Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática

Distância Semântica entre Mapas Conceptuais:

Uma Abordagem Experimental

Ana Cristina da Costa Oliveira Alves

Coimbra

Março de 2004

Dissertação submetida à
Universidade de Coimbra
para obtenção do grau de
Mestre em Engenharia Informática
Área de Especialização em Sistemas e Tecnologias do Conhecimento

Distância Semântica entre Mapas Conceptuais: Uma Abordagem Experimental

Ana Cristina da Costa Oliveira Alves

Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Informática
Março de 2004

Dissertação realizada sob a orientação do
Professor Doutor Fernando Amílcar Bandeira Cardoso
Professor Associado
do Departamento de Engenharia Informática
da Faculdade de Ciências e Tecnologia
da Universidade de Coimbra

Agradecimentos

Quero agradecer ao meu orientador, Prof. Dr. Amilcar Cardoso, pela atenção dispensada ao longo do trabalho efectuado e, especialmente, à minha família pelo apoio e incentivo nos momentos mais difíceis.

Estou grata também ao Engº Franscico Câmara Pereira pelos seus conselhos e orientações que me ajudaram, principalmente, nesta última fase a delinear os traços finais desta Dissertação.

Índice Geral

1. Introdução 1

2. Fundamentos Teóricos 5

2.1. Mapas Conceptuais 5

2.1.1. Origem 5

2.1.2. Utilização 6

2.1.3. Organização dos Mapas: Hierarquia vs. Rede de Conceitos 7

2.1.4. Criação e Dimensão dos Mapas Conceptuais 9

2.1.5. Semantic Web: Uma representação baseada nos mapas conceptuais 10

2.2. Conhecimento Semântico: Taxonomia de conceitos 11

2.2.1. WordNet 11

2.2.1.1. Organização 12
2.2.1.2. Relações Semânticas 14
2.2.1.3. Conhecimento sobre o uso de conceitos 18

2.2.2. Desambiguação dos Significados de Palavras 20

2.2.2.1. Representação Simbólica: Restrição Seleccionadora 21
2.2.2.2. Técnicas Robustas 24

3. Similaridade Semântica 29

3.1. Similaridade ou Distância Semântica entre Conceitos 29

3.1.1. Abordagem baseada na Distância Semântica 30

3.1.2. Abordagem baseada no Conteúdo de Informação 33

3.1.3. Resumo das Medidas Apresentadas 37

3.2. Similaridade entre estruturas conceptuais: Aplicações em diferentes áreas 38

3.2.1. A Similaridade como um Processo Cognitivo: Contraste de Características vs. Alinhamento Estrutural 39

3.2.1.1. Contraste de Características (*Feature-Constrast Model*) 39
3.2.1.2. Alinhamento Estrutural (*Structural Alignment*) 41

3.2.2. Avaliação da Aprendizagem por Mapas Conceptuais 42

3.2.3. Modelo Computacional: Comparação de Grafos Conceptuais 44

4. Propostas de Similaridade Semântica entre Mapas Conceptuais: Uma abordagem experimental 49

4.1. Esquema Geral 49

4.2. Contextualização: Semântica de um Mapa Conceptual 51

4.3. Comparação Semântica: Similaridade entre Mapas Conceptuais 60

4.3.1. Similaridade entre Conceitos 61

4.3.2. Similaridade Global entre Mapas Conceptuais. 72

4.3.2.1. Contraste de Características entre Mapas Conceptuais 76
4.3.2.2. Procura Simples pela Maior Similaridade entre Conceitos 90
4.3.2.3. Procura Ponderada pela Maior Similaridade entre Conceitos 94
4.3.2.4. Conceito Central e Média Ponderada das Similaridades entre Conceitos 96

5. Testes e Resultados 103

5.1. Contextualização 103

5.2. Comparação Semântica 105

5.2.1. Contraste de Características 106

5.2.2. Abordagens baseadas na similaridade entre conceitos 108

6. Conclusões e Trabalho Futuro 113

7. Referências 117

Anexos A.1

A.1 Introdução à Teoria dos Grafos A.1

A.2 Mapas Concepuais Utilizados A.11

A.3 Código Desenvolvido A.31

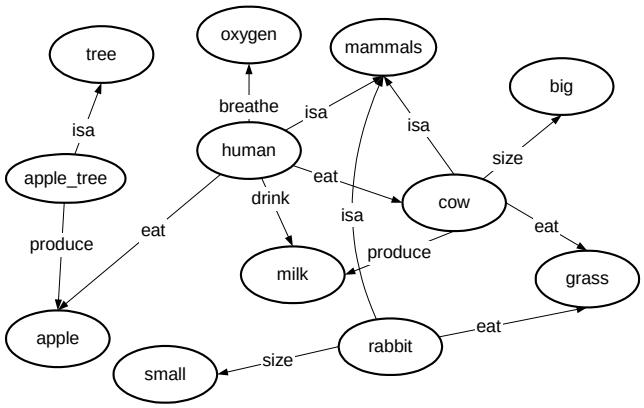
Índice de Figuras

Figura 1.1 ..... *1*
Figura 2.1 ..... *7*
Figura 2.2 ..... *8*
Figura 2.3 ..... *13*
Figura 2.4 ..... *23*
Figura 3.1 ..... *35*
Figura 3.2 ..... *36*
Figura 3.3 ..... *45*
Figura 3.4 ..... *46*
Figura 4.1 ..... *51*
Figura 4.2 ..... *55*
Figura 4.3 ..... *63*
Figura 4.4 ..... *66*
Figura 4.5 ..... *67*
Figura 4.6 ..... *74*
Figura 4.7 ..... *74*
Figura 5.1 ..... *104*
Figura 5.2 ..... *107*
Figura 5.3 ..... *107*
Figura 5.4 ..... *108*
Figura 5.5 ..... *109*
Figura 5.6 ..... *110*
Figura 5.7 ..... *110*

1. Introdução

A representação do conhecimento tem sido uma área de grande investimento científico nos últimos anos devido ao aumento exponencial de informação electrónica sobre diversas áreas do conhecimento. Depois do advento da Internet, a oferta de dados tem sido cada vez mais vasta e numerosa, ao ponto de ser quase impossível actualmente de uma forma rápida e precisa encontrar realmente aquilo que procuramos. Visto ser um conhecimento muitas vezes textual e pouco normalizado, o que impede a sua catalogação e disponibilização directa por bases de conhecimento, é cada vez mais necessário encontrar uma forma de oferecer esta informação de uma forma eficaz e precisa através de meios de representação intuitivos e automáticos.

Este estudo é o seguimento de um trabalho sobre representação do conhecimento e aprendizagem que tem sido desenvolvido no Laboratório de Inteligência Artificial do CISUC, cujo objectivo principal é construir um sistema inteligente que disponibilize e enriqueça conhecimento utilizando como ferramenta mapas conceptuais. Um mapa conceptual é, basicamente, um grafo dirigido onde os nós são conceitos interligados por relações (arcos) que representam um dado domínio. Na Figura 1.1 é apresentado um exemplo de mapa conceptual.



isa(rabbit,mammals).	eat(human, cow).
isa(cow, mammals).	eat(cow, grass).
isa(human, mammals).	eat(rabbit, grass)
isa(apple_tree, tree).	eat(human, apple).
produce(apple_tree, apple)	drink(human, milk).
produce(cow, milk).	size(rabbit, small).
breathe(human, oxygen).	size(cow, big).

Figura 1.1: Mapa conceptual representando uma parte da cadeia alimentar.

O sistema em desenvolvimento, sobre o qual assenta o trabalho aqui descrito, é composto por dois módulos principais: um de extracção de conhecimento denominado *TextStorm* [Alves et. al, 2001] e um outro de aprendizagem dedutiva e interactiva denominado *Clouds* [Pereira et. al, 2000]. O primeiro destina-se a criar mapas conceptuais a partir de informação textual; mais concretamente, o *TextStorm* interpreta um documento fornecido criando automaticamente o

mapa aproximado do domínio representado. Posteriormente, o *Clouds* irá, interactivamente, deduzir novas relações e conceitos através de dedução lógica.

Neste trabalho procuramos estabelecer uma medida de comparação entre mapas conceptuais de diferentes domínios, a fim de verificar a proximidade ou distância dos seus significados. Propomos um mapeamento entre mapas conceptuais baseado na sua semântica comum, com um reduzido custo de processamento computacional, e que se assemelhe ao julgamento humano com determinado grau de confiança. Para tal, será feito um estudo exaustivo através de uma bateria de testes de medidas semânticas já existentes e analisaremos a sua aplicabilidade aos Mapas Conceptuais. Iremos, também, adaptar algumas destas medidas e verificar o seu desempenho face a um conjunto variado de exemplos de mapas conceptuais recolhidos de diferentes fontes.

