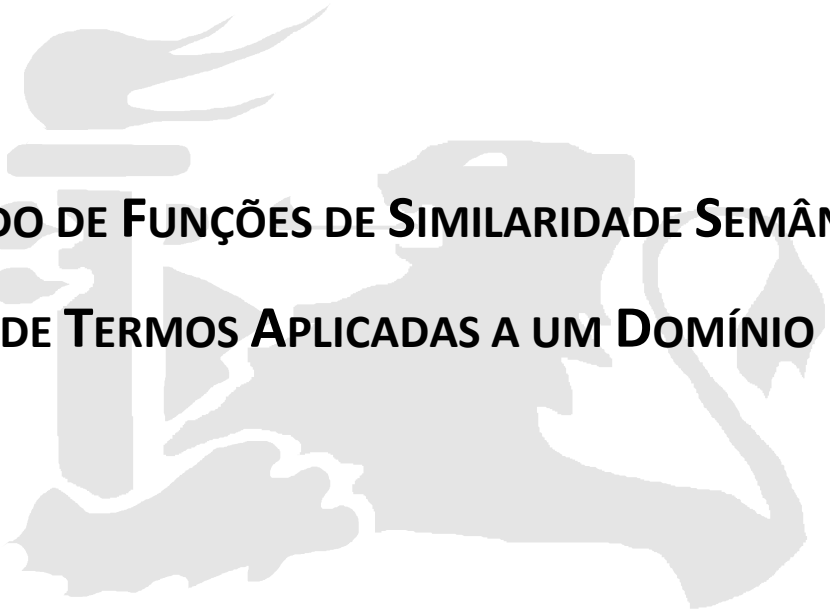
UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

2007.2



ESTUDO DE FUNÇÕES DE SIMILARIDADE SEMÂNTICA DE TERMOS APLICADAS A UM DOMÍNIO

TRABALHO DE GRADUAÇÃO

Aluno – Daniel Ferreira da Silva (dfs3@cin.ufpe.br)
Orientadora – Ana Carolina Salgado (acs@cin.ufpe.br)
Co-orientadora – Rosalie Barreto Belian (rbb@cin.ufpe.br)

Janeiro de 2008

UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

2007.2

DANIEL FERREIRA DA SILVA

ESTUDO DE FUNÇÕES DE SIMILARIDADE SEMÂNTICA DE TERMOS APLICADAS A UM DOMÍNIO

Este trabalho foi apresentado à graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora – Ana Carolina Salgado (acs@cin.ufpe.br)
Co-orientadora – Rosalie Barreto Belian (rbb@cin.ufpe.br)

Janeiro de 2008

Agradecimentos

Primeiramente agradeço a Deus, meus pais, familiares e amigos por terem me dado toda a estrutura de *background* necessária para realização do projeto.

Também agradeço a professora Ana Carolina Salgado pela disposição e por ter me dado esta oportunidade de aprendizado conjunto.

Agradecimento também muito especial a minha esposa Fanávida Almeida e minha co-orientadora Rosalie Barreto Belian, por todo o apoio, paciência, dedicação, ajuda e principalmente por terem acreditado na conclusão deste trabalho.

Daniel Ferreira

Resumo

O principal objetivo de um ambiente de integração de dados é fornecer ao usuário uma visão integrada de diversas fontes de dados distribuídas e heterogêneas, criando a impressão no usuário de se estar utilizando um sistema centralizado e homogêneo. Um dos problemas cruciais dos sistemas de integração de dados diz respeito à integração semântica dos esquemas das fontes, um processo reconhecidamente difícil de ser realizado de forma automática, sendo necessária muitas vezes, a intervenção do usuário.

Dentre os diversos sistemas de integração de dados, este trabalho está inserido no contexto que foi proposto por Bernadette Farias Lóscio sob a orientação da professora Ana Carolina Salgado no Centro de Informática da Universidade Federal de Pernambuco, o Integra [Lóscio, 2003], que visa o desenvolvimento de um sistema de integração de várias fontes de dados distribuídas na Web.

O objetivo deste trabalho é estudar funções de similaridade existentes para empregar num processo de identificação da similaridade semântica a ser utilizado na integração de esquemas do Integra [Belian, 2007]. Um processo de integração de esquemas recebe dois ou mais esquemas como entrada e produz um único esquema como resultado [Rahm *et al.* 2001]. Neste caso, é preciso identificar a similaridade semântica dos elementos dos esquemas para realizar a sua integração.

Palavras-chaves: Banco de Dados, Integração de esquemas, Semântica, Contexto, Similaridade semântica.

Abstract

The main purpose of a data integration environment is to supply the user an integrated vision of several sources of data, distributed and heterogeneous, creating the illusion of a centralized and homogeneous system. One of the crucial problems of data integration systems concerns the semantic integration of the source schemas, a difficult process to accomplish automatically, often requiring the user's intervention.

This work is part of Integra, a data integration system proposed by Bernadette Farias Lóscio under Ana Carolina Salgado orientation in the Center of Computer Science of the Federal University of Pernambuco [Lóscio, 2003]. Integra aims to deal with several Web data sources.

The objective of this work is to study existing functions of similarity to employ in a process of identification of the semantic similarity to be used in the integration of data source schemas in Integra [Belian, 2007]. A process of schema integration receives two or more schemas as input and produces a single schema [Rahm *et al.* 2001]. In this case, it is necessary to identify the semantic similarity between schema elements to carry out their integration.

Keywords: Data bases, Schema Integration, Semantics, Context, Semantic Similarity.

Índice

Índice de Figuras

1. Introdução

O crescente surgimento de informações a cada dia na web e em qualquer meio de armazenamento compartilhado tem feito surgir a necessidade de sofisticados mecanismos de busca dessas informações. Apenas a busca por palavras-chave, por exemplo, não tem sido satisfatória em alguns casos. Além disso, a grande heterogeneidade das bases de dados disponíveis atualmente faz surgir também a necessidade de sistemas que integrem tais bases num sistema unificado para facilitação de posteriores consultas com maior precisão e eficiência.

Como a maioria das informações disponíveis hoje se encontra espalhada na web, o uso das tecnologias preconizadas pela "Web Semântica" torna-se cada vez mais desejável. A Web Semântica consiste numa web com toda sua informação organizada de forma que não somente seres humanos possam entendê-la, mas principalmente máquinas. Através desse sistema, o processo de integração obtém um melhor resultado no processo de extração dos dados da web, uma vez que os mesmos já se apresentam estruturados. Sistemas de integração de informações na Web compõem o cenário da Web semântica [Berners, 2001] 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.

Muito se tem feito para que a integração dos dados dessas bases heterogêneas seja feita da forma mais automática possível, sem intervenção humana. Porém, um dos principais problemas dos sistemas de integração de dados diz respeito à integração semântica dos esquemas das fontes de dados, e uniformizá-los é um processo reconhecidamente difícil de ser realizado automaticamente. Como resultado, obtemos conceitos como *ontologias*, *metadados*, *contextos* e similaridade semântica, que exercem uma importante influência no entendimento e no desenvolvimento do resultado final num processo de integração de dados utilizando bases heterogêneas. *Ontologias* são especificações explícitas de uma *conceitualização* em algum domínio

[Gruber, 1993]. Já *metadados* [Kashyap, 1996] são comumente definidos como "dados sobre dados", e podem descrever significado, conteúdo, organização ou objetivos de algum conjunto de dados. *Contextos,* por sua vez*,* "contém metadados relacionados ao seu significado, propriedades (tais como fonte, qualidade e precisão), e organização" [Goh, 1997; Wache, 2001]. De acordo com o contexto que o termo está inserido, um termo pode receber diversos significados distintos já pré-estabelecidos, por exemplo, em bases de conhecimentos distribuídas na web. Sendo assim, contextos são considerados ferramentas eficazes no tratamento da heterogeneidade da informação [Wache, 2001]. Neste cenário um conceito fundamental é o de similaridade semântica [Lin, 2000]. Através do cálculo de similaridade é possível identificar que elementos dos esquemas das fontes de dados são semanticamente similares e que então devem ser utilizados no processo de integração 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 [Kashyap, 1996]. O estabelecimento da similaridade entre objetos, baseada em princípios puramente esquemáticos e estruturais, foi discutido na literatura existente e considerado ineficiente para determinar a integração destes objetos [Ouksel, 1999]. Como resultado, devemos considerar fortemente o uso de funções de similaridade semântica para realizar a ligação entre os objetos desejados.

