informações Geográficas - por si não proporciona ao usuário e suas aplicações toda a flexibilidade e cognição necessárias a uma interface cooperativa e independente da camada de recursos de dados. A indexação por meio de índices diretos muitas vezes não é o bastante [Wie92; Wie96; WG97].

A linguagem SQL e sistemas gerenciadores de bancos de dados relacionais foram escolhidos como ferramentas para dotar um mediador de um mecanismo cognitivo independente de esquema de banco de dados: o MIR [Fle94; Fle97]. MIR tem como requisito ser rápido, estruturalmente leve e robusto o bastante para permitir processamento semântico cooperativo sobre grandes volumes de dados (há mais de uma implementação com dezenas de milhões de tuplas), uso concorrente ininterrupto, 7 dias por semana, 24 horas por dia em ambientes corporativos, orientados a internets , redes ou ambientes legados de tecnologias mais antigas.

Entre aplicações que utilizam este componente mediador semântico encontra-se uma base alfanumérica de dados sobre localidades de todo o Brasil, disponível 24 horas para os diversos sistemas que dela fazem uso, mediados semanticamente.

A participação de um especialista humano pode ser necessária para modificar bases de conhecimento semântico. Essa atividade requer intimidade com o léxico e a semântica de um ou mais domínios para atribuição de sinônimos e criação de contextos, por exemplo.

Cabe à interface da aplicação do usuário utilizar de forma criativa o mediador, buscando oferecer maneiras fáceis e ricas para o usuário explorar suas bases de dados. O exemplo a seguir ilustra uma sessão de consulta que usa um mediador especializado em dados postais e informações de localidade até o nível de logradouro.

Nome do Atributo	Descrição
NúmeroLocalidade	Número seqüencial de Localidade
UF	Sigla de Estado (MG, RJ, SP, ... etc.)
NomeMunicípio	Nome de Município
NomeBairro	Nome de Bairro
TipoLogradouro	Nome de Logradouro
NúmeroLogradouroDe	Numeração de
NúmeroLogradouroAté	Numeração até
CódigoCEP	Código de CEP

Tabela 1- Exemplo de atributos da base de localidade[1]

Um usuário deseja incluir um endereço de um cliente e vai obter um número de localidade para armazenar em sua base de cliente. Informa os seguintes dados de logradouro: Rua Baia, 2277, Sion, BH, MG.

A identificação do usuário é fornecida pela camada de aplicação. Os termos das perguntas da consulta serão fornecidos pelo usuário e o conector de argumentos assumido implicitamente é "E" (*default*). Aspectos da interface com o usuário não são importantes aqui, pois fazem parte da camada de aplicação.

Os dados de localidade estão armazenados na base de localidades de uma determinada forma, a qual não é conhecida pelo usuário; a maneira pela qual este digitará não coincidirá com os conteúdos da base, além de os argumentos estarem

[1] Esta é uma entidade descrita de forma simplificada para exemplificação, não representa todos os atributos que deveriam existir numa base real, nem está estruturada da melhor maneira.

incompletos e com algum erro. Na verdade o usuário está procurando pela Rua da Bahia, 2277, no bairro de Lourdes em BH, MG.

Os conteúdos dos dados na base de localidade que atendem ao que o usuário deseja não são exatamente iguais aos termos que ele informará através da consulta (usando a interface do usuário na camada de aplicação). Os dados na base são os seguintes:

Nome do Atributo	Conteúdo
NúmeroLocalidade	445 (único índice da base)
UF	MG
NomeMunicípio	Belo Horizonte
NomeBairro	Lourdes
TipoLogradouro	Rua
NomeLogradouro	Bahia
NúmeroLogradouroDe	0000
NúmeroLogradouroAté	4000
CódigoCEP	30120000

Tabela 2- Instância da base de localidade

A pergunta pode ser formulada como na ilustração a seguir:

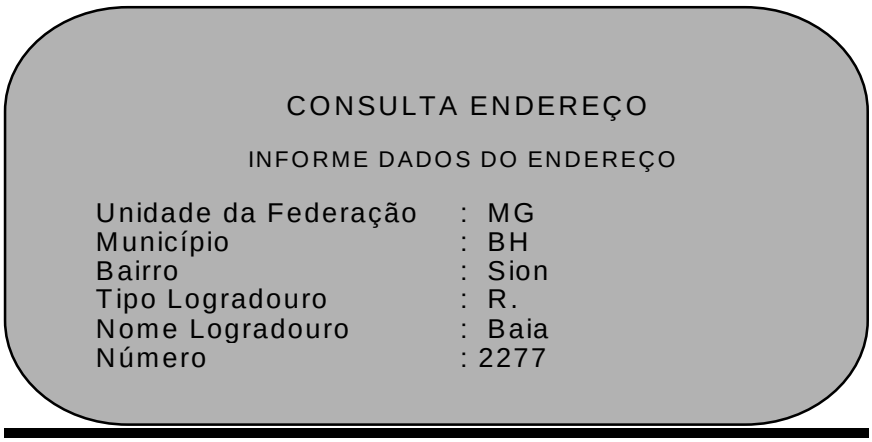


Figura 2- Formulação de pergunta da camada de aplicação

Os termos que compõem as perguntas de uma consulta devem ser contextualizados da forma como isto ocorreu no processo de aquisição de conhecimento sobre a base de localidade consultada.

O termo é contextualizado para adquirir uma significação precisa, salvaguardando conflitos de significado, reduzindo ambigüidades. Um termo contextualizado pode ser a concatenação de vários termos básicos ou de outros termos contextualizados. Os termos contextualizados complexos são também uma forma de obtenção de maior seletividade e conseqüentemente melhor desempenho do motor de inferência

O número de vezes que um termo ocorre num acervo é obtido para determinação da forma mais eficiente de busca pelo motor de inferência. As perguntas são insertas uma a uma numa base de perguntas. Um contador de perguntas é incrementado e tornado disponível para o mediador para uso posterior quando da obtenção de respostas ordenadas de forma aproximativa (podendo ser na forma ascendente ou descendente de proximidade), com base numa ordem de correspondência dos conjuntos de respostas.

Uma vez terminada a inserção, é feita uma seleção entre as perguntas para obter aquela cujo número de termos ocorre o menor número de vezes no acervo consultado. Com isto encerra-se a etapa de formulação de pergunta(s). O conceito de termo mais seletivo assumido é o do que ocorre o menor número de vezes, portanto o que restringe mais o universo de respostas da consulta a partir do termo que ocorrer menos.

A obtenção de respostas se faz pelo acionamento de uma visão SQL, a qual é o resultado da superposição de outras visões SQL que executam de forma transparente e independente da camada de aplicação do usuário, no interior do software gerenciador de banco de dados relacional que as hospeda. O mecanismo de inferência é em última análise a interface “para baixo” do mediador.

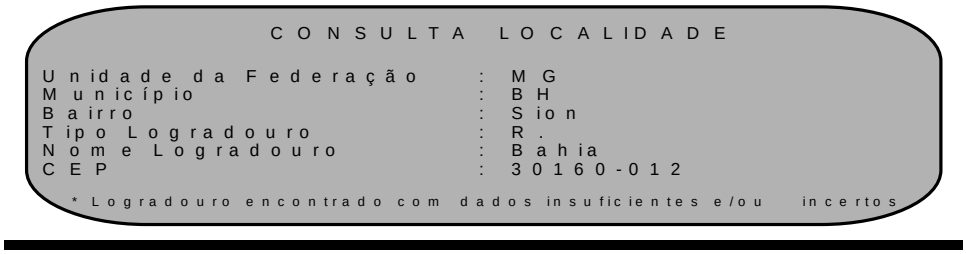
Uma resposta é um elemento de uma consulta que se caracteriza por um conjunto de perguntas e respostas. As respostas identificadas por usuário/acervo/número de ordem de pergunta são o resultado da inferência realizada. Os identificadores de instâncias de acervos, retornados pelo mecanismo de inferência, são as respostas encontradas.