A Similaridade é um conceito cognitivo amplamente aplicado em diversas áreas humanas e tecnológicas. Na psicologia, William James [1890] já referia a importância da nossa capacidade intrínseca de julgar a proximidade de algo: *"This sense of Sameness is the very keel and backbone of our thinking"*. E de facto os humanos são os mais hábeis a resolver problemas actuais recorrendo a situações semelhantes ocorridas no passado. Esta capacidade cognitiva é alvo de constantes modelações pela ciência de forma a tornar o processo mais independente possível do julgamento humano: a Categorização depende da similaridade dos objectos a classificar; a busca de informação *(Information Retrieval*) baseia-se na similaridade da questão formulada a dados armazenados em memória; o Raciocínio Indutivo é fundamentado no princípio de que se um facto é similar a um anterior, então é previsível para acontecimentos futuros semelhantes.

Os mapas conceptuais são interpretados com recurso a uma base de conhecimento semântico (como veremos de seguida), para posteriormente serem relacionados através de uma medida relativa de similaridade. Decidimos valorizar a eficiência do sistema ao propor algoritmos com uma complexidade reduzida (tanto a nível temporal como em armazenamento) a fim de garantir a sua aplicabilidade em larga escala. Desta forma, não foi, apesar de ponderada, feita qualquer experimentação ou proposta de algoritmos que se baseassem no processamento de isomorfismo entre grafos.

Uma vez que o próprio conceito de similaridade é subjectivo e dependente do contexto apresentado, a atribuição de um valor absoluto de similaridade ente conceitos (ou mapas conceptuais) não nos parece viável na medida em que a mesma pode variar quando estamos perante um conjunto mais abrangente de objectos a relacionar. Desta forma, o resultado esperado consiste em conseguir relacionar semanticamente os mapas conceptuais dentro de um dado

universo, a fim de ser possível dizer que dado mapa A, B, C, D e E, por exemplo, B é mais próximo de A do que C. Ou ainda, dizer quais os mapas mais próximos de A criando um ordenamento decrescente quanto à similaridade semântica em relação a este.

A utilização de uma medida de similaridade semântica sobre objectos estruturados (como é o caso dos mapas conceptuais) que beneficie de uma reduzida complexidade computacional e de uma eficácia razoável poderá vir a enriquecer diversos mecanismos que se fundamentem na comparação: indexação e pesquisa inteligente de arquivos, pesquisa de casos antigos mais similares semanticamente no raciocínio baseado em casos, classificação de conteúdos, entre outros.

Devido a uma extensa pesquisa e estudo na área de Processamento Semântico em que este trabalho se baseia, as medidas aqui propostas poderão ainda servir de base para experiências futuras que utilizem outras abordagens na comparação de estruturas conceptuais. Perante os resultados obtidos, iremos averiguar se o facto de se considerar como principal objecto de análise no processo de comparação os conceitos, e consequentemente a semântica, em detrimento da estrutura sintáctica, irá influenciar o seu desempenho e a aplicabilidade em aplicações reais.

De forma a contextualizar o leitor nos domínios onde este trabalho se encontra inserido, no Capítulo 2, é dada uma introdução aos conceitos básicos necessários para uma melhor compreensão dos Mapas Conceptuais e o conhecimento por eles representados. Posteriormente, no Capítulo 3, é feito um estudo exaustivo sobre a investigação actual no domínio da Similaridade, tanto a nível mais específico (entre conceitos) como de forma mais abrangente a estruturas mais complexas. No Capítulo 4, é apresentado o sistema de comparação semântica, assim como, os resultados obtidos (no Capítulo 5). Finalmente, o Capítulo 6 apresenta os melhoramentos e possível trabalho posterior a ser desenvolvido, incluindo o balanço global do trabalho desenvolvido.

2. Fundamentos Teóricos

A representação do conhecimento tem como principal objectivo codificar os dados de uma forma facilmente manuseável e compreensível pela máquina. A necessidade de uma linguagem de representação do conhecimento de âmbito geral para utilização na resolução de qualquer tipo de problema foi primeiramente levantada por [Quillian, 1968]. Como uma evolução deste pensamento, surgem os mapas conceptuais como uma linguagem visual e informal para representar factos. Neste Capítulo, procuraremos descrever os principais aspectos dos mapas conceptuais como linguagem de representação de conhecimento, tanto a nível semântico como sintáctico.

1. Mapas Conceptuais

A utilização de mapas conceptuais é uma técnica visual de representação do conhecimento baseada em grafos, cujos nós são interpretados como conceitos e os arcos como relações entre estes conceitos. Nos mapas conceptuais, os nós e, eventualmente, os arcos são etiquetados, isto é, são identificados por palavras que exprimem significado relacionado ao conceito/relação. O nome "mapa" poderia estabelecer uma analogia com os mapas genéricos aos quais estamos habituados (geográficos, rodoviários, etc.), no entanto, neste caso em particular, os conceitos não são pontos espacialmente absolutos em relação a um referencial, nem tão pouco as ligações entre estes conceitos representam distâncias mensuráveis quantitativamente. As ligações existentes num mapa conceptual procuram relacionar os conceitos (ou entidades conceptuais) na forma de proposições.

Os mapas conceptuais foram primeiramente utilizados como técnicas de aprendizagem na área das Ciências da Educação com o objectivo de facilitar a compreensão e retenção de conhecimento, e auxiliar o processo de escrita. Um mapa conceptual é uma imagem das ideias ou tópicos presentes na informação a ser representada e como estes estão relacionados. Pode ser assim denominado um sumário visual que mostra a estrutura do conteúdo que vai ser escrito pelo autor [Crandell et. al, 1996].

1.1. Origem

Os mapas conceptuais fundamentam-se na teoria de Aprendizagem Significativa de David Ausubel [1963], que afirma o seguinte: "as novas ideias e informações podem ser aprendidas e retidas na medida em que conceitos relevantes estejam adequadamente claros e disponíveis na estrutura cognitiva do indivíduo e sirvam dessa forma de plataforma a novas ideias e conceitos". Esta mesma teoria sugere que as pessoas pensam recorrendo a conceitos, revelando a sua

importância para a aprendizagem. Na mesma década de 60, Joseph D. Novak [1965] começou o seu estudo dos mapas conceptuais utilizando a teoria de Ausubel como base. Mais especificamente, tal como Ausubel, Novak defendia que um novo conhecimento só ganha significado quando este pode ser relacionado a uma estrutura de conhecimento já existente, em vez de ser isoladamente processado de acordo com critérios arbitrários [Jonassen, 1996].

1.2. Utilização

Os Mapas Conceptuais têm sido utilizados em diferentes áreas como uma ferramenta de visualização principalmente devido à sua facilidade de manuseio e criação. Na área das Ciências da Educação, o uso de mapas conceptuais tem sido promovido para verificar a percepção de um aluno sobre um determinado tópico aprendido [Novak e Gowin, 1984]. Novak e Gowin defendem, ainda, que os mapas conceptuais são o método mais adequado para representar conhecimento, sendo uma das ferramentas mais eficientes para a aprendizagem.

Uma das primeiras formas de representação de conhecimento na Inteligência Artificial, as redes semânticas [Quillian, 1968] são consideradas um tipo de mapa conceptual, na medida em que representam a organização da memória semântica humana através de conceitos. Numa rede semântica, tal como nos mapas conceptuais, os conceitos ou objectos estão interligados através de relações que no seu todo representam a parte objectiva e declarativa do conhecimento. No entanto, uma rede semântica é considerada uma estrutura mais complexa, já que pode representar, geralmente, além das ideias principais, informação mais detalhada por ter uma organização multi-dimensional; onde cada conceito pode ser visto em detalhe através de redes embebidas. Esta característica permitiu que as redes semânticas fossem aplicadas na análise aprofundada de textos e redes de ideias assemelhando-se à estruturação obtida em documentos de hipertexto[1] [Murray, WWW].

Um mapa conceptual tipicamente representa as ideias principais e as suas inter relações num espaço bidimensional de uma forma simples e informal. O tamanho de um mapa conceptual é no máximo o tamanho de uma única folha de papel ou do ecrã (até no máximo 30 conceitos). Deste modo, os mapas conceptuais são também empregues na geração de ideias (*brainstorming*) e denominados mapas cognitivos na Gestão, sendo utilizados como recurso no processo de tomada de decisão [Axelrod, 1976]. Em [Thadgard, 1992] e [Nersessian, 1989], é proposta uma analogia sobre a dinâmica dos mapas conceptuais como modelo para os processos de alterações conceptuais nas revoluções científicas estudadas na História da Ciência. Com um maior grau de

[1] Hipertexto é uma tecnologia informática que consiste em associar a blocos de texto individuais ligações electrónicas que os relacionam, apresentando muitos pontos em comum com a teoria literária.

formalismo, os grafos conceptuais [Sowa, 1984] são uma forma de representação de conhecimento complexa com menor cobertura que os mapas conceptuais já que cada grafo representa apenas uma proposição de cada vez. Os grafos conceptuais são considerados como uma representação mais formal que os mapas conceptuais por ter um conjunto definido e restrito de relações possíveis entre os objectos; e por estes objectos estarem classificados na rede semântica que acompanha cada estrutura quanto ao seu tipo [Kremer, 1994] (ao contrário dos mapas conceptuais, onde não há quaisquer restrições nem significados para os objectos /relações).