Este trabalho está organizado da seguinte forma:

- O **Capítulo 2**, Integração de Esquemas – Conceitos Básicos, realiza uma breve explicação de conceitos básicos necessários para se entender todo o processo de integração de Esquemas, contexto no qual o trabalho está inserido;
- O **Capítulo 3**, Medidas de Similaridade Semântica, mostra algumas das principais e mais utilizadas abordagens para se obter similaridade semântica entre os termos de domínios similares ou distintos;
- O **Capítulo 4**, Funções de Similaridade Semântica, faz um estudo das principais funções e métodos propostos por alguns autores conceituados na

literatura disponível atualmente, inclusive fazendo uso de diversas medidas abordadas no Capítulo 5;

- O **Capítulo 5**, Comparativo das Funções, realiza uma comparação detalhada das vantagens e desvantagens das funções citadas no Capítulo 4. O objetivo desta seção é mostrar qual a melhor função a ser utilizada no contexto abordado por este trabalho; e
- O **Capítulo 6**, Conclusão e Trabalhos Futuros, apresenta a conclusão do trabalho e as próximas etapas que podem ser abordadas como continuidade deste trabalho.

2. Integração de Esquemas – Conceitos Básicos

Um esquema é uma coleção de objetos de um banco de dados que estão disponíveis para um determinado usuário ou grupo [Elsmari *et al.* 1999]. Os objetos de um esquema são estruturas lógicas que se referem diretamente aos dados do banco de dados. Eles incluem estruturas, tais como tabelas, visões, seqüências, procedimentos armazenados, sinônimos, índices, agrupamentos e links de banco de dados. Os esquemas neste projeto serão codificados no formato X-Entity [Lóscio *et al.* 2003], o qual é um pouco mais detalhado na seção 2.2 deste trabalho.

Nomes ou *labels* dos elementos dos esquemas são usualmente formados por palavras ou conjunto de palavras, os quais têm a função de representar de forma léxica os respectivos elementos. Entretanto, tais palavras, antes da função de similaridade semântica entrar em ação, precisam ser normalizadas.

Tal normalização consiste em realizar um pré-processamento nas palavras para retirar caracteres especiais, acentos, hífens, em alguns casos os espaços em branco ou caracteres não pertencentes à linguagem que se está trabalhando. Outro exemplo de tarefa importante na normalização é o processo de expandir as abreviaturas encontradas, como “id”, “num”, entre outros. Além disso, se faz necessário separar individualmente as várias palavras que compõem os *labels* dos elementos (*tokens)*, para daí então obtermos os termos desejados dentro do texto que foi passado na entrada dos dados. O trabalho que está apresentado já considera que os termos estão normalizados para se aplicar as funções de similaridade estudadas.

2. 1 Definição de Similaridade Semântica

Similaridade é um conceito fundamental e amplamente utilizado. Pessoas do mundo real identificam sinônimos mesmo que as palavras não tenham a mesma grafia ou escrita semelhante, como por exemplo, agachar – abaixar, pôr – colocar, roupa – vestuário, ou em alguns casos, mesmo que as palavras não sigam a regra gramatical

que define o que são sinônimos, são identificadas como tal, como por exemplo, Anjo – Querubim, Automóvel – Carro, Voar – Liberdade, e assim por diante.

Miller e Charles (1991) empregam uma definição formal de similaridade, porém pouco precisa, que geralmente é atribuída a Leibniz: “Duas palavras são ditas sinônimos, se numa frase ou proposição, uma pode ser substituída pela outra sem perda de significado”. Porém, sistemas NLP geralmente estabelecem o grau de sinonímia entre duas palavras através da sua similaridade semântica. A similaridade semântica entre dois ou mais termos pode ser calculada através de diversas funções, que consideram as mais variadas informações. McGill et al. (1979) pesquisaram e compararam cerca de 67 métricas de similaridade para recuperar informações.

No presente trabalho, estão sendo abordadas algumas das funções mais relevantes e estudadas na literatura para se desenvolver o processo de integração de esquemas, nomeadas pelos seus respectivos autores.

2.2 O Sistema Integra

O sistema de integração de dados proposto por Lóscio (2003), chamado Integra, tenta sobrepor a dificuldade de integrar informações de múltiplas fontes com estruturas heterogêneas através do uso de um modelo de dados comum para representação do conteúdo e da estrutura das fontes. Ela adota o XML [Bray, 1999] para troca de dados e integração. Em [Lóscio, 2003] também foi proposto o modelo X-Entity, que é usado para prover uma abstração de alto nível para as informações descritas nos esquemas XML, o qual está descrito na seção 2.4. O X-Entity é utilizado para descrever tanto o esquema global quanto o esquema das fontes locais.

Arquitetura do Integra

O Integra está basicamente dividido em quatro ambientes: Ambiente Comum, Ambiente de Geração das Consultas de Mediação, Ambiente de Integração de Dados e o Ambiente do Usuário [Lóscio, 2003]. Na Figura 1 é apresentada uma arquitetura resumida do sistema:

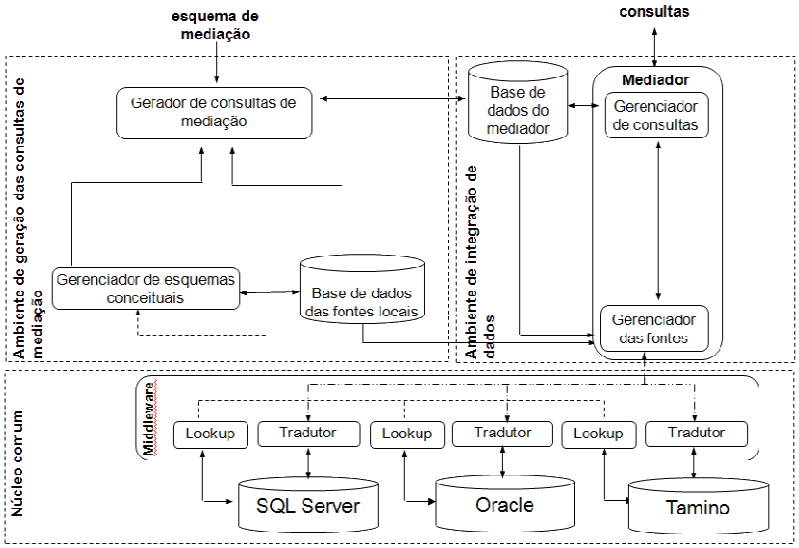


Figura 1 – Arquitetura resumida do Sistema Integra. [Lóscio, 2003]

O calculo da similaridade semântica é realizado em algumas etapas específicas do processo de integração de esquemas do Integra. O módulo no qual este trabalho está inserido é o Ambiente de Geração de Consultas de Mediação, onde a similaridade semântica entre termos é necessária para o processo de unificação de esquemas das fontes. Neste processo, é necessário distinguir e escolher quais os melhores termos a serem usados para nomear os elementos e atributos no esquema de mediação que é gerado como saída para etapas posteriores de integração.

2.3 Integração de Esquemas no Integra

O Processo de integração de esquemas do Integra, procura resolver as diferenças semânticas entre os elementos dos esquemas, identificando o significado de cada elemento antes de sua integração [Belian, 2007]. Neste sentido, ele utiliza uma ontologia de contextos que representa informações sobre o vocabulário das fontes de dados e também informações contextuais que auxiliam no esclarecimento do significado dos elementos, como discutido no Capítulo 1. Neste processo, o Integra necessita de uma função para cálculo da similaridade semântica entre termos da ontologia que representam os elementos dos esquemas das fontes de dados.

2.4 O Modelo Conceitual X-Entity

X-Entity [Lóscio, 2003] é um modelo conceitual para representação de XML Schemas e é uma extensão do modelo ER [Chen, 1976]. O modelo X-Entity se baseia principalmente no conceito de Entidade que representa a estrutura de elementos de um esquema XML composto por outros elementos e atributos. Este modelo apresenta o tipo relacionamento que permite representar os relacionamentos entre elemento e sub-elemento, bem como a associação entre elementos. O X-Entity também dispõe de uma representação gráfica para os esquemas.