Por meio dos identificadores se pode ter acesso aos acervos e obter os demais dados. O destino que será dado às respostas depende da camada de aplicação, pois o trabalho do mediador foi cumprido.

Nome do Atributo no SGBD	Descrição
Idusuário	Identificação do usuário (login, URL etc)
NúmeroOrdemConsulta	Número que indica a que consulta pertence a pergunta, quando mais de uma consulta simultânea ocorre
NúmeroAcervo	Número Identificador de acervo de dados geográficos
IdObjetoAcervo	Número seqüencial de objeto no acervo (chave de acesso)

Tabela 3- Atributos da visão Resposta

A inferência é acionada após insertas as perguntas. No exemplo utilizado na etapa de formulação de perguntas, alguns dos conteúdos dos termos de consulta não coincidem na forma, mas se aproximam semanticamente com os da base de



localidade. Alguns dos termos estão errados ou apresentam grafia errônea, como é o caso de bairro e logradouro.

Figura 3- Representação da interface de resposta da camada de aplicação

A camada de aplicação neste exemplo é a consulta a uma base de localidade. Neste nível deve ser explorado com criatividade o potencial de utilização da

inferência, privilegiando a qualidade da interface oferecida ao usuário.

Figura 4- Fluxo genérico de obtenção de respostas. A interação entre a interface "para baixo" e “para cima” por meio da visão Resposta

O conjunto-resposta é proveniente da interseção entre o repertório de termos do usuário e o da base de conhecimento sobre o universo de discurso do domínio consultado. É produzido com os termos correspondentes, relacionados às instâncias da(s) base(s) de dados. As respostas são ordenadas da maior para a menor proximidade, utilizando o anteriormente já mencionado contador de pergunta, o qual permite agrupar as respostas pelo número de termos encontrados em cada conjunto.

A conexão entre o mediador e a base de dados se faz através de relacionamentos entre identificadores de termos contextualizados e das instâncias da base de dados, estabelecendo-se uma indexação indireta através de um índice semântico. Por esta razão, os dados gráficos de informações geográficas podem ser obtidos por meio dos termos utilizados pelo usuário ou agregados automaticamente por um programa de consulta. O inverso é obtido da mesma forma, ou seja, com termos oriundos das imagens geográficas se pode obter os dados alfanuméricos.

O que se passa entre a inserção de perguntas e a obtenção de respostas está sob a custódia do mecanismo inferencial. Cabe à interface de mediação "para baixo" proporcionar o nível adequado de abstração em relação à base de dados específica, tornando independentes o software mediador e seus demais componentes.

O cerne do MIR é um conjunto de tabelas e visões SQL que trabalham juntas como componentes que provêem serviços básicos de inferência semântica [Fle94; Fle97]. O cerne compreende as seguintes tabelas e visões constituintes[2]: tabela *Base de Conhecimento Semântico*; tabela *Estatísticas de Utilização de Termo*; tabela *Consulta*; visão *Conjunto Resposta Intermediário* (conjunto de identificadores de instâncias de bases de fatos relacionadas com o termo contextualizado mais seletivo dentre os termos de uma consulta de usuário); visão *Resposta* (conjunto de identificadores de instâncias, organizado em função de conjuntos de respostas mais aproximadas até os mais distantes).

Identificadores de tuplas apontam para instâncias de uma base de dados que podem estar armazenadas sob qualquer modelo de SGBD ou sistema de arquivos. O mecanismo de indexação indireta provido pelo MIR permite liberar o mediador da dependência de esquemas de bases de dados ou especificidades de implementação de software de gerência de base de dados geográficos.