Podemos concluir que os mapas conceptuais são uma poderosa ferramenta de organização e representação do conhecimento que permite, de uma forma simples, apresentar os conceitos envolvidos e a estrutura de um domínio. E, o mais importante, não existe *o* mapa sobre um dado assunto, mas sim, mais uma representação possível construída sob a interpretação do autor do mapa. Para se ter uma ideia da simplicidade e ao mesmo tempo poder de representação desta ferramenta, é apresentado na Figura 2.1 um mapa conceptual típico do domínio educacional feito por um estudante a quem se pediu para expor o que pensava sobre o conceito *water* [Novak e Gowin, 1984]. O mapa tem dois tipos de nós: conceitos expressos em elipses e exemplos em rectângulos, que no seu conjunto comunicam algumas características físicas e biológicas da água.

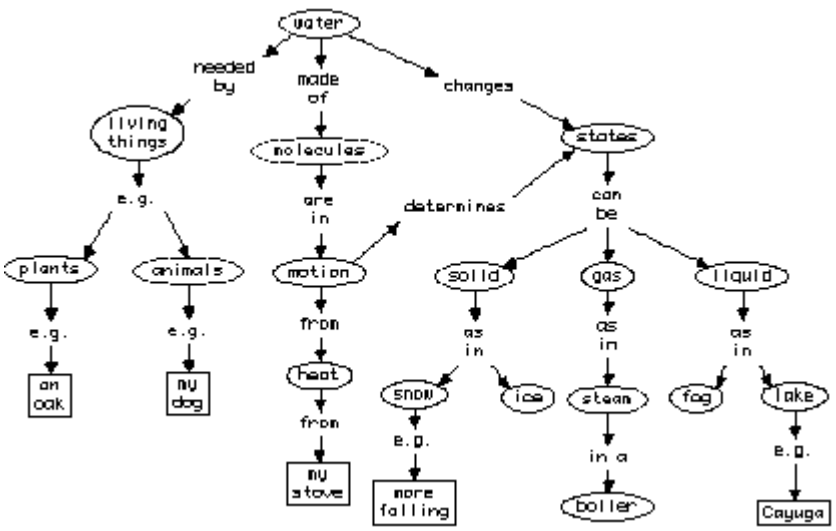


Figura 2.1: Mapa conceptual feito por um estudante sobre o conceito *water*.
(Extraído de [Kremer e Gaines, 1994])

1.3. Organização dos Mapas: Hierarquia vs. Rede de Conceitos

A definição original de Joseph Novak sobre os mapas conceptuais aponta para uma estrutura hierárquica dos conceitos (como pode ser observado na Figura 2.1), tal que o conceito mais inclusivo e geral deva aparecer no topo do mapa, descendendo deste os mais específicos e

subordinados [Jonasson e Grabowski, 1993]. A hierarquia inerente a um dado domínio também depende do contexto em que o conhecimento está a ser considerado. Sendo, assim, Novak [2002] aconselha a construção de mapas conceptuais utilizando como referência uma questão em particular a responder ou alguma situação ou evento que tentamos entender através da organização do conhecimento.

Por outro lado, outros investigadores na área educacional enfatizam que em certos domínios não é possível determinar se uma estrutura é hierárquica ou não, tendo sido um tópico de discussão já há três décadas sem uma definição formal e consensual [Whyte et. al, 1969] e [Dress e von Haeseler, 1990]. Acrescentam ainda que os mapas conceptuais hierárquicos não são uma representação válida da forma como o conhecimento está armazenado na memória humana, que defendem estar mais próximo de uma rede de conceitos [Lawson, 1994]. Deste modo, [Ruiz-Primo et. al, 1997] defendem que os mapas conceptuais devam ser estruturados de acordo com o domínio representado, formando *clusters* à volta dos conceitos mais importantes.

É importante notar que Novak [Novak e Gowin, 1984] não descreve uma estrutura modelo dos mapas conceptuais estritamente hierárquica: é possível interligar ramos que descendem dos conceitos mais gerais (ver um exemplo na Figura 2.2).

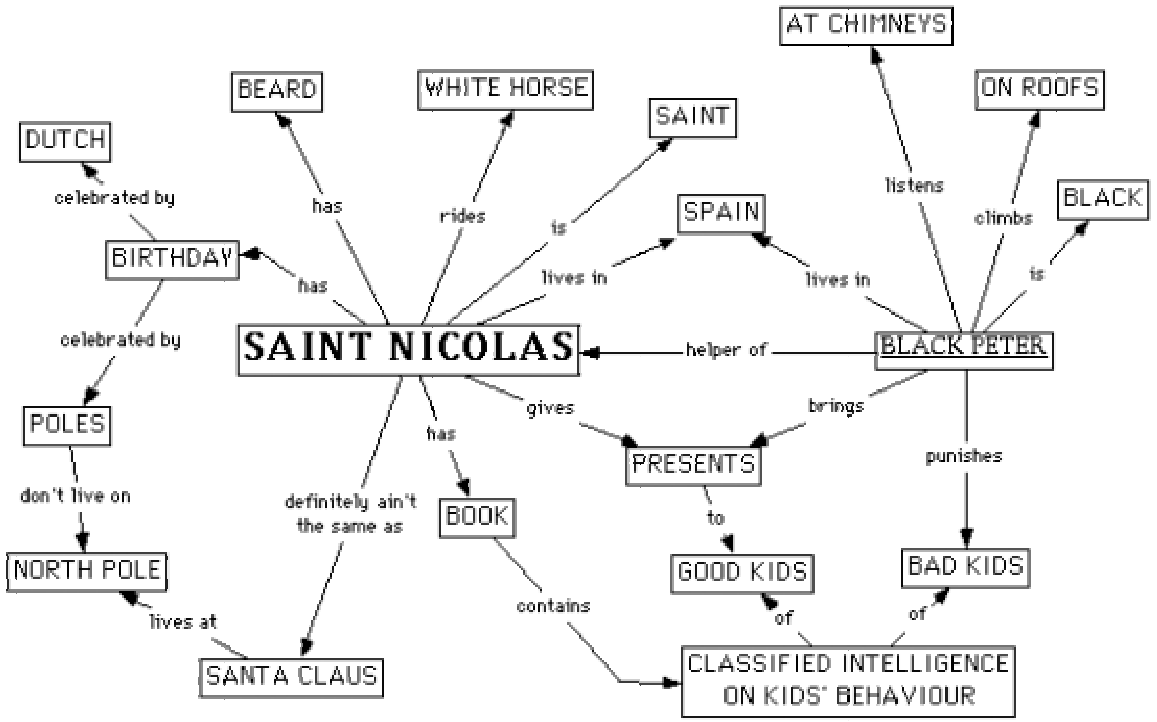


Figura 2.2: Mapa conceptual não-hierárquico sobre o domínio *saint nicolas*.
(Extraído de [Lanzing, 1997])

1.4. Criação e Dimensão dos Mapas Conceptuais

Sem restringir a estrutura do mapa conceptual a uma hierarquia, mas sim associando-o a uma rede de conceitos, podemos seguir alguns passos na construção manual de mapas conceptuais [Zimmaro e Claweya, 1998]:

i. Identificar os conceitos ou termos importantes a incluir no mapa;

ii. Organizar os conceitos que melhor representem a informação;

iii. Utilizar círculos ou elipses para encapsular cada conceito;

iv. Utilizar setas para ligar os termos que estão relacionados;

v. Utilizar uma palavra ou proposição como etiquetas das setas para indicar a relação entre dois termos;

vi. Refinar o mapa de modo a escolher as principais ideias e determinar quais relações devem existir entre elas.

A grande questão está em saber o limiar que separa um mapa incompleto de um impossível de compreender devido ao grande número de nós e arcos representados. Para uma melhor compreensão, alguns investigadores sugerem que o conteúdo de um mapa conceptual que represente um dado domínio deva ser limitado tanto nos conceitos como no tipo de relações que possam existir entre estes [Hibberd et. al, 2002]. Existem abordagens para a construção de mapas conceptuais que oferecem um conjunto de relações previamente definidas, denominadas *canonical links,* com o objectivo de auxiliar o processo de criação do mapa [Holley et. al, 1979] e [Lambiotte et. al, 1989].