Conceitos Básicos

O principal elemento de um modelo X-Entity é a entidade. Ela representa a estrutura dos elementos XML composta por outros elementos e atributos. No modelo X-Entity, os relacionamentos podem ser do tipo *contém* (um elemento contém outro elemento) ou *referência* (um elemento referencia outro elemento). De acordo com a descrição para o X-Entity [Costa, 2005], temos a definição formal dos principais elementos a seguir:

- **ENTIDADE** – uma entidade E, denotada por E({A_1,..., A_n}, {R_1,..., R_m}), é composta por uma entidade de nome E, um conjunto de atributos A_1,..., A_n e

um conjunto de relações R_1,..., R_m. Uma entidade representa um conjunto de elementos com uma estrutura complexa, composta de atributos e outros elementos (chamados de sub-elementos). Uma instância de uma entidade é um elemento particular no documento XML fonte. Cada entidade tem atributos $\{A_1,..., A_n\}$ que a descreve. Um atributo A_i representa tanto um atributo como também um sub-elemento que não é composto por outros elementos ou atributos;

- **RELACIONAMENTO *CONTÉM*** – um relacionamento contém entre duas entidades E e E_1 especifica que cada instância de E contém instâncias de E_1. Isso é denotado por R(E, E_1, (min, max)), onde R é um nome de relacionamento e (min, max) define o número mínimo e máximo de instâncias de E_1 que podem ser associadas com uma instância de E. Nos diagramas X-Entity, relacionamentos *contém* são exibidos com um losango rotulado com o texto *contains*. A linha conectando o relacionamento com as entidades participantes é direcionada da entidade E para a entidade E_1; e

- **RELACIONAMENTO *REFERÊNCIA*** – o relacionamento *referência*, denotado por $R_j(E_1, E_2, \{A_{11},..., A_{1n}\}, \{A_{21},..., A_{2n}\})$, especifica que a entidade E1 referencia a entidade E2. $\{A_{11},..., A_{1n}\}$ e $\{A_{21},..., A_{2n}\}$ representam os atributos de referência entre as entidades E_1 e E_2 tal que o valor de A_{1i}, $1 \leq i \leq n$, em qualquer entidade E_1 deve ter o mesmo valor de A_{2i}, $1 \leq i \leq n$, em qualquer entidade de E_2. Nos diagramas X-Entity, relacionamentos de *referência* são representados como losangos rotulados com o texto *refers*. A linha que conecta as entidades participantes é direcionada da entidade que referencia para a entidade referenciada.

Um modelo exemplo extraído de [Lóscio, 2003] que apresenta os principais elementos de um esquema X-Entity é apresentado na Figura 2.

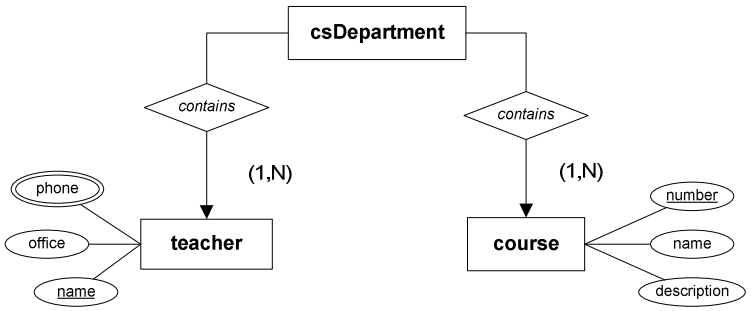


Figura 2 – Exemplo de Esquema X-Entity

No esquema apresentado podemos visualizar os seguintes elementos:

- **Entidades:** professor, csDepartment, course;
- **Relacionamentos:** *contém* (csDepartment *contém* professor e csDepartment *contém* course);
- **Atributos da entidade professor:** name (chave), office e phone;
- **Atributo multivalorado:** phone da entidade professor; e
- **Atributo obrigatório da entidade professor:** name e phone.

Além do relacionamento "contains" (containment relationship) exemplificado no esquema anterior, o X-Entity apresenta o relacionamento "refers" (reference relationship) que representa associações entre elementos do esquema (exemplificado na Figura 3).

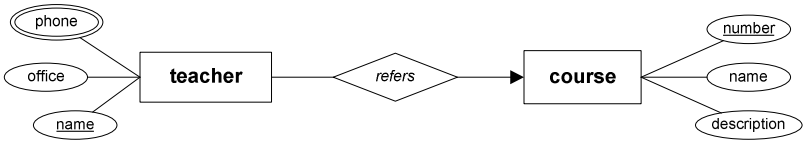


Figura 3 – Exemplo de relacionamento "refers" (reference relationship) no X-Entity

O X-Entity permite representar estruturas hierárquicas em XML Schema transformando-as em representações planas que evidenciam entidades e seus relacionamentos. Este formato privilegia os conceitos que são relevantes para a integração de esquemas ocultando detalhes de implementação, tais como os

aninhamentos entre elementos. Um processo de conversão do formato XML Schema para X-Entity foi proposto e descrito em detalhes em [Lóscio 2003].

Notação XML do X-Entity

Também em [Lóscio, 2003] foi proposto o uso de um documento XML para especificação de esquemas X-Entity.

A especificação XML para o X-Entity possui um elemento `XENTITY_SCHEMA` (o elemento raiz), que é composto pelos elementos `ENTITY`, `CONTAINMENT_RELATIONSHIP`, e `REFERENCE_RELATIONSHIP`. Um elemento ENTITY descreve os atributos e relacionamentos associados com uma entidade através dos elementos `ATTRIBUTE` e `RELATIONSHIP_NAME`, respectivamente. Um elemento `CONTAINMENT_RELATIONSHIP` é composto por dois elementos: `ELEMENT_ENTITY` e `SUBELEMENT_ENTITY`, os quais representam as entidades envolvidas no relacionamento. Um elemento `REFERENCE_RELATIONSHIP` também é composto por dois elementos que representam as entidades envolvidas: `REFERENCING_ENTITY` e `REFERENCED_ENTITY`. Além disso, o elemento `REFERENCE_RELATIONSHIP` possui os elementos `KEY` e `KEYREF`, que especificam os atributos envolvidos no relacionamento de referência.

As notações XML para os exemplos da Figura 2 e Figura 3 estão descritas nas Figuras 4 e 5, respectivamente.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<XENTITY_SCHEMA name="csDepartment Description">
  <ENTITY name="csDepartament">
    <RELATIONSHIP_NAME name="csDepartment_professor" />
    <RELATIONSHIP_NAME name="csDepartment_course" />
  </ENTITY>
  <ENTITY name="professor">
    <ATTRIBUTE name="name" type="string" cadMin="1" cardMax="1" key="key1" />
    <ATTRIBUTE name="phone" type="string" cadMin="1" cardMax="n" />
    <ATTRIBUTE name="office" type="string" cadMin="0" cardMax="1" />
    <KEY name="key1">
  </ENTITY>
  <ENTITY name="course">
    <ATTRIBUTE name="name" type="string" cadMin="1" cardMax="1" />
    <ATTRIBUTE name="number" type="string" cadMin="1" cardMax="1" key="key2" />
    <ATTRIBUTE name="description" type="string" cadMin="0" cardMax="1" />
    <KEY name="key2">
  </ENTITY>
  <CONTAINMENT_RELATIONSHIP name="csDepartment_professor" cardMin="1" cardMax="n">
    <ELEMENT_ENTITY name="csDepartment" />
    <SUBELEMENT_ENTITY name="professor" />
  </CONTAINMENT_RELATIONSHIP>
  <CONTAINMENT_RELATIONSHIP name="csDepartment_course" cardMin="1" cardMax="n">
    <ELEMENT_ENTITY name="csDepartment" />
    <SUBELEMENT_ENTITY name="course" />
  </CONTAINMENT_RELATIONSHIP>
</XENTITY_SCHEMA>
```

Figura 4 - Especificação XML para o esquema do Exemplo 1