3.Avaliações & medições

Há implementações de mediadores utilizando SC e MIR em vários sistemas, tais como pesquisa bibliográfica, pesquisa criminal, consultas por nome de pessoas, instituições e órgãos entre outras. Cada uma delas utiliza diferentes ontologias, bases de conhecimento e MIR. Nenhuma das implementações já feitas utilizou linguagem

[2] Os nomes de tabelas e visões não são os implementados segundo a sintaxe SQL, mas nomes mais significativos

de programação orientada a objeto, o que gerou situações indesejáveis devidas à baixa reutilização, herança e ocultamento de detalhes [Fle97].

Um exemplo significativo do porte das implementações, dentre outros, é o de um Sistema de Informações Policiais da Secretaria de Segurança Pública de Minas Gerais o qual tem uma base de fatos de cerca de onze milhões de cidadãos cadastrados, uma base de conhecimento de 60 milhões de tuplas e uma ontologia de 20 milhões de termos. Um outro sistema de porte expressivo mantém informação sobre aproximadamente quatrocentos mil funcionários públicos, uma base de conhecimento de 40 milhões de tuplas e uma ontologia de dez milhões de termos.

Uma comparação entre dois sistemas foi feita para avaliar e medir o grau de cooperação com usuários consultantes (ver tabelas 4 e 5[3]). Um deles usa um mediador como componente, o outro não. Os sistemas comparados são especializados em localidades de todo o Brasil e seus dados têm a mesma fonte de origem: os correios. As informações vão do nível de unidade de federação até o de logradouros [Fle97].

Alguns mediadores, a exemplo dos dois sistemas citados acima, têm bases de conhecimento maiores que as do mediador de localidades, mas ele é o mais requisitado e de uso concorrente no ambiente em que foi avaliado. Sua base de conhecimento é de cerca de dez milhões de tuplas e sua base de termos possui aproximadamente três milhões de termos básicos.

O sistema não mediado, cujo nome é CEP Digital, é um produto de software para uso pessoal que executa em equipamentos *stand-alone*. Ele oferece uma interface

---

para o entendimento.

[3] Tanto na tabela 4 quanto na tabela 5, "S" significa Sim, "N" significa Não, "LP" significa **L**ista **P**equena (menos de 40 itens encontrados), "LG" significa Lista Grande (mais de 40 itens encontrados).

gráfica externa e usa técnicas de pesquisa baseadas em igualdades de cadeias de caracteres para recuperação de dados postais. O sistema mediado é usado em um ambiente interativo, concorrente e multiusuário. Ele é suportado por um equipamento de grande porte que hospeda a base de fatos e os componentes de dados e programas do mediador, os quais são invocados intensivamente por diversas aplicações corporativas.

A informação de localidade obtida pelo mediador é usada em uma variedade de formas pelas aplicações de usuário, por exemplo: para exibição de código postal, validação e padronização de entrada de dados, obtenção de identificador seqüencial de localidade para substituir dados genéricos de localidade, representando campos de dados tais como local de nascimento, endereço residencial e comercial etc.

Os sistemas comparados executam em cenários bastante distintos, por isto a comparação é restrita à mesma coleção de dados, usando os mesmos termos de consulta em diferentes ambientes.

Não foi nossa meta responder questões relacionadas a características de interface externa, performance de hardware, performance de SGBDR ou questões ambientais. Entretanto, questões técnicas do MIR, relacionadas à seletividade e performance em grandes bases de fatos, da ordem de dezenas de milhões de tuplas, demandaram soluções. A superação de tais restrições produziu drástica redução de tempo de resposta, tempo de CPU e número de leituras/gravações [Fle97].

O código de programação utilizado na interface “para cima” foi implementado em linguagem de quarta geração (L4G) sob um paradigma não orientado a objeto. A implementação da interface “para baixo” usa somente implementação SQL para um

conjunto interoperativo de tabelas e visões que é ativado a partir da interface “para cima”.