Para uma maior automatização deste processo de criação, foram desenvolvidos os programas *Clouds* [Pereira et. al, 2000] e *TextStorm*[Alves et. al, 2001] que possibilitam a construção interactiva de mapas conceptuais, A primeira constrói mapas conceptuais por aprendizagem lógica indutiva e pela interacção com utilizador, enquanto que o *TextStorm* serve de interface em Língua Natural a esta primeira, na medida em que extrai mapas conceptuais de textos. O objectivo global do sistema é permitir ao utilizador que queira representar um dado domínio criar inicialmente uma base de conhecimento por mapas conceptuais a partir de textos sobre o assunto e, de seguida, em sessão com o *Clouds*, completar a representação respondendo algumas questões propostas durante o processo de aprendizagem. O resultado final é um mapa conceptual com uma abrangência muito maior do que o conhecimento introduzido (seja através de repostas ou textos) pelo utilizador, incluindo também deduções e algumas regras aprendidas.

Utilizando o princípio de representação por mapas conceptuais, foram desenvolvidas outras ferramentas alternativas à utilização de informação textual. Naturalmente, cada ferramenta tem a sua própria notação e foi desenvolvida para um certo propósito, mas a característica

comum é o facto de representarem conhecimento através de grafos. Alguns exemplos podem ser encontrados em diversos trabalhos de investigação: KRS [Gaines, 1991] é uma linguagem visual baseada na linguagem de representação de conhecimento textual CLASSIC; em [Cuena e Molina, 1996] é proposto um sistema de modelação do conhecimento denominado KSM (*Knowledge Structure Manager*) que cria mapas conceptuais a partir de uma estrutura base de conhecimento; e, por fim, o método de desenvolvimento de sistemas multi-agente baseados em conhecimento DESIRE (*DEsign and Specification of Interacting REasoning components*) utiliza um tradutor automático para converter textos em representações gráficas.

Outros investigadores propõem uma visão alternativa para a representação estática de conhecimento através de mapas conceptuais. Na abordagem Construtivista [Jonassen et. al, 1998], os mapas conceptuais também permitem auxiliar o processo de evolução da aprendizagem. Esta natureza temporal dos mapas conceptuais não é enfatizada na maior parte da literatura sobre o assunto. A tentação de ver um mapa como uma imagem fixa e acabada é constante, em vez disso, ele é nada mais que a representação de um estado da aprendizagem. Ou seja, um mapa conceptual nunca está terminado, já que o conhecimento por ele representado está sempre a evoluir.

1.5. *Semantic Web*: Uma representação baseada nos mapas conceptuais

Ultimamente tem-se observado um grande interesse na introdução de informação semântica na estrutura hipertexto das páginas *Web.* Este significado inerente aparece na forma de uma ligação de cada conceito referido na página actual a uma nova página onde é definido o seu significado. A *World Wide Web* poderia assim se transformar numa imensa rede de conceitos à escala global. Esta nova abordagem é denominada *SemanticWeb*, que representa uma colecção evolutiva de conhecimento, construída para permitir que qualquer utilizador da *Internet* possa adicionar e inferir nova informação.

Até então, os conteúdos *on-line* eram apenas compreensíveis pelos humanos e somente legíveis (*readable*) pela máquina, o que tornava impraticável qualquer processamento automático de dados na *World Wide Web.* Através da *SemanticWeb*, a informação poderá também ser manipulada e compreendida de uma forma pelos sistemas informáticos. Em vez da linguagem natural, a representação dos dados é composta por factos descritivos, que definem as relações existentes entre um e outro conceito.

Na linguagem humana utilizamos frequentemente um mesmo termo para referir diferentes coisas, mas a automatização não admite ambiguidades A solução proposta é desambiguar cada termo referindo um URI (*Universal Resource Identifier*) específico para cada

significado possível. A *Semantic Web*, ao referir para cada conceito um URI, permite a criação de novos conceitos apenas ligando-os a uma página onde está a sua definição, tornando-se uma linguagem lógica unificada e possibilitando a integração destes novos conceitos na *World Wide Web*.

2. Conhecimento Semântico: Taxonomia de conceitos

Antes de propriamente comparar semanticamente mapas conceptuais, devemos descobrir o significado do domínio que é representado por cada mapa em si, ou seja, tentar contextualizá-lo tendo como base o conhecimento sobre o mundo. Por apenas apresentar o conhecimento de uma forma simples e intuitiva, as regras de construção dos mapas conceptuais não abrangem a especificação do significado de cada conceito ou relação aquando da sua criação. E porque será que não se deveria criar tal regra? A questão talvez seja que, quando Novak [Novak e Gowin, 1984] começou a desenvolver este novo conceito de representação do conhecimento para utilização primária na aprendizagem e ensino, não tinha como objectivo tratá-lo computacionalmente, mas sim utilizar os principais instrumentos que um professor tem a sua disposição: lápis e papel.

Como, neste trabalho, o objectivo é comparar mapas conceptuais de uma forma automática através de um sistema computacional, precisamos, de alguma forma, contextualizar o conhecimento apresentado no mapa para que de alguma forma possamos saber o seu significado. Esta contextualização utiliza como recurso um dicionário electrónico amplamente utilizado no Processamento da Língua Natural e Anotação Semântica denominado WordNet [Miller, 1990] que será amplamente discutido na Secção 2.2.1. A contextualização ou desambiguação dos significados das palavras, assim como as diferentes abordagens para o problema, serão abordadas na Secção 2.2.2.

2.1. WordNet

O WordNet é um recurso lexical electrónico utilizado pela maioria dos projectos científicos em processamento da língua natural, tendo sido desenvolvido (e mantido em contínua evolução) pela Universidade de Princeton. Reunindo um grande número de palavras da língua inglesa introduzidas e classificadas manualmente, o WordNet é um léxico sistematicamente estruturado onde é possível determinar o que as palavras podem significar e como podem ser usadas. É aceite actualmente como uma base de conhecimento padrão para a consulta de informação semântica pela sua disponibilidade (sem qualquer encargo para quem o queira utilizar) e actualização (constante inclusão de novas palavras).

Antes de começarmos a aprofundar a análise de estrutura do WordNet devemos definir alguns termos utilizados de menor ambiguidade, já que o próprio significado de "palavra" é muito abrangente e vago. Quando nos referirmos a cada entrada de um dicionário utilizaremos o termo **lexema**. Um lexema pode ser pensado como um triplo composto pela forma ortográfica, forma fonológica e alguma forma de representação de significado de palavras ou termos compostos. Desta forma um **léxico** é uma lista finita de lexemas [Jurafsky e Martin, 2000].

2.1.1. Organização

A estrutura interna do WordNet consiste em relações entre os lexemas e seus significados, tendo como unidade fundamental um *synset*. Cada *synset* representa um conjunto de lexemas da mesma classe gramatical que expressam o mesmo significado (sinónimas). Um mesmo lexema (mas com diferentes significados) pode pertencer a diferentes *synsets* dependendo do contexto onde está inserida. Sendo assim, dois lexemas com a mesma ortografia mas com significados diferentes (**homógrafos**) nunca poderão pertencer ao mesmo *synset*, já que cada lexema representando um dado conceito é identificado univocamente através da sua inclusão num determinado *synset*, por exemplo, *mouse* com o significado de pequeno roedor está enquadrado num *synset* diferente de *mouse* significando um dispositivo electrónico utilizado nos computadores.

Esta organização à volta de conceitos (e não por simples ordem alfabética) foi inspirada nas teorias psico-linguísticas da memória humana, onde o agrupamento de palavras é conceptual [Miller et. al, 1993]. O WordNet também pode ser simplesmente utilizado como um dicionário electrónico, que abrange as quatro classes gramaticais *abertas* (substantivos, adjectivos, verbos e advérbios) da língua inglesa. Estas classes são assim denominadas por evoluírem todos os dias recebendo constantemente novos termos e palavras, ao contrário das outras classes consideradas *fechadas* que permanecem estáveis por umas dezenas de anos (artigos, conjunções, preposições, pronomes, etc.). Cada item do dicionário, representado por um conceito (e internamente por um *synset*) tem uma definição associada (*gloss*) que, de uma forma sucinta, descreve cada significado para a palavra pesquisada. Na Figura 2.3 é apresentada uma imagem da utilização do WordNet como dicionário na pesquisa de palavras.

A versão actualmente utilizada do WordNet é a 1.7.1 onde existem 195817 significados para um total de 140446 palavras diferentes (76653 simples e 63793 compostas) distribuídas por 111223 *synsets*. A distribuição das palavras, significados e *synsets* entre as quatro classes gramaticais abrangidas pelos WordNet é apresentada na Tabela 2.1. Através destes números podemos verificar que a classe dos verbos é a mais ambígua mesmo em relação aos substantivos.