```xml
<?xml version="1.0" encoding="UTF-8"?>
<XENTITY_SCHEMA name="csDepartment Description">
  <ENTITY name="professor">
    <ATTRIBUTE name="name" type="string" cadMin="1" cardMax="1" key="key1" />
    <ATTRIBUTE name="phone" type="string" cadMin="1" cardMax="n" />
    <ATTRIBUTE name="office" type="string" cadMin="0" cardMax="1" />
    <ATTRIBUTE name="courseNumber" type="string" cadMin="1" cardMax="n" />
    <RELATIONSHIP_NAME name="professor_ref_course" />
    <KEY name="key1">
  </ENTITY>
  <ENTITY name="course">
    <ATTRIBUTE name="name" type="string" cadMin="1" cardMax="1" />
    <ATTRIBUTE name="number" type="string" cadMin="1" cardMax="1" key="key2" />
    <ATTRIBUTE name="description" type="string" cadMin="0" cardMax="1" />
    <KEY name="key2">
  </ENTITY>
  <REFERENCE_RELATIONSHIP name="professor_ref_course">
    <REFERENCING_ENTITY name="professor" />
    <REFERENCED_ENTITY name="course" />
    <KEY name="key2">
      <ATTRIBUTE_NAME name="number" />
    </KEY>
    <KEYREF name="keyref1">
      <ATTRIBUTE_NAME name="courseNumber">
    </KEYREF>
  </REFERENCE_RELATIONSHIP>
</XENTITY_SCHEMA>
```

Figura 5 - Especificação XML para o esquema do Exemplo 2

3. Medidas de Similaridade Semântica

Em praticamente todo o processo de integração de esquemas se faz necessário o uso de uma medida de similaridade semântica entre termos. Conseqüentemente, problemas de *matching*, mapeamentos, agrupamentos e principalmente a integração final dos elementos dos esquemas são operações que necessitam do grau de similaridade entre os termos para realizar a diferenciação (ou junção) entre eles.

Para calcular o grau de similaridade semântica entre dois termos, se faz necessário o uso de algumas técnicas apropriadas de acordo com o problema em questão. Atualmente, existem diversas formas disponíveis que têm sido estudadas e propostas ao longo dos anos, como por exemplo, a estimativa do grau de similaridade semântica através da "contagem de nós" ou métodos baseados em índices.

As medidas de similaridade semântica são formadas estabelecendo o grau de *relacionamento semântico*, da *similaridade semântica* ou estabelecendo o cálculo da *distância semântica* entre dois termos. O conceito de *relacionamento semântico* é mais abrangente do que o de *similaridade semântica*, enquanto que *distância semântica* (ou dissimilaridade) é justamente o oposto de relacionamento semântico [Rodriguez *et al.* 1999].

Geralmente, as medidas de similaridade semântica são classificadas dentro de quatro categorias principais [Wang, 2005, Petrakis *et al.* 2006], as quais serão detalhadas nas próximas seções:

a) Abordagem baseada em ontologias;
b) Abordagem baseada no índice de informações compartilhadas;
c) Abordagem baseada em características;
d) Abordagem híbrida (algum tipo de combinação das três anteriores).

3.1 Abordagem baseada em ontologias

Esta categoria abrange todas as abordagens que usam recursos e bases de conhecimento (como ontologias, dicionários e vocabulários) para melhorar o cálculo do grau de similaridade semântica entre os termos. Como conseqüência, essas abordagens estão geralmente baseadas em redes ou estruturas de grafos, usando propriedades de caminho (tamanho) para calcular o grau de similaridade semântica ou a distância semântica. Usualmente, esse tipo de abordagem se utiliza de relacionamentos do tipo *is-a* para definir relações de subclasses e superclasses entre os conceitos presentes na hierarquia da ontologia. A Figura 6 exemplifica uma taxonomia simples.

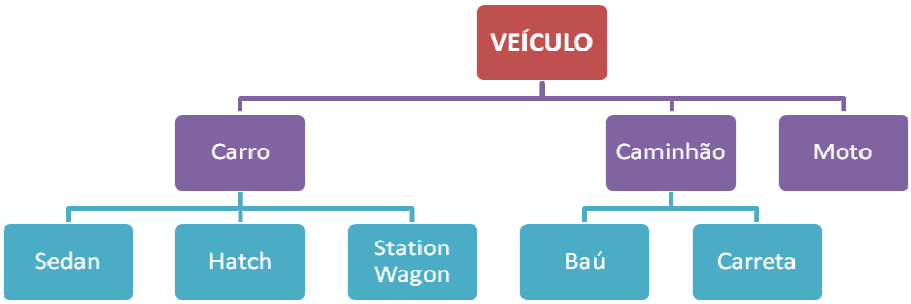


Figura 6 – Exemplo de uma taxonomia simples

Também estão incluídas nessa categoria, outros tipos de cálculo do grau de similaridade usando o WordNet (http://wordnet.princeton.edu) e outras redes semânticas disponíveis. O WordNet [Miller *et al.* 1990, Miller, 1995, Richardson *et al.* 1994] é um dicionário léxico on-line desenvolvido pelo Laboratório de Ciências Cognitivas da Universidade de Princeton. Algumas funções descritas no Capítulo 4 fazem uso do WordNet para realizar suas medições. O WordNet e outras taxonomias similares são vistas como uma estrutura de grafo. As figuras 7 e 8 são fragmentos da WordNet. O Relacionamento semântico, neste caso, pode ser obtido usando o tamanho do caminho entre os termos (nós do grafo). “Um nó que tiver o menor

caminho entre outro nó, é mais similar a ele" [Resnik, 1995]. Um exemplo de trabalho que usa essa abordagem esta descrito em [Rada *et al.* 1989], no qual define distância semântica usando o MeSH (Medical Subject Headings – http://www.nlm.nih.gov/mesh), um sistema de indexação de arquivos e um banco de dados da área médica. Nesta abordagem, a distância conceitual entre os termos é medida considerando "o número de ligações entre os termos na hierarquia do MesH".

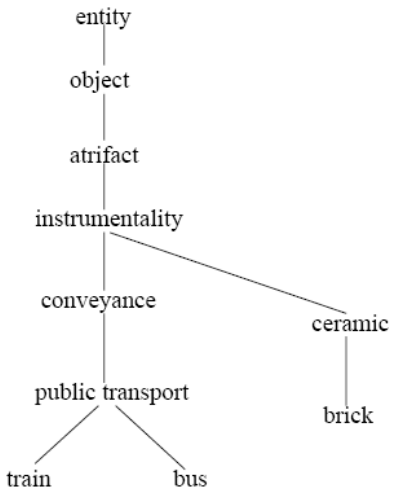


Figura 7 – Fragmento da hierarquia is-a da WordNet. [Petrakis et al. 2006]

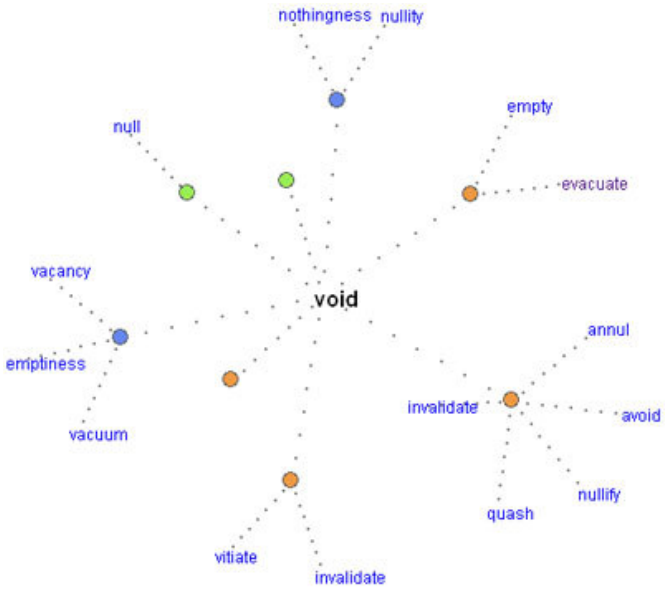


Figura 8 – WordNet visual para o termo "void". [Ajaxian, 2006]

O método de “contar o número de ligações entre os dois termos” para calcular o grau de similaridade é fortemente defendida pela teoria que diz que “as ligações presentes numa taxonomia representam distancias uniformes”. Porém, isto nem sempre é verdade se levarmos em consideração que algumas ligações possuem pesos associados, tornado-as mais “densas” que outras [Resnik, 1995]. Com isso, surgem diversos outros trabalhos que exploram mais as redes de taxonomias a fim de resolver este problema [Sussna, 1993].

3.2 Abordagem baseada no índice de informação compartilhada

Essa abordagem compreende todas as técnicas que basicamente confiam no fato de que, a similaridade semântica entre dois termos pode ser calculada através do grau de informações que eles têm em comum, ou seja, o grau de informações que elas compartilham [Resnik, 1995]. “Quanto mais informações os termos compartilharem, mais similares eles são.” As informações compartilhadas por dois conceitos, A e B, é denotada pelo conteúdo da informação do conceito mais específico segundo a taxonomia. Este tipo de abordagem também leva em consideração a seguinte teoria: “palavras semanticamente similares tem comportamento distributivos semelhantes no *Corpus*[1]” [Resnik, 1995]. De acordo com esta teoria, palavras que co-ocorrem bastante próximas de outra palavra específica são consideradas como sendo “características” ou “propriedades” desta palavra.

Conseqüentemente, relacionamento entre palavras são freqüentemente extraídos de sua co-ocorrência[2] distributiva no *Corpus* [Jiang *et al.* 1997]. Nestes casos, um conjunto de classes de palavras hierárquicas pode ser extraído através de distribuição e agrupamentos. Isto é conseguido percorrendo as classes da taxonomia para descobrir os níveis hierárquicos, classes e subclasses das quais o termo pertence.

[1] ***Corpus (pl. Corpora)*** – Grande e estruturado conjunto de textos usado para análises estatísticas, verificação de ocorrências ou validação de regras lingüísticas num domínio específico [The American Heritage® Dictionary of the English Language, 4ª edição, copyright ©2000. Atualizado em 2003. Publicado por Houghton Mifflin Company].

[2] **Co-ocorrência** – Quando duas ou mais palavras ocorrem simultaneamente num texto ou base de dados.

Exemplos de medidas de similaridade que usam a abordagem baseada no índice de informação compartilhada são: i) [Resnik, 1995] que considera taxonomias do tipo *is-a* e ii) [Jiang *et al.* 1997] método que se baseia na distância dos caminhos e usa estatísticas do Corpus como fator corretivo.

O maior inconveniente encontrado para se calcular similaridades através do índice de informação compartilhada é baseado na determinação de probabilidades dos conceitos usando o *Corpus*. Diferentes *Corpora (plural de Corpus)* podem informar diferentes probabilidades. Além disso, essas medidas são freqüentemente baseadas nos termos e não nos significados dos termos, o que pode causar graves diferenças nos valores obtidos.