3.1 Pesquisa cidade

A tabela 4 apresenta dados produzidos durante o teste dos softwares CEP Digital e mediador de localidade para avaliar o grau de cooperatividade de cada um. Três cidades foram pesquisadas usando diferentes escritas de nome: Belo Horizonte, São Thomé das Letras e Rio de Janeiro. Além de nome de cidade a camada de aplicação permite que as iniciais de unidade de federação (UF) sejam informadas pelo usuário ou não.

Nomes de Cidades	**CepDigital**		**Mediador**	
	Sem UF	Com UF	Sem UF	Com UF
BH	N	N	S	S
BHZ	N	N	S	S
B Horizonte	N	N	S	S
B Orizonte	N	N	S	S
Belo Horizonte	S	S	S	S
Belo	N	N	N	N
São Thomé das Letras	N	N	S	S
São Tomé das Letras	S	S	S	S
São Tomé Letras	N	N	S	S
São Tomé	LG	N	LP	LP
Rio	N	N	S	S
R de Janeiro	N	N	S	S
R Janeiro	N	N	S	S
R J	N	N	S	S
Rio de Janeiro	S	S	S	S
São Sebastião do Rio de Janeiro	N	N	S	S

Tabela 4- Consulta por nome de cidade e resultados retornados por software avaliado

3.2 Pesquisa logradouro

A tabela 5 mostra resultados de pesquisa de nome de logradouro utilizando diferentes escritas em relação ao nome existente na base de dados. Foi utilizada como exemplo a Rua Professor Aníbal de Matos. Além de nome de logradouro, as iniciais de unidade de federação, nome de cidade, tipo de logradouro e nome de bairro podem ter sido informados ou não na consulta.

O teste consistiu na formulação de consultas típicas sobre localidades do tipo logradouro, utilizando variações de escrita de nome de logradouro.

UF	Cidade	Tipo de Logradouro	Nome de Logradouro	Nome de Bairro	CepDigital achou?	Mediador achou?
—	—	—	Aníbal Matos*	São Pedro**	N	LP
—	—	R	Aníbal Matos	São Pedro	N	LP
—	—	Rua	Aníbal Matos	São Pedro	N	LP
—	—	Rua	Professor Aníbal Matos	São Pedro	N	S
—	—	Rua	Professor Aníbal de Matos	São Pedro	N	S
—	—	Avenida	Prof.Aníbal Matos	São Pedro	N	S
MG	Belo Horizonte	Rua	Professor Aníbal de Matos	Santo Antônio	N	S
MG	Belo Horizonte	Rua	Prof Anïbal de Matos	Santo Antônio	N	S
MG	Belo Horizonte	R.	Professor Aníbal de Matos	Santo Antônio	S	S
MG	Belo Horizonte	Rua	Professor Aníbal de Matos	— ou S Antônio	N N	LG LG
MG	Belo Horizonte	Rua	Professor	— ou S Antônio	LG LG	LG LP
MG	Belo Horizonte	Rua	Anïbal	— ou S Antônio	N	LP

Tabela 5- Consulta por nome de localidade e resultados retornados por software avaliado

As medições feitas buscaram abordar aspectos intrínsecos ao mediador [Fle97]. Focalizaremos aqui apenas a importância da seletividade dos termos para desempenho eficiente do motor de inferência relacional . Nos grandes sistemas, sejam as grandezas relativas ao volume de dados, ao número de transações ou à importância dos serviços prestados, estes problemas que em pequenos e médios[4] sistemas podem não se manifestar, se tornam críticos.

O conceito de seletividade é simples: dados *n* termos utilizados numa consulta, o mais seletivo deles é o que tem o menor número de relacionamentos com instâncias de acervos. Assim, todos os demais termos deverão ser encontrados a partir do universo de instâncias obtidas com o termo mais seletivo. Os números de ocorrências de todos os termos são mantidos e atualizados numa base de dados e são obtidos dali quando as perguntas estão sendo processadas.

A definição do que é um termo e como deve ser formado afeta diretamente a seletividade. Os termos podem ter existência individual ou podem ser complexos, concatenados uns aos outros.