Isto porque, apesar de todas as frases necessitarem de pelo menos um verbo e não necessariamente de um substantivo, a língua Inglesa possui menos verbos que substantivos, sugerindo assim, que os significado dos verbos são mais flexíveis do que os dos substantivos. Gentner e France [1988] conseguiram demonstrar o que afirmam ser a alta mutabilidade dos verbos: "podem alterar o seu significado dependendo do tipo de argumentos que o acompanham", enquanto que os significados dos substantivos tendem a ser mais estáveis na presença de diferentes verbos. Os verbos mais frequentemente usados (*have*, *be*, *run*, *make*, *set*, *go*, *take* e outros) são também os mais ambíguos ao passo que a grande maioria possui significado único.

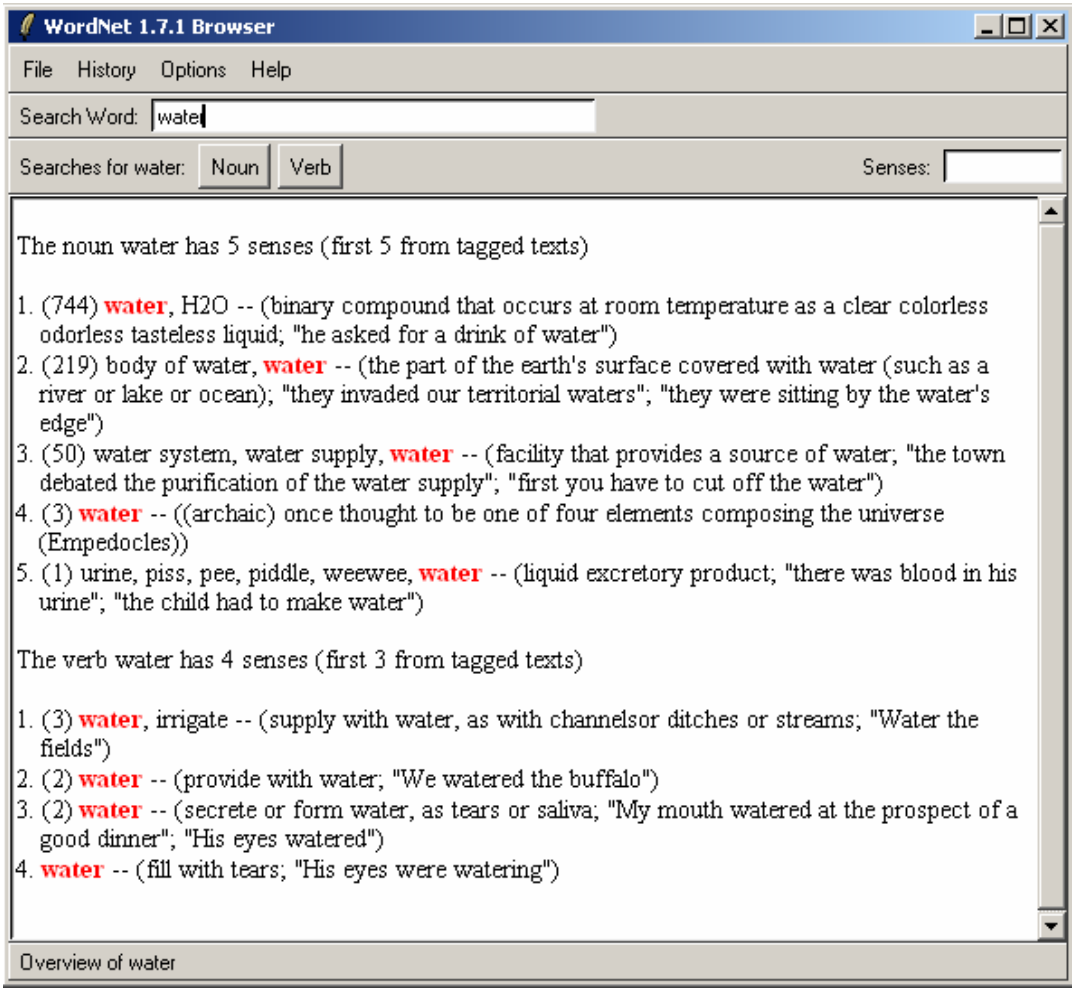


Figura 2.3: Resultado de uma pesquisa efectuada no *browser* do WordNet (versão 1.7.1) para a palavra *water*.

A utilização em larga escala deste recurso lexical levou à necessidade de construir diferentes versões do WordNet para outras línguas além do Inglês. O *EuroWordNet* [Vossen, 1997] é uma base de dados multilingue contendo diversos *wordnets* de línguas europeias (Alemão, Holandês, Espanhol, Francês, Italiano, Checo e Estónico) estruturados e possuindo relações semânticas básicas da mesma forma que o WordNet original desenvolvido em

Princeton. Além das relações semântica inerentes a cada língua, esta versão conjunta inclui um lista central que mapeia todos os *synsets* ao seu significado equivalente na versão do WordNet 1.5, denominado *Inter-Lingual-Index* (ILI). Desta forma, é possível interligar todos os *wordnets* e utilizar esta informação em diversas tarefas de PLN e *Information Retrieval*, tais como pesquisas multilingue, tradução automática, transferência de conhecimento lexical e relações semânticas entre *wordnets*. Actualmente, o *EuroWordNet* encontra-se em fase final de desenvolvimento estando disponível, tal como o WordNet original, livremente para a comunidade científica. Neste mesmo modelo, está a ser desenvolvida a versão portuguesa do WordNet denominada *WordNet.PT* [Marrafa, 2001] tornando possível, desta forma, a integração da nossa língua neste projecto europeu de processamento semântico.

O WordNet é considerado diferente dos dicionários tradicionais, não só por organizar as lexemas em *famílias de sinónimos* (tradução à letra de *synsets*) mas também por interligar estas unidades conceptuais através de uma variedade de relações semânticas próprias de cada classe gramatical. Estas relações semânticas serão estudadas em maior profundidade na próxima Sub Secção. Desta forma, os *synsets* não se encontram isolados, tendo cada classe gramatical presente no WordNet uma organização e relações próprias implementadas. Os substantivos estão organizados em hierarquias, enquanto que os verbos estão vinculados através de relações formando uma rede semântica. Os adjectivos e advérbios organizam-se de uma forma semelhante formando um hiperespaço multi-dimensional. Esta distribuição foi assim concebida, pois ao tentar impor um único esquema de organização para todas as categoriais gramaticais, a complexidade do conhecimento lexical estaria a ser representada de uma forma equívoca [Miller et. al, 1993].

	Total	**Substantivos**	**Verbos**	**Adjectivos**	**Advérbios**
Número de palavras	147769	110176	11090	21904	4599
Número de significados	195817	134716	24169	31184	5748
Significados/palavra	1,33	1,22	2,18	1,42	1,25
Número de *synsets*	111223	75804	13214	18576	3629
Significados/*synset*	1,76	1,78	1,83	1,68	1,58

Tabela 2.1: Números e abrangência do WordNet1.7.1

2.1.2. Relações Semânticas

As relações semânticas implementadas pelos projectistas do WordNet incluem: sinónimos, antónimos, hiperónimos (*isa*), hipónimos (*kind_of*), holónimos (*part_of*), merónimos (*has_part*), atributos, tropónimos, funções, entre outras.

Sinónimo

Comum às quatro classes gramaticais e implícita na própria organização do WordNet, a relação **sinónimo** existe entre todos os lexemas que pertencem ao mesmo *synset*. Desta forma, esta relação semântica não existe entre conjuntos (*synset*) mas sim entre lexemas individuais. Segundo Leibniz (grande filósofo do século XVII), duas expressões são consideradas sinónimas somente se a substituição de uma por outra não alterar o significado da frase onde se encontram. Por esta definição seria muito difícil encontrar lexemas verdadeiramente sinónimos que pudessem ser permutáveis em qualquer contexto. Por exemplo, o substantivo *plank* possui os seguintes significados segundo o WordNet:

```
1. a stout length of sawn timber; made in a wide variety of sizes and
used for many purposes
2. an endorsed policy in the platform of a political party
```

Este lexema só poderá ser considerado sinónimo de *board* no contexto de objectos de madeira (primeiro significado). Segundo [Miller et. al, 1993], a concepção do WordNet adaptou esta definição de sinónimos para uma versão mais flexível em que: ”Duas expressões são consideradas sinónimas num dado contexto, somente se a substituição de uma por outra não alterar o significado deste contexto”.

Uma vez que dois lexemas sinónimos pertencem ao mesmo *synset* e podem ser substituídos entre si sem qualquer modificação do contexto onde estão inseridos, estes devem pertencer à mesma classe gramatical. Esta é a principal razão para a divisão do WordNet em classes gramaticais (substantivos, verbos, adjectivos e advérbios), ou seja, um *synset* só pode pertencer a uma classe e consequentemente todas as palavras que o constituem também serão categorizadas de igual modo.