3.3 Abordagem baseada em características

Este método não considera a posição real da palavra dentro da taxonomia que está sendo utilizada. A abordagem baseada em características considera o conjunto de informações que se referem à palavra desejada. Sendo assim, quanto mais características os termos têm comum, mas similares eles são. Esta abordagem é baseada justamente no conjunto de palavras que descrevem o termo, chamado tais palavras de "características" ou "propriedades" [Tversky, 1977]. Logo, este método estabelece que duas palavras são semanticamente relacionadas considerando a combinação de características em comum que elas possuem (ou vice-versa).

O trabalho descrito em [Pedersen *et al.* 2005], visto mais detalhado na seção 4.3, introduz o conceito de relacionamento semântico baseado num vetor de contextos extraídos de um *Corpus* da área médica. Um conceito c_1 é representado como sendo um vetor de contextos. Esta abordagem constrói uma matriz de co-ocorrências onde cada célula representa uma pontuação, que é calculada baseada na similaridade entre o termo encontrado na descrição do conceito e cada palavra que co-ocorre no *Corpus*. As linhas da matriz correspondem aos termos usados para descrever o conceito. Os vetores de contextos foram criados baseados em dados precisos em linguagem natural que foram obtidos de frases presentes em diagnósticos da *Clínica Mayo*. Depois que

todos os vetores de contexto são criados, os conceitos são representados por palavras descritivas que são representadas através de uma média de todos os vetores associados com as palavras.

3.4 Abordagem híbrida

Na abordagem híbrida, que é comumente utilizada, se combinam algumas das abordagens descritas nas seções anteriores para se calcular a similaridade semântica entre dois termos “A” e “B”. A maioria delas combina a abordagem do tamanho do caminho que conecta dois temos na estrutura da taxonomia, relacionamentos *is-a* entre os termos e seus nós pais, e a abordagem usando as características entre os termos. Foi observado que alguns trabalhos consideram os métodos de Resnik, Jiang, Conrath e Lin como abordagens híbridas por utilizarem estruturas de ontologias e informações compartilhadas em suas métricas [Nguyen *et al.* 2006]. Porém, na maioria dos trabalhos encontrados na literatura, está sendo usada a classificação que considera os autores mencionados acima enquadrados na abordagem baseada no índice de informação compartilhada.

Rodriguez *et al.* definem uma abordagem híbrida para calcular similaridade semântica entre classes inteiras dentro de uma ontologia simples ou através de várias ontologias mapeadas, aplicadas para obter dados geográficos [Rodriguez *et al.* 2004]. O trabalho de Rodriguez considera relacionamentos *is-a* e características distintivas (como atributos e funções) para determinar a similaridade semântica entre as classes. Este método combina a abordagem baseada em características com a abordagem baseada em distâncias do modelo. Informações contextuais também são utilizadas para estabelecer a importância relativa de certas características distintivas e usar no processo de comparação dos resultados produzidos pela métrica com julgamentos humanos.

4. Funções de Similaridade Semântica

Como já foi citado, um processo de integração de esquemas recebe dois ou mais esquemas como entrada e produz um único esquema como resultado [Rahm *et al.* 2001]. Porém, para que esta tarefa seja realizada, se faz necessário em muitos casos realizar o calculo da similaridade semântica entre termos dos esquemas utilizados.

No processo de integração, este cálculo é indispensável para que os sistemas possam escolher qual o melhor termo a ser utilizado, dentro de uma gama de termos heterogêneos existentes nos esquemas que são recebidos como entrada. O processo de cálculo de similaridade semântica também dá suporte para que as funções se tornem cada vez mais automáticas, com cada vez menos necessidade da intervenção do usuário.

De acordo com o que foi visto no Capítulo 3, as funções que se propõe a calcular a similaridade semântica entre termos geralmente estão enquadradas em alguma classificação, de acordo com o método que os autores utilizam para se chegar a um resultado. Abordamos nas próximas seções, algumas das principais funções já propostas por alguns autores e com resultados reconhecidos pela comunidade acadêmica.

4.1 Função de Caviedes e Cimino

Em 1989, Rada publicou o primeiro trabalho sobre medidas de similaridade semântica através de termos do MeSH utilizando a abordagem de tamanho do menor caminho, conseguindo obter uma ótima medida de distância entre dois termos [Rada *et al*. 1989]. Recentemente Caviedes e Cimino, baseados no trabalho de Rada, introduziram uma medida de distância chamada **CDist**, baseada na menor distância entre dois termos numa árvore de taxonomia [Caviedes *et al.* 2004]. Uma das características estudadas no trabalho de Rada [Rada *et al*. 1989] foram as propriedades que mostram ser, a distância entre os termos, uma métrica válida, a saber:

1. **Propriedade Não-Negativa**

   $d(C_1, C_2) \geq 0$, e $d(C_1, C_2) = 0 \leftrightarrow C_1 = C_2$.
2. **Propriedade da Simetria**

   $d(C_1, C_2) = d(C_2, C_1)$
3. **Propriedade Triangular**

   $d(C_1, C_2) \leq d(C_1, C_3) + d(C_3, C_2)$

Tendo como pré-requisitos essas propriedades, foi possível criar a função **CDist** também como sendo uma métrica válida para cálculos de similaridade semântica entre termos, introduzindo o uso de relacionamentos *is-a* da hierarquia da taxonomia para se obter os resultados. Caviedes e Cimino também ampliaram o conceito de Rada permitindo que esta função também trabalhasse com terminologias SNOMED-CT, ICD9CM e MeSH, enquanto que a proposta de Rada trabalha apenas com a MeSH.

4.2 Função de Nguyen e Al-Mubaid

Utilizando como base de estudo diversas funções de similaridade semântica existentes baseadas em ontologias e taxonomias, Nguyen *et al*. (2006) propõem uma função com resultados mais otimizados e próximos da realidade desejada. Os autores também se utilizaram do *framework* disponível pelo Sistema de Linguagem Médica Unificada, a UMLS (Unified Medical Language System) [UMLS, 2008]. A UMLS é um framework que provê uma base de conhecimento bastante rica, desenvolvida para suportar grande quantidades de consultas e pesquisas no domínio da medicina, incluindo mais de 100 terminologias de fontes médicas, como a MeSH [MeSH, 2008], ICD9CM [ICD9CM, 2003] e a SNOMED-CT [SNOMED, 1993], além de possuir uma grande variedade de semântica nativas e estruturas sintáticas [Caviedes *et al.* 2004]. Dentro da UMLS, os autores usam mais especificamente as terminologias MeSH e SNOMED-CT.

A função de cálculo de similaridade semântica proposta por Nguyem *et al.* se baseia em ontologias e considera a profundidade de cada nó na hierarquia da taxonomia como uma boa distância (tamanho do caminho) entre eles. Para calcular a distância da similaridade semântica entre dois conceitos, a função recupera a profundidade de todos os nós menos comuns abaixo deles (lcs) e a distância do menor caminho entre eles. A função atribui um alto valor de similaridade quando os dois conceitos estão no nível mais baixo da hierarquia [Nguyen *et al.* 2006]. A medida da similaridade é calculada da seguinte forma:

$$\boldsymbol{Sim(C_1, C_2)} = log_2 ([l(C_1, C_2) - 1) \times [D - depth(lcs(C_1, C_2))] + 2)$$

Onde $l(C_1, C_2)$ é a menor distância entre C_1 e C_2, $depth(lcs(C_1, C_2))$ é a profundidade do $lcs(C_1, C_2)$ usando a contagem de nós e finalmente o $lcs(C_1, C_2)$ são todos os mais baixos nós abaixo de C_1 e C_2. Além disso, ainda temos D como sendo a profundidade máxima da taxonomia. Na árvore MeSH por exemplo, D é igual

a 12 quando nós adicionamos um nó raiz para conectar todas as 15 categorias de árvores, e o mínimo valor da distância é igual a 1.