Se a escolha inicial dos termos não for adequada, em sistemas com grandes bases de conhecimento e muitas consultas simultâneas o desempenho do MIR pode ser comprometido pela carência de termos com boa seletividade.

Relação das 10 consultas utilizadas para aferir os ganhos de seletividade após mudança na forma de montagem dos termos para formulação de perguntas. Cada consulta foi medida de forma isolada (LOPES, 1996).

1. Rua Professor Aníbal de Matos
2. Rua Conceição do Mato Dentro
3. Rua João César de Oliveira
4. Rua João Pinheiro
5. Rua Coronel José Nogueira Duarte
6. Praça Engenheiro Nogueira de Sá
7. Rua Doutor José Silva Martins
8. Rua Doutor José Esteves Rodrigues
9. Rua Capitão José Carlos Vaz de Melo
10. Av. Contorno

O fundamento do aumento de seletividade se baseia na combinação dos termos isolados para formar termos complexos com a concatenação dos termos isolados a outros termos complexos, de forma que se criem termos mais seletivos.

Por exemplo: a localidade de São Thomé das Letras estava associada inicialmente aos termos: São, Thomé e Letras. Após a combinação de termos passou a estar associada também aos termos: São Thomé, São Letras, Thomé Letras e São Thomé Letras. Estes termos complexos, ocorrem muito menos vezes que cada um dos termos isoladamente.

Argumentos de Pesquisa	**Termos Simples e Complexos Produzidos**
São Thomé das Letras	São
	Thomé
	Letras
	São Thomé
	São Letras
	Thomé Letras
	São Thomé Letras

Tabela 6- Idéia básica do processo de formação de termos complexos para produzir maior seletividade e melhoria de desempenho

Esta solução aumenta a localização direta de localidades com nomes compostos com conseqüente redução de consumo de recursos. O aumento de seletividade também contribui para a redução do travamento de páginas de dados, pois as consultas passam a ser atendidas mais rapidamente, reduzindo a contenção de recursos do ambiente.

4.Conclusão

Tratamento de contexto, termos sinônimos, consultas com faixas de termos, múltiplas bases de fatos e usuários, consultas simultâneas por usuário, respostas aproximativas e abstração de esquema são alguns aspectos já implementados, testados e em uso.

Uso de *caching* (explorando o conceito de localidade pela alta possibilidade de repetição de determinados termos avizinhados semanticamente), estatísticas de seletividade de termos buscando trabalhar a partir do termo que reúne o menor número de instâncias e controles de concorrência, gargalos e explosão combinatória foram implementados no cenário real mencionado na seção 3 ([Fle97] discute detalhes).

É provável que, no futuro, agentes-mediadores sejam o elemento de integração nas redes; incrementando a densidade e relevância da informação por meio de interfaces cooperativas, capazes de ocultar a complexidade e sustentar o processo decisório diante da rapidez das mutações.

Referências Bibliográficas

[Bus45] BUSH, Vannevar. As We May Think. *The Atlantic Monthly*. July, 1945, p. 13. Ottawa, Canada: University of Ottawa. DUCHIER, Denys (ed). Versão HTML disponível em <www.isg.sfu.ca/~duchier/misc/vbush/vbush.all.shtml>

[Dam97] DAMACENO, Eduardo Teixeira. *Relatório de Uitilização do MIR*. Belo Horizonte, MG: Prodemge, DTP/STP/GPP, 30 de jul.,/ 1997. (Relatório de Desempenho).

[Fis94] FISCHER, Gerhard. Domain-Oriented Design Environments. *Automated Software Engineering*, , 1994 p.177-203.

[Fle94] FLECHA, Mário L. O. . Processamento Semântico de Informação. Construindo um Motor de Inferência Relacional - MIR -, para Desenvolvimento de Sistemas Cognitivos. Belo Horizonte, MG: Prodemge, 1994, p.68. *(Relatório Técnico n. 25*).