Antónimo

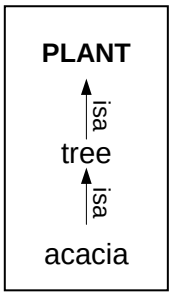
Esta é uma relação semântica, tal como os sinónimos, interliga lexemas ao invés de conceitos (*synsets*). Isto porque nem todos os lexemas de um *synset* podem ter o mesmo antónimo, como por exemplo: os dois significados expressos pelos verbos {*rise*, *ascend* (*come up, of celestial bodies*)} e {*fall, descend* (*come as if by falling*)} são considerados opostos mas não antónimos. Enquanto, que, *rise/fall* e *ascend*/*descend* são naturalmente referidos como lexemas antónimos.

Nem sempre é fácil definir esta relação, já que nem sempre o antónimo de uma lexema **x** é uma **não-x**. Um exemplo disto é o facto de haver um meio-termo, ou seja, quando dizemos que *rich* e *poor* são antónimos não significa que uma pessoa que não é pobre seja rica.

Tal como acontece com os sinónimos, as quatro classes gramaticais possuem implementadas a relação de antónimo.

Hiperónimos/Hipónimos

Estas relações expressam a hierarquia existente entre conceitos (*synsets*): pares de lexemas onde um é subclasse de outro, e é a base da organização dos substantivos no WordNet. Por exemplo: a palavra *tree* é considerada uma generalização (ou hiperónimo) de *acácia*, por outro lado, também é uma especialização (ou hipónimo) de *plant*. Estas duas relações semânticas são transitivas e simétricas entre si, formando uma estrutura semântica hierárquica também denominada sistema de herança por [Touretzky, 1986]: "um hipónimo herda todas as características e um conceito mais geral de qual descende e é distinguido deste último (e de outros hipónimos do mesmo conceito mais geral) por possuir pelo menos mais uma característica própria". No exemplo dado, *tree* possui todas as características de *plant*, mas destaca-se das outras plantas por ser alta e possuir um tronco principal com ramos a formar uma elevada copa.

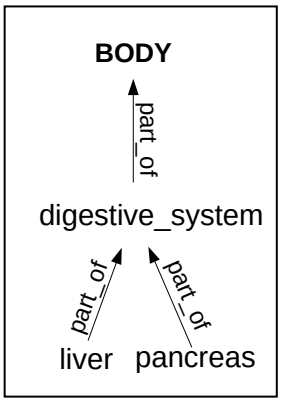


Somente os substantivos e verbos se organizam de uma forma hierárquica no WordNet formando uma taxonomia própria para cada uma destas classes gramaticais. Os substantivos possuem hipónimos para referir os conceitos mais específicos, enquanto que, nos verbos, os modos particulares de implementar uma dada acção são denominados *tropónimos* (e.g., *walk* é tropónimo de *locomote*). Tanto substantivos como verbos definem os conceitos/acções mais gerais como sendo hiperónimos (e.g., *walk* é hiperónimo de *march*).

A taxonomia hierárquica dos substantivos é considerada, em termos visuais, muito mais profunda e estreita que a dos verbos, sendo estes últimos organizados de uma forma muito mais larga e superficial que a primeira. Existem apenas 9 conceitos mais gerais na hierarquia de substantivos constituídos pelos seguintes *synsets*: {*entity*, *physical_thing*}, {*psychological_feature*}, {*abstraction*}, {*state*}, {*event*}, {*act*, *human_action*, *human_activity*}, {*group*, *grouping*}, {*possession*} e {*phenomenon*}; e uma profundidade máxima de 18 conceitos, como por exemplo, ligando o conceito mais específico {*rock_hind*, *Epinephelus adscensionis*} até o mais geral {*entity, physical_thing*}. Na classe gramatical dos verbos, existem 357 *synsets* considerados como os mais gerais e uma profundidade máxima de 12 *synsets*, por exemplo na generalização do verbo *{rumpus}* até *{move, displace}*.

Holónimos/Merónimos

Tal como os sinónimos, antónimos e generalizações/especializações, os holónimos/merónimos são relações semânticas familiares não sendo necessário nenhum conhecimento linguístico para as reconhecer. Um merónimo é considerado como sendo parte integrante do conceito a que se relaciona, enquanto que holónimo é a relação simétrica desta, onde um conceito possui diversos elementos constituintes. Por exemplo: *body* é holónimo de *digestive system*, enquanto que *pâncreas* é um merónimo deste último conceito.



Visto que somente os substantivos podem ser decompostos em outros mais elementares, o WordNet apresenta a relação de holónimos/merónimos implementada apenas para esta classe gramatical.

Atributos

Esta é uma das duas relações semânticas (e a única actualmente implementada) entre diferentes classes gramaticais implementada no WordNet (a seguir veremos a outra relação). Representa as características (adjectivos) que um dado conceito (substantivo) pode representar. Esta é uma relação simétrica em que os adjectivos modificam os substantivos, enquanto que os substantivos servem de argumento aos adjectivos. Por exemplo, *size* e *weight* são dois substantivos cujos atributos podem ser *small* e *light*. Actualmente, somente 325 substantivos e 631 adjectivos têm esta relação implementada, formando 650 ligações entre as duas classes.

Funções

Ainda não implementada no WordNet mas com disponibilização prometida para breve, esta é a outra relação semântica que relaciona diferentes classes gramaticais: substantivos e verbos. Uma função (verbo) de um conceito (substantivo) é entendida como a descrição de algo que as instâncias do conceito normalmente praticam, ou actividades em que geralmente são utilizadas como instrumentos, ou ainda, acções frequentemente perpetradas a estas instâncias. Algumas associações parecem mais naturais que outras, por exemplo é comummente aceite que a função de um lápis seja escrever, enquanto que a de uma faca seja cortar, mas afirmar que a função de um canário seja voar ou cantar parece um pouco forçado.

O Quadro 2.1 resume as principais relações semânticas aqui discutidas, assim como as classes gramaticais envolvidas.

Relação	Significado	Classes gramaticais	Transitiva?	Relação simétrica
sinónimo	lexemas do mesmo *synset* com significado igual ou muito semelhante	substantivos, verbos, adjectivos e advérbios	X	sinónimo
antónimo	lexemas pertencentes a diferentes *synsets*, opostos no significado	substantivos, verbos, adjectivos e advérbios		antónimo
hiperónimo	*synset* que é a generalização de outro	substantivos: page isa leaf verbos: page isa number	X	hipónimo tropónimo
hipónimo	*synset* que é a especificação de outro	substantivos jeep *isa_kind_of* car	X	hiperónimo
tropónimo	*synset* que é a especificação de outro	verbos sleep *is_one_way_to* rest	X	hiperónimo
holónimo	*synset* que é um conjunto de partes	substantivos hand *has_part* finger	X	merónimo
merónimo	*synset* que é uma parte de outro	substantivos hand *part_of* arm	X	holónimo
atributo	*synsets* (adjectivos) que representam valores para um outro (substantivo)	entre adjectivos e substantivos: {poor} *value_of* {financial condition, economic condition}		atributo
função	*synsets*(verbos) que representam a funcionalidade de outros *synsets* (substantivos).	entre substantivos e verbos: {chair} *has_function* sit		função

Quadro 2.1: Algumas relações semânticas do WordNet1.7.1.
(a relação *função* ainda não se encontra implementada, somente especificada).

2.1.3. Conhecimento sobre o uso de conceitos

Existe, actualmente, uma distinção na classificação do conhecimento linguístico entre lexical (sobre as características das palavras) e contextual (sobre a utilização das palavras) [Fellbaum, 1998]. Como repositório de tais conhecimentos, os dicionários são incluídos na primeira classe e as enciclopédias na segunda. Por exemplo, a palavra *hit* é conhecida como um verbo irregular e sinónima de *strike* (léxico). Em relação ao seu significado, podemos afirmar que o acto de bater em alguém é considerado hostil. Ambos os tipos de conhecimentos são necessários para um completo entendimento do significado e utilização das palavras.

O WordNet possui um conjunto de relações semânticas que são baseadas na similaridade e contraste entre palavras e conceitos, e não sobre a sua familiaridade ou utilização conjunta no dia-a-dia. Uma vez que o conhecimento sobre o mundo é encarado como sendo vasto e complexo de organizar, os projectistas do WordNet propuseram-se a conceber um léxico totalmente estruturado que inclui alguma informação, em formato texto, sobre a utilização dos conceitos através das definições (*glosses*) que acompanham os *synsets*. Harabagiu e Moldovan [1998] propõem ampliar o conhecimento oferecido pelo WordNet utilizando as palavras contidas

nos *glosses*. Como cada *synset* possui obrigatoriamente uma definição (cada definição é constituída de palavras, e por sua vez cada palavra pode pertencer ao seu próprio *synset* no WordNet) é possível criar um *microcontexto* à volta do *synset* primeiramente referido, permitindo inferir mais conhecimento. Por exemplo, *racquet*, *ball*, e *net*[2] não estão actualmente ligadas por qualquer relação no WordNet, no entanto ao examinarmos as suas definições isoladamente é possível estabelecer um *caminho semântico* entre as três palavras, tal como é apresentado no Quadro 2.2. Neste exemplo, cada uma das palavras presentes nas definições dos *synsets* podem ter, por sua vez, mais do que um significado (como é o caso da palavra *game* que possui 8 significados distintos no WordNet). Por esta razão, para ser possível implementar propostas como a de Harabagiu e Moldovan de utilização dos *glosses*, estes mesmo autores desenvolveram recentemente (disponibilizado na altura da escrita deste documento) na Universidade do Texas, o *Extended* WordNet [Harabagiu et. al, 1999; Moldovan, 2003] que inclui informação extraída dos glosses anotada semanticamente[3] em relação aos *synsets* que os compõem.