Quando dois conceitos estão no mesmo conjunto de nós (*cluster*) ou quando a distância é igual a 1 usando a contagem dos nós, atribuímos o valor da similaridade igual a 1. O máximo valor da medida de distância ocorre quando um nó se encontra na posição mais afastada à direta da árvore e o outro nó encontra-se na posição à esquerda mais afastada. Na terminologia MeSH, por exemplo, o máximo valor desta medida é igual a:

$$\boldsymbol{D(max)} = log_2 ([23 - 1) \times [12 - 1] + 2) = \boldsymbol{7{,}9307}$$

Sendo assim, podemos concluir que os valores de distância na terminologia MeSH encontram-se no intervalo [1,0000, 7,9307]. A técnica de usar o tamanho do caminho considera somente a distância entre dois conceitos. Conseqüentemente, quando dois pares de conceitos têm a mesma distância do caminho, eles possuem o mesmo valor de similaridade semântica. Por exemplo, de acordo com esse método, na Figura 9, a similaridade entre (n_1, n_5) é igual à similaridade de (n_2, n_4). Mas, na realidade, a similaridade entre (n_2, n_4) deve ser bem maior do que (n_1, n_5), já que os nós (conceitos) n_2 e n_4 compartilham mais informações (mesmos nós estão conectados com o outro) e atributos do que n_1 e n_5.

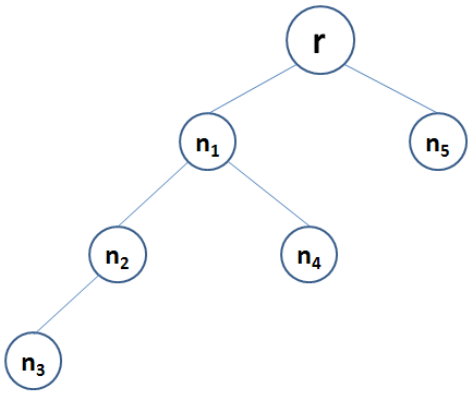


Figura 9 – Árvore de hierarquia entre seis conceitos.

Agora consideremos um exemplo dentro da terminologia MeSH. A categoria 70ª da árvore de categorias é a "Biological Science" e é representada pela letra G na versão 2006 da terminologia MeSH. Uma parte da árvore pode ser vista a seguir:

- **Biological Sciences [G]**
  - Biological Sciences [G01]+
  - Health Occupations [G02]+
  - Environment and Public Health [G03]+
  - ……..
  - Genetic Phenomena [G13] +
  - Genetic Structures [G14] +

O símbolo "+" indica que o conceito pode ser expandido e possui sub-conceitos associados a ele. Por exemplo, *Biological Sciences [G01]* pode ser expandido e ser visto da seguinte forma:

- **Biological Sciences [G01]**
  - Anatomy [G01.100]+
  - Biochemistry [G01.201]+
  - Biology [G01.273]+
  - Biophysics [G01.344] +
  - Biotechnology [G01.550] +
  - Neurosciences [G01.610] +
  - Pharmacology [G01.703] +
  - Physiology [G01.782] +

Neste exemplo, fica claro que a similaridade entre Biological Sciences [G01] e Environment and Public Health [G03] é menor que a similaridade entre Biology [G01.273] e Biotechnology [G01.550]. Entretanto, o método comum baseado no tamanho do caminho atribui a mesma similaridade para os dois pares citados. De acordo com os cálculos, o método baseado no tamanho do caminho calcula a mesma distância **0.33** para os dois pares, enquanto que o método proposto por Nguyem e Al-Mubaid obtém o valor de distância **4.32** para o par [G01, G03] e o valor **4.17** para o par [G01.273, G01.550].

4.3 Função de Pedersen

A proposta sugerida no trabalho de Pedersen [Pedersen *et al.* 2005] é uma função que calcula similaridade semântica baseada em vetores de contexto que são extraído de dicionários da área médica (Corpora[3]) da língua inglesa. Para realizar os testes, os autores utilizaram a terminologia SNOMED-CT disponível na UMLS, pela grande quantidade de ontologias disponíveis nesta fonte. Também foi utilizado o conceito de pares, presente na hierarquia *is-a* da SNOMED-CT, sempre que as medidas utilizadas necessitassem disto, porém, em geral, as medidas utilizadas são mais abrangentes e não necessitavam fazer esta adaptação.

O Corpus utilizado para os testes e desenvolvimento da função foi extraído de aproximadamente 1.000.000 de notas e diagnósticos médicos da Clínica Mayo. Diagnósticos médicos são ricos por possuírem uma grande quantidade de palavras, termos e caracteres específicos que não são facilmente encontrados em outras fontes mais gerais. Nas clínicas e hospitais dos EUA, este tipo de documento é obrigatório e todos os médicos geram boa quantidade deste documento por dia. Como resultado, obtemos diversos relatórios espontâneos (em linguagem natural) com grande quantidade de termos relevantes.

[3] **Corpora –** Plural de Corpus. Definição na seção 3.2.

****CC****
Review recent progress.
****CM****
Aspirin 81 mg q.d.
Imdur 30 mg q.d.
Lisinopril 5 mg q.d. (increased to 10 mg q.d. today)
****HPI****
Her vocal cord examination yesterday was unremarkable. She broke her ankle toward the end of YEAR and is still limping but it is improving. While she was hospitalized for aspiration pneumonia after her vocal cord biopsy in DATE, she developed tachycardia with ECG changes. Echocardiogram showed EF of 30-35% with regional wall motion abnormalities. She was started on Lisinopril and Imdur.
****IP****
#1 Probable CAD
#2 ASO
Plan: Because of some elevated blood pressure, we will increase her Lisinopril to 10 mg q.d.
****SI****
DISM 1/13/99
****DX****
#1 Probable CAD
#2 ASO

Figura 10 – Parte de um típico diagnóstico [Pedersen et al. 2005].

Os diagnósticos da Clínica Mayo são armazenados diretamente no prontuário eletrônico do paciente, fato que facilita o arquivamento digital dos documentos. Outro fator importante na extração dos dados se deve ao fato destes documentos serem semi-estruturados com algumas subseções bem definidas em cada diagnóstico. Um exemplo típico deste tipo de diagnóstico pode ser observado na Figura 10. Cada subseção é representada por uma sigla (exemplo: *SI – Special Instructions, CM - Current Medications*, entre outros) e algumas delas foram bastante utilizadas pelo processo proposto pelos autores, dependo do caso que se estivesse estudando.

Como já foi citado, a função proposta se baseia em vetores de contexto, ou seja, cada conceito é representado por seu respectivo vetor. Esta abordagem é considerada mais flexível pelo fato de não necessitar que os conceitos estejam necessariamente relacionados em alguma taxonomia específica de uma ontologia. Tal abordagem proposta por Pedersen já é considerada uma adaptação da que foi proposta em 1998 por Schütze's [Schütze, 1998].

Esta abordagem constrói uma matriz de co-ocorrências onde cada célula representa uma pontuação, a qual é calculada baseada na similaridade entre o termo

encontrado na descrição do conceito e cada palavra que co-ocorre no *Corpus*. As linhas da matriz correspondem aos termos usados para descrever o conceito, enquanto que as colunas são as palavras que mais ocorrem no *Corpus*. Os vetores de contexto criados são resultados de mais de dez anos de dados coletados da Clínica Mayo, com informações ricas, diversificadas e com diferentes níveis de *expertise* dos usuários. Esta base de dados possui aproximadamente 16 milhões de frases únicas expressas em linguagem natural, o que representa mais de 21 mil diagnósticos. Toda essa informação foi classificada pelo sistema HICDA [4]em basicamente quatro níveis. O nível mais abrangente possui 19 categorias como *Neoplams*, *Diseases of the Circulatory System*, entre outros. Os outros três níveis são grupos de diagnósticos mais específicos.

A base de informações da Clínica Mayo foi construída assumindo que a maioria das frases de diagnóstico foram classificadas na mesma categoria na hierarquia do HICDA. Sendo assim, essas frases podem ser consideradas sinônimos no nível de granularidade obtido do HICDA. Depois de todo o processo de se retirar ambigüidades e palavras repetidas ou com significados semelhantes, foi realizada a junção de toda a informação os dados da UMLS (terminologia SNOMED). O resultado foi a obtenção de 3.665.721 frases de diagnósticos organizadas dentro de 594.699 agrupamentos (*clusters*). Todo esse conjunto de informações formam o ***thesaurus*** que será utilizado no cálculo da similaridade, representado cerca de 95% dos conceitos do SNOMED-CT.

Então, para representar os conceitos que ocorrem tanto no *thesaurus* como no SNOMED-CT para as medidas de similaridade semântica, são obtidos os termos descritivos no thesaurus e construída uma matriz de co-ocorrências. Logo após o processo de criação dos vetores, os conceitos representados por termos descritivos são então refeitos através da média de todos os vetores associados com todas as palavras descritivas.