[Fle97] FLECHA, Mário L. O. . *Processamento Semântico na Mediação de Consultas a Bases de Dados Públicas.* Dissertação de Mestrado, Belo Horizonte: Fundação João Pinheiro - Escola de Governo/Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, 1997, 223p.

[Gen95] GENESERETH, Michael et al. *Reference Architecture for the Intelligent Integration of Information Retrieval. Prepared by the Program on Intelligent Integration of Information (**I**^{3}).* (Version 2.0 (Draft). DARPA - Defense Advanced Research Projects Agency, Aug., 2, 1995, 90p.

[GF92] GENESERETH, Michael, FIKES, Richard E. et al., *KIF - Knowledge Interchange Format. Version 3.0 Reference Manual.* Stanford, CA: Computer Science Department, Stanford University. (Report Logic-92-1) June, 1992, 68p.

[HG97] HART, Peter E., GRAHAM, Jamey. Query-free Information Retrieval. In: Cooperative Information Systems. *IEEE Expert*, p.32-37, Sep/Oct, 1997.

[MP97] MYLOPOULOS, John, PAPAZOGLOU, Michael. Cooperative Information Systems. In: Guest Editors' Introduction. *IEEE Expert*,. p.28-37, Sep/Oct, 1997.

[Qui68] QUILLIAN, Ross. Semantic Memory. In: *Semantic Information Processing*. Cambridge, Mass.: M.I.T.. Minsky, Marvin (ed). Ed. The MIT, 1968, p.227-270.

[Rho98] RHODES, Philip. Abundance of Information. How do Designers use Information? Belo Horizonte, MG: Universidade Federal de Minas Gerais - UFMG, Escola de Design, 1998, 7p.

[Rum94] RUMBAUGH, J. et al. *Modelagem e projetos baseados em objetos*. Trad. Dalton Conde de Alencar: Rio de Janeiro: Campus, 1994, 654p.(Tradução de: Object-oriented modeling and design. Englewood Cliffs, New Jersey: Ed. Prentice-Hall, 1991).

[She95] SHNEIDERMAN, Ben. *Designing the user interface* , Rio de Janeiro: Campus, 1994, 654p.

[Ull77] ULLMANN, Stephen. *Semântica - uma introdução à ciência do significado.* 4.ed. (português). Trad. J. A. Osório Mateus. Lisboa: Fundação Calouste Gulbenkian, 1977, 578p. (Tradução de: Semantics - An Introduction to the Science of Meaning. Oxford: Ed. Basil Blackwell, 1964).

[Web86] WEBBER, Bonnie L.. Questions, Answers and Responses: Interacting with Knowledge Base Systems. In: *On Knowledge Base Management Systems.* Integrating Artificial Intelligence and Database Technologies. Topics in

Information Systems. Michael L. Brodie/John Mylopoulos (ed). Ed. Springer-Verlag, 1986, p.365-402.

[WG95] WIEDERHOLD, Gio, GENESERETH, Michael. Basis for Mediation. May, 1995, In: PROC. INTERNATIONAL CONFERENCE ON COOPERATIVE INFORMATION SYSTEMS (CoopIS95). Viena, Austria: available in <coopis@cs.toronto.edu>, May, 1995, p.138-155.

[WG97] WIEDERHOLD, Gio, GENESERETH, Michael. The Conceptual Basis for Mediation Services. In: Cooperative Information Systems. *IEEE Exper,*. p.38-47, Sep/Oct, 1997.

[Wie96] WIEDERHOLD, Gio. Foreword and Glossary: Intelligent Integration of Information. In: *Intelligent Integration of Information.* Norwell, Mass.: Gio Wiederhold (ed). Ed. Kluwer Academic Publishers: p.5-9, p.193-201May/June, 1996. (Journal of Intelligent Information Systems, Integrating Artificial Intelligence and Database Technologies, 6(2/3)).

[Wie92] WIEDERHOLD, Gio. Mediators in the Architecture of Future Information Systems. *Computer,* p.38-49, 1992.