Visto que o WordNet por si só é pobre no que se refere à informação sobre a utilização das palavras e conceitos que contém, e, tal como Miller e Charles [1991] haviam observado que o conhecimento sobre as palavras deve consistir, além dos seus significados, dos contextos onde estas podem ocorrer, foi desenvolvida uma base de dados auxiliar denominada ***SemCor*** (*semantic concordance*) por Landes et. al [1998]. O *SemCor* combina um subconjunto dos textos (*corpus*) extraídos do *Standard Corpus of Present-Day Edited American English* (mais conhecido como *Brown Corpus*[4] por ter sido desenvolvido na Universidade de Brown [Kučera e Francis, 1967]) com o léxico do WordNet através de uma relação bidireccional entre estes dois, de forma que todas as palavras dos textos foram manualmente identificadas quanto à sua sintaxe e significado (através de ligações aos *synsets*) e grande parte dos *synsets* (os que aparecem no *corpus*) possuem frases-exemplo de utilização com um grande potencial para aplicações educacionais e de desambiguação automática de significados.

[2] Conhecido como "*The Tennis Problem"* por Roger Chaffin (comunicação pessoal) em que afirma que o principal desafio dos dicionários electrónicos é conseguir armazenar informação estruturada sobre tópicos de discurso.
[3] Este processo de identificação de significados é denominado Desambiguação dos Sentidos das Palavras que abordaremos na próxima Sub Secção .
[4] Foi construído a partir de artigos do *Wall Street Journal*, abrangendo um total de 1 milhão de palavras.

The noun racket has 4 senses:
1. a loud and disturbing noise
2. an illegal enterprise (such as extortion or fraud or drug peddling or prostitution) carried on for profit
3. the auditory experience of sound that lacks musical quality; sound that is a disagreeable auditory experience
4. a sports implement (usually consisting of a handle and an oval frame with a tightly interlaced network of strings) used to strike a ball (or shuttlecock) in various games

The noun ball has 11 senses:
1. **round object that is hit or thrown or kicked in games**
2. a solid ball shot by a musket
3. an object with a spherical shape
4. the people assembled at a lavish formal dance
5. one of the two male reproductive glands that produce spermatozoa and secrete androgens
6. a spherical object used as a plaything
7. a compact mass
8. a lavish formal dance
9. a more or less rounded anatomical body or mass; ball of the human foot or ball at the base of the thumb
10. a ball game played with a bat and ball between two teams of 9 players; teams take turns at bat trying to score run
11. a pitch that is not in the strike zone

The noun net has 6 senses:
1. an interconnected or intersecting configuration or system of components
2. a trap made of netting to catch fish or birds or insects
3. the excess of revenues over outlays in a given period of time (including depreciation and other non-cash expenses)
4. a goal lined with netting (as in soccer or hockey)
5. game equipment consisting of a strip of netting dividing the playing area in tennis or badminton
6. an open fabric woven together at regular intervals

Quadro 2.2: Alguns significados para os substantivos *racket*, *ball*e *net.* Extraído do WordNet1.7.1.

2.2. Desambiguação dos Significados de Palavras

Cada termo, ou palavra, expresso no mapa conceptual (seja representando um nó ou arco) quando visto isoladamente pode descrever vários significados. Para utilizarmos uma taxonomia como recurso lexical, como o *WordNet*, precisamos determinar exactamente (ou com um certo grau de confiança) que elementos desta hierarquia estão a ser referenciados no mapa conceptual.

A *Desambiguação* do sentido das palavras (*Word Sense Disambiguation*) é uma subárea de investigação do Processamento da Língua Natural que procura estabelecer metodologias para a correcta identificação dos conceitos representados por palavras observando o seu contexto. Assume-se que uma palavra possa ter um número finito de significados (recorrendo-se a um dicionário, thesaurus ou outra fonte de referência) e a tarefa do sistema é escolher um destes significados para cada utilização possível da palavra ambígua. Por exemplo, o lexema *'dish'*,

dependendo do contexto onde está inserido (ver Quadro 2.3), pode representar a ideia de uma peça de loiça, um prato preparado, o prato de uma antena parabólica, etc.

```
The noun dish has 6 senses:

1. (9) dish -- (a piece of dishware normally used as a container for holding
or serving food; "we gave them a set of dishes for a wedding present")
2. (3) dish -- (a particular item of prepared food; "she prepared a special
dish for dinner")
3. dish, dishful -- (the quantity that a dish will hold; "they served me a
dish of rice")
4. smasher, stunner, knockout, beauty, ravisher, sweetheart, peach, lulu,
looker, mantrap, dish -- (a very attractive or seductive looking woman)
5. dish, dish aerial, dish antenna, saucer -- (directional antenna consisting
of a parabolic reflector for microwave or radio frequency radiation)
6. cup of tea, bag, dish -- (an activity that you like or at which you are
superior; "chemistry is not my cup of tea"; "his bag now is learning to play
golf"; "marriage was scarcely his dish")
...
```

Quadro 2.3: Alguns significados da palavra *dish* como substantivo. Extraído do resultado de uma pesquisa efectuada pelo WordNet Browser1.7.1 [Miller et. al, 1993]

Esta tarefa auxiliar pode ser facilitada quando um lexema apresenta significados claramente distintos entre si (**homonímia**) ao invés de significados que estão muito próximos semanticamente, não havendo assim necessidade de dois lexemas separados mas apenas um com significados múltiplos relacionados (**polissemia**). Estes fenómenos podem ser melhor explicados se observarmos a sua distinção na prática. Ainda utilizando a palavra 'dish', os conceitos de prato de loiça e um prato confeccionado estão relacionados ambos com o conceito comida, tendo assim uma semelhança muito maior (sendo consideradas palavras polissémicas, ou seja, ortografia igual e significados próximos) do que se comparados com o significado de 'dish' como sendo parte de uma antena parabólica (neste caso, palavras homónimas, ou seja, ortografia igual, e significados completamente não-relacionados).

Existem actualmente, várias abordagens para o problema, sem no entanto haver uma padrão ou fortemente estabelecida, mas é comummente aceite uma classificação quanto à simbologia e base de conhecimento das mesmas. De seguida veremos as principais abordagens e suas variantes para a desambiguação do significado das palavras.

2.2.1. Representação Simbólica: Restrição Seleccionadora

Esta abordagem analisa a estrutura e o significado das frases em paralelo, utilizando a estrutura sintáctica da frase para validar a análise semântica da mesma. O método utilizado baseia-se em regras sintáctico-semânticas (denominadas **seleccionadoras**) que são aplicadas às frases durante o processo de *parsing* inserido numa análise linguística integrada. Assim que é detectada uma violação a uma dada restrição seleccionadora não é efectuada a análise sintáctica da frase.

Geralmente, uma restrição seleccionadora é formulada recorrendo-se à concordância semântica imposta por uma palavra sobre os conceitos adjacentes que podem servir de argumento a determinados predicados, como por exemplo:

wash the dishes

serve delicious dishes

Na primeira frase, o verbo *wash* requer um objecto que possa ser lavável, o que restringe a sua aplicabilidade aos significados de *dish* como sendo loiça ou um elemento de uma antena parabólica. No segundo exemplo, o verbo *serve* impõe uma restrição de comestibilidade ao objecto que o acompanha, neste caso, somente atendida pelo significado de prato de comida confeccionada.

Os significados que não atendem às restrições impostas não são considerados na fase seguinte de análise, eliminando-se assim grande parte da ambiguidade no uso das palavras descartando representações semânticas mal-formadas. Uma representação semântica é considerada mal-formada quando, por exemplo, um objecto não corresponde a uma determinada restrição imposta por um verbo.