[4] HICDA – Hospital International Classification of Diases Adaptation

4.4 Função de Petrakis

Em seu trabalho, [Petrakis *et al.* 2006] implementou algumas das principais abordagens de cálculo de similaridade semântica existentes até então para serem estudadas e seus resultados analisados. Porém, o foco do seu trabalho foi desenvolver um método que permite trabalhar com "ontologias cruzadas" que é capaz de calcular a similaridade entre termos, mesmo que estes estejam em ontologias diferentes. Para realização de testes, foram utilizadas as ontologias WordNet e MeSH.

O desenvolvimento desta função tem especial motivação por ser este um dos mais difíceis problemas (trabalhar com ontologias cruzadas) de serem resolvidos e não tão estudados na literatura existente. Os autores chamaram sua função de **X-Similarity**.

Um dos principais trabalhos analisados foi inicialmente proposto por Rodriguez [Rodriguez *et al.* 2003], que consiste num framework que possibilita a comparação de termos da mesma ou de diferentes ontologias. A similaridade entre os termos *a* e *b* é calculada como a soma dos pesos das similaridades entre os conjuntos de sinônimos do WordNet (*synsets*), características e vizinhança:

$$\boldsymbol{Sim_{(a,\ b)}} = w \cdot S_{synsets}(a, b) + u \cdot S_{features}(a, b) + v \cdot S_{neighborhoods}(a, b) \qquad (1)$$

onde w, u e v denotam a importância relativa dos três componentes de similaridade. Características são classificadas como sendo partes, atributos ou funções associados ao termo. Por exemplo, no WordNet, o conjunto $S_{features}$ é caracterizado como sendo os relacionamentos "*Part-Of*" dos termos. Assumindo que todos os termos da vizinhança de a e b também possuem suas características (partes, atributos e funções), estes termos também podem ser representados por *synsets*, onde cada componente de similaridade é calculado da seguinte maneira [Tversky 1997]:

$$\frac{|A \cap B|}{|A \cap B| + \gamma(a,b)|A \setminus B| + (1-\gamma(a,b))|B \setminus A|} \quad (2)$$

onde A e B denotam os *synsets* dos termos *a* e *b* e A\B o conjunto de termos que estão em A mas não estão em B (o contrário para B\A). O parâmetro $\gamma(a, b)$ é calculado através de uma função de profundidade dos termos *a* e *b* na sua taxonomia:

$$\gamma(a,b) = \begin{cases} \dfrac{depth(a)}{depth(a)+depth(b)}, & depth(a) \le depth(b); \\ 1 - \dfrac{depth(a)}{depth(a)+depth(b)}, & depth(a) > depth(b), \end{cases} \quad (3)$$

A função proposta *X-Similarity* realiza comparações entre os *synsets* e o conjunto de descrições dos termos. Dois termos são se seus *synsets* ou seu conjunto de descrições ou, os *synsets* dos termos da suas vizinhanças são lexicamente similares. Os autores ainda propõem substituir a equação (2) por um conjunto de similaridades plana:

$$S(a,b) = \frac{|A \cap B|}{|A \cup B|}, \quad (4)$$

onde A e B denotam *synsets* ou conjunto de descrições do termo. Como nem todos os termos na vizinhança de um termo estão conectados pelos mesmos relacionamentos, o autor sugere que os conjuntos de similaridades sejam calculados pelo tipo do relacionamento (i.e, *is-a* e *part-of* para WordNet e somente *is-a* para MeSH):

$$S_{neighborhoods}(a,b) = \max \frac{|A_i \cap B_i|}{|A_i \cup B_i|}, \quad (5)$$

Onde *i* representa o tipo de relacionamento. A equação acima sugere um cálculo de similaridade entre termos da vizinhança pela comparação de mesmos tipos de

relacionamentos entre os *synsets* dos mais específicos e dos termos mais abrangentes. Finalmente, todas as idéias e princípios sugeridos podem ser resumidos na equação abaixo:

$$Sim(a,b) = \begin{cases} 1, & if\ S_{synsets}(a,b) > 0; \\ \max\{S_{neighborhoods}(a,b), S_{descriptions}(a,b)\}, & if\ S_{synsets}(a,b) = 0. \end{cases} \quad (6)$$

$S_{descriptions}$ representa a comparação dos conjuntos de termos. $S_{descriptions}$ e $S_{synsets}$ são calculados de acordo com a equação 4. Com isso, também é possível notar que dois termos com mais sinônimos em comum são 100% similares. É importante ressaltar que tanto o método proposto por Rodriguez [Rodriguez *et al.* 2003] quanto o método de Petrakis podem ser usados para calcular similaridade entre termos de uma mesma ontologia.

5. Comparativo das Funções

No Capítulo 4, foram analisados os métodos de algumas das funções de cálculo de similaridade semântica existentes na literatura e já utilizadas. Segue a baixo um resumo das principais diferenças e comparações entre as funções analisadas neste trabalho:

1.

AUTORES: Caviedes e Cimino

Medida: Tamanho do menor caminho;

Terminologia fonte: MeSH, SNOMED-CT, ICD9CM;

Vantagens: Simplicidade; Dá suporte a grande número de terminologias fontes da UMLS; Usa relacionamentos *is-a*.

Desvantagens: Produz poucos resultados significativos. Não calcula a profundidade ou diferentes densidades na taxonomia.

2.

AUTOR: Pedersen

Medida: Vetor de contexto; Medida baseada em *Corpus*.

Terminologia fonte: SNOMED-CT;

Vantagens: Função não depende da estrutura da taxonomia;

Desvantagens: É necessário pelo menos um *Corpus* para obter informações estatísticas.

3.

AUTOR: Petrakis

Medida: Baseada em características;

Terminologia fonte: MeSH;

Vantagens: Função não depende da estrutura da taxonomia.

Desvantagens: Necessita realizar a extração das características previamente para estimar a similaridade semântica.

4.

AUTORES: Nguyen e Al-Mubaid

Medida: Tamanho do caminho; Profundidade dos nós.

Terminologia fonte: MeSH, SNOMED-CT;

Vantagens: Trabalha com a especificidade de conceitos em comum; Usa profundidade dos nós e sua granularidade;

Desvantagens: Necessário pré-processamento das informações antes do calculo da similaridade.

Através do estudo e das comparações realizadas com as funções abordadas, devemos levar também em consideração a forma de funcionamento do sistema em que utilizaremos a função, antes de escolhermos qual a melhor função a ser utilizada. Como foi visto nos capítulos anteriores, este trabalho se propôs a escolher a função para cálculo da similaridade semântica mais adequada para o processo de integração de esquemas do Integra [Belian, 2007].

O processo de integração de esquemas realiza a aquisição de informações semânticas utilizando um dicionário ou terminologia. Ou seja, sistemas como a UMLS são acessados para obter os dados e termos necessários nas etapas iniciais e então armazenados dentro de uma estrutura interna própria do Integra. Tal estrutura é baseada em ontologias que armazenam os valores obtidos em instâncias de classes dentro da própria ontologia. O processo de comparação da similaridade entre os termos é então realizado navegando-se na estrutura da ontologia do Integra, através de mecanismos próprios ou usando APIs disponíveis para tal, como por exemplo, a existente no software Protégé [Protégé 2007].

Levando em consideração como ponto importante a arquitetura já definida do Integra, a função que se apresentou melhor enquadrada para ser aplicada ao sistema é a que foi proposta por Petrakis [Petrakis *et al.* 2006], principalmente pela vantagem dela ser independente da estrutura da taxonomia. A ontologia de contextos do Integra não reproduz a taxonomia dos vocabulários que definem o significado dos termos utilizados. Na realidade, esta ontologia recupera as informações semânticas

necessárias e incorpora na sua estrutura interna de conceitos [Belian, 2007]. Por este motivo, é bastante viável e interessante realizar a comparação dos termos definidos na sua própria ontologia, considerando suas características, como atributos, funções.

6. Conclusões

O processo de integração de esquemas ainda têm sido um desafio para diversos pesquisadores, principalmente no que se refere a tornar o processo com menos intervenção humana. Neste trabalho abordamos diversas técnicas e medidas com resultados já comprovados pela comunidade acadêmica, porém muitas delas, dependendo do contexto que serão utilizadas, têm sua aplicação bastante limitada, sendo necessário muitas vezes, artifícios, considerações e fusões com outras técnicas para se obter um resultado satisfatório.

Também verificamos que a quantidade de informações distribuídas na Web que ainda se encontram desestruturadas é muito grande, o que vêm motivar ainda mais para que os métodos e técnicas de integração de dados também levem em consideração, cada vez mais, dados não estruturados. Os métodos estudados muitas vezes se encontravam bastante dependentes de estruturas e taxonomias para funcionar corretamente.

Porém, o objetivo maior deste trabalho, que era de analisar, estudar e escolher uma função de cálculo de similaridade semântica para o processo de integração de esquemas do Integra foi bem sucedido, uma vez que encontramos uma técnica aceitável que se enquadrasse nos requisitos necessários para conclusão da tarefa de integração. Segue como sugestão de trabalhos futuros, implementar a função escolhida neste trabalho (de Petrakis) dentro do algoritmo do Integra, para que se possa gerar resultados para estudos de eficiência e de precisão dos resultados produzidos.

Referências Bibliográficas

[Ajaxian, 2006] **Ajaxian**, 2007. Trying to generate more hype than Rails. Disponível em <http://ajaxian.com/archives/visual-wordnet>. Acesso em 29 janeiro 2008.

[Belian, 2007] Belian, R..: **A Context-based Name Resolution Approach for Semantic Schema Integration**, PHD Thesis proposal, Center of Informatics, Federal University of Pernambuco, Brazil, 2007.

[Berners, 2001] Berners-Lee, T., Hendler, J., Lassila, O.: **The Semantic Web**, Scientific America184, no. 5, pp. 34-43, 2001.

[Caviedes *et al.* 2004] Caviedes, J., Cimino, J.: **Towards the development of a conceptual distance metric for the UMLS**. Journal of Biomedical Informatics, 37: 77-85, 2004.

[Chen, 1976] Chen, P.: **The entity-relationship model - toward a unified view of data**. In ACM Transactions on Database Systems, (1), 1976.

[Costa, 2005] Costa, T.A. **O Gerenciador de Consultas de um Sistema de Integração de Dados**, Tese de mestrado, CIn – UFPE, 2005.

[Elis, 2007] Elis, Diego. A Web Semântica. Disponível em < http://www.tableless.com.br/a-web-semantica/>. Acesso em 01 outubro 2007.

[Elmasri *et al.* 2000] Elmasri, R., Navathe, S.: **Fundamentals of Database Systems**. Addison-Wesley, Third Edition, 2000.

[Goh, 1997] 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.

[Gruber, 1993] Gruber, T.: **A Translation Approach to Portable Ontologies**. Knowledge Acquisition, V.5, n.2, p.199-200, 1993.

[ICD9CM, 2003] International Classification of Diseases, ninth revision, Clinical Modifications, 2003. Disponível em < http://icd9cm.chrisendres.com/>. Acesso em 29 janeiro 2008.

[Jiang *et al.* 1997] Jiang, J., Conrath, D.: **Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy**. In Proceedings of The International Conference Research on Computational Linguistics, Taiwan, 1997.

[Kashyap, 1996] 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.

[Lin, 2000] Lin, Dekang.: **An Information-Theoretic Definition of Similarity**. Departament of Computer Science, University of Manitoba, Canada, 2000.

[Lóscio, 2003] Lóscio, B.: **Managing the Evolution of XML-based Mediation Queries**. PHD Thesis, Federal University of Pernambuco, Brazil, 2003.

[Lóscio *et al.* 2003] Lóscio, B., Salgado, A., Galvão, L.: **Conceptual Modeling of XML Schemas**. In Proceedings of the International Conference on Conceptual Modeling ER, 2003.

[McGill *et al.* 1979] McGill *et al.*, M. (1979). **An evaluation of factors affecting document ranking by information retrieval systems**. Project report, Syracuse University School of Information Studies.

[MeSH, 1999] Medical Subject Headings, 1999. Disponível em < http://www.nlm.nih.gov/mesh/>. Acesso em 28 janeiro 2008.

[Miller *et al.* 1990] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J. : **Introduction to WordNet : An on-line Lexical Database**. International Journal of Lexicography, 3(4): 235-244, 1990.

[Miller 1995] Miller, G.: WordNet: **A lexical database for English**. Communications of the ACM, 38(11), 1995.

[Nguyen *et al.* 2006] Nguyen, H., Al-Mubaid, H.: **New Ontology-based Semantic Similarity Measure for the Biomedical Domain**. IEEE conference on Granular Computing GrC-2006, pp. 623-628, 2006.

[Ouksel, 1999] Ouksel, A., Sheth, A.: **Semantic Interoperability in Global Information Systems**.

[Pedersen *et al.* 2005] Pedersen, T., Pakhomov, S., Patwardhan, S.: **Measures of semantic similarity and relatedness in the medical domain**. University of Minnesota Digital Technology Center Research Report DTC 2005/12, 2005.

[Petrakis *et al.* 2006] Petrakis, E., Varelas, G., Hliaoutakis, A., Raftopoulou, P. : **Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies**. 4th Workshop on Multimedia Semantics (WMS'06), pp. 44-52, 2006.

[Protégé 2007] Protégé ontology editor and knowledge-base framework. Stanford Medical Informatics. Disponível em <http://protege.stanford.edu/index.html>. Acesso em 28 dezembro 2007.

[Rada *et al.* 1989] Rada, R., Mili, H., Bicknell, E., Blettner, M.: **Development and application of a metric on semantic nets**. IEEE Transactions on systems, man and cybernetics. 19 (1):17-30, 1989.

[Rahm *et al.* 2001] Rahm, E., Bernstein, P.: **A survey of approaches to automatic schema matching**. In The VLDB Journal, (10) 334-350, 2001.

[Resnik, 1995] Resnik, P.: **Using information content to evaluate semantic similarity**. In proceedings of the 14th International Joint Conference on Artificial Intelligence, pp 448-453, Canada, 1995.

[Richardson *et al.* 1994] Richardson, R., Smeaton, A., Murphy, J.: **Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words**. Technical Report CA-1294, Dublin City University, School of Computer Applications, 1994.

[Rodriguez *et al.* 1999] Rodriguez, M., Egenhofer, M., Rugg, R.: **Assessing semantic similarities among geo-spatial feature class definitions**. In Interoperating geographic information systems, LNCS (1580) 189-202, Springer-Verlag, March, 1999.

[Rodriguez *et al.* 2003] Rodriguez M.A. and Egenhofer M.J. (2003). **Determining Semantic Similarity among Entity Classes from Different Ontologies**. IEEE Trans. on Knowledge and Data Engineering, 15(2), 442-456.

[Rodriguez *et al.* 2004] Rodriguez, M., Egenhofer, M., Rugg, R.: **Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure**. International Journal of Geographical Information Science, 18(3): 229-256, 2004.

[Schütze, 1998] Schütze, H. (1998). Automatic Word Sense Discrimination; Computational Linguistics; 24 (1): 97-123.

[SNOMED, 1993] SNOMED-CT: **Systematized Nomenclature of Medicine, Clinical Terminology**, 1993 Disponível em <http://www.snomed.org/ >. Acesso em 29 janeiro 2008.

[Sussna, 1993] Sussna, M.: **Word Sense disambiguation for free-text indexing using a massive semantic network**. Proceedings of the Second International Conference on Information and Knowledge Management, pp 67-74, 1993.

[Tversky, 1997] Tversky, A.: **Features of similarity**. Psycological review, 84(4):327-352, 1997.

[UMLS, 2008] **Unified Medical Language System**, 2008. Disponível em: <http://www.nlm.nih.gov/research/umls/about_umls.html>. Acesso em: 02 janeiro 2008.

[Bray, 1999] Bray, T., Paoli, J., Sperberg-Mcqueen, C. M. **Extensible Markup Language (XML) 1.0**, World Wide Web Consortium.

Disponível em http://www.w3.org/TR/REC-xml. Acesso em: 20 novembro de 2007.

[Wache, 2001] 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.

[Wang, 2005] Wang, Y.: **An Empirical Evaluation of Semantic Similarity Measures Using the WordNet and UMLS Ontologies**. Master of Computer Science thesis, Miami University, Oxford, Ohio, 2005.

Assinaturas

_____

Ana Carolina Salgado
Orientadora

_____

Rosalie Barreto Belian
Co-orientadora

_____

Daniel Ferreira da Silva
Aluno