Para aplicar esta abordagem é necessário uma grande base de conhecimento para:

- conter a especificação exaustiva de todas as regras para cada predicado possível ligados aos significados de cada verbo do léxico;
- abranger a taxonomia de todos os conceitos que podem atender a restrição estabelecida.
- O espaço ocupado por esta última necessidade poderia ser atenuado utilizando a informação hierárquica do WordNet, i.e., determinar regras sobre conceitos mais gerais, e inferir que estas mesmas regras também se aplicam aos conceitos que descendem deste conceito mais geral. Utilizando ainda o exemplo das frases com a palavra '*dish'*, se examinarmos a árvore taxonómica do WordNet no que respeita aos múltiplos significados que a palavra poderia ter, examinamos que a primeira frase '*wash the dishes*' é aplicável principalmente ao primeiro significado desta palavra: uma peça de loiça. Criando uma regra semântica para o efeito, poderíamos utilizar o conceito mais geral '*tableware'* (artigos usados à mesa) para englobar outros conceitos tais como: '*dishware*' (louçaria), '*cutlelery'* (talheres), '*silverware'* (baixela de prata), '*tea service'* (serviço de chá), etc., ou ainda mais longe, utilizar o conceito *'physical_object'* que abrange todos os objectos concretos da taxonomia (ver Figura 2.4).

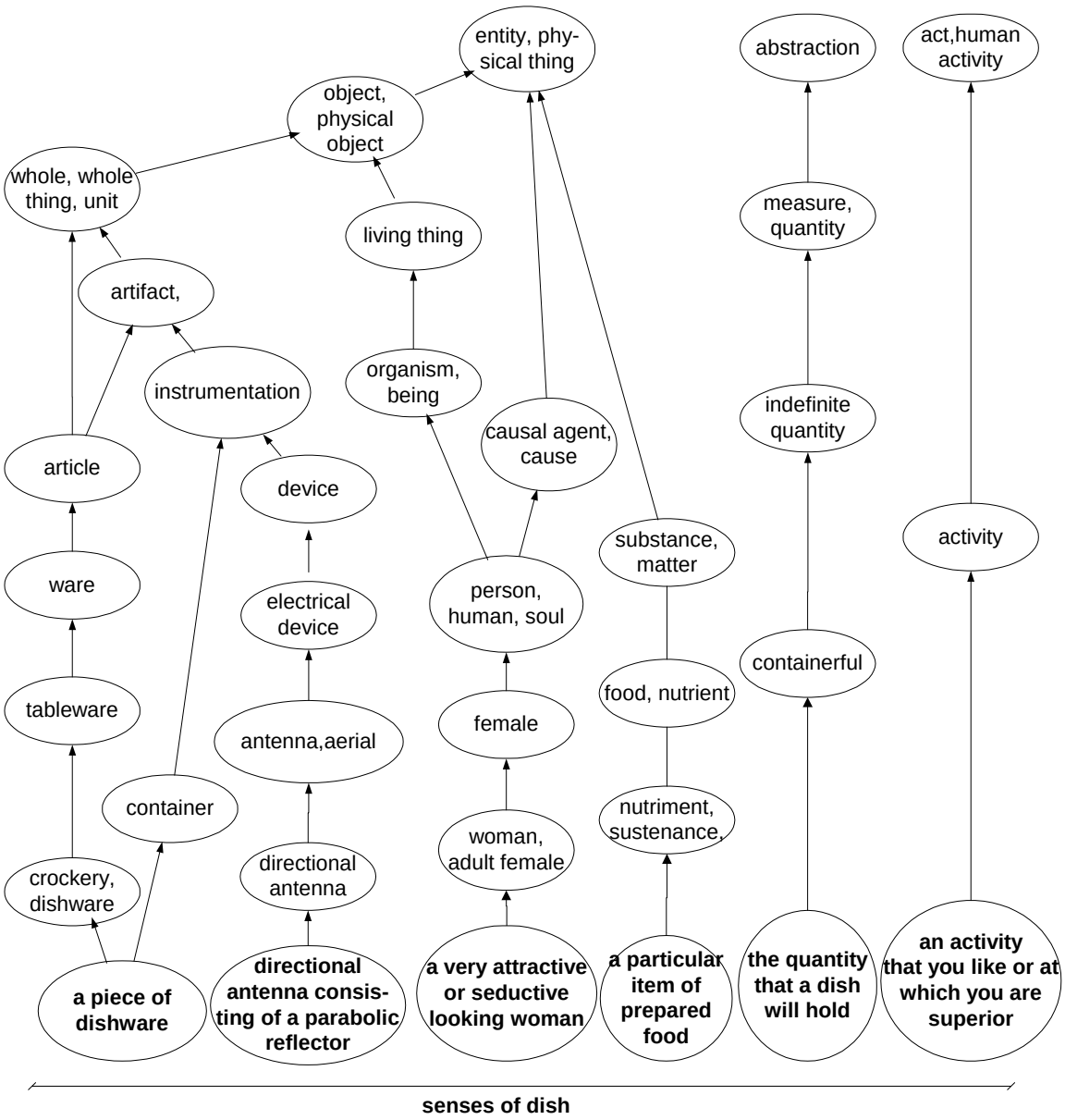


Figura 2.4: Conceitos mais gerais dos significados possíveis da palavra *dish.* Extracto da taxonomia WordNet 1.7.1 (relações *isa* entre conceitos*)*.

Esta abordagem obriga a um conjunto de limitações que não podem ser totalmente resolvidas por regras de concordância sintáctico-semântica, tais como:

- a impossibilidade de se decidir pela utilização mais comum de uma dada palavra para situações em que, analisadas isoladamente, não é claro o significado expresso:

Which dishes do you prefer?

- alguma inflexibilidade para admitir violações que são comummente aceites:

Eat dirt, worm!

- desconsideração da utilização metafórica das palavras:

At this semester, he ate the math books!

Em suma, a elaboração de um grande número de regras não é aplicável à maioria das aplicações de grandes dimensões que processam a língua natural. Mesmo com a utilização do WordNet, torna-se impraticável a formulação exaustiva das restrições seleccionadoras que abranja:

- todos os predicados originados por verbos que só por si são amplamente ambíguos;
- significados possíveis dos objectos que possam integrar estes predicados.

2.2.2. Técnicas Robustas

Têm como objectivo escolher correctamente um dos possíveis significados de uma palavra aplicando técnicas conhecidas de aprendizagem ou a partir do conhecimento presente em léxicos, e surgem, face à necessidade de utilizar métodos mais realistas e modestos nos seus requisitos que as restrições seleccionadoras [Jurafsky e Martin, 2000]. Ao contrário destas últimas, as técnicas robustas para a desambiguação de sentidos, justificando a sua designação, não está condicionada pela disponibilização prévia de informação por outros processos, como a análise sintáctica. Sendo assim, uma ferramenta independente que pode ser facilmente utilizada por aplicações cujo objectivo não é uma análise linguística profunda das palavras nem a sua organização em frases, mas sim apenas encontrar os conceitos-alvo que são expressos no conhecimento textual.

- **Aprendizagem automática baseada em Corpus**

Permite criar um classificador para a escolha de um entre um conjunto limitado de significados possíveis de uma palavra. O classificador é treinado sobre um conjunto de palavras (que podem já estar manualmente desambiguadas ou não) para que possa determinar os sentidos de novas palavras ainda não examinadas. Existem actualmente várias abordagens que utilizam aprendizagem e que variam entre si em aspectos técnicos tais como:

- origem do material de treino: provém geralmente de *corpus* ao invés de recursos lexicais já previamente construídos;
- classificação prévia ou não: estes conjuntos de textos podem estar em bruto ou total/parcialmente desambiguado, em que cada item lexical (ou palavra) é, numa primeira fase, classificado manualmente quanto ao seu sentido;
- resultado produzido;
- pré-processamento e conhecimento linguístico requerido: consiste, geralmente, em classificar gramaticalmente as palavras e extrair os seus radicais (*stemming*). Eventualmente, as frases

do texto onde as palavras se encontram, são guardadas como contexto depois de analisadas sintacticamente;

- quantidade de palavras necessárias para treino;
- necessidade de supervisão humana ou não.

Numa segunda etapa, os exemplos para treino e teste do classificador, respeitantes a cada palavra a ser desambiguada, serão construídos com base nas características linguísticas de colocação (***collocation*)** e/ou co-ocorrência (***co-ocurrence***). O primeiro tipo representa a informação linguística obtida no pré-processamento sobre a palavra-alvo e sobre as palavras a sua volta, enquanto que, na co-ocorrência, é contado o número de ocorrências de palavras que normalmente são associada aos significados possíveis da palavra-alvo (tendo como base frases onde a palavra em questão aparece). A codificação da informação linguística em valores numéricos ou nominais dará origem a um vector de características (***feature vector***) que será utilizado directamente pelo algoritmo de aprendizagem. Segue-se um exemplo de codificação de um vector de características para a palavra-alvo *bass* no seguinte contexto [Jurafsky e Martin, 2000]:

> *An electric guitar and **bass** player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations peharps.*

Supondo criar um vector com as característica de colocação da palavra *bass* e as duas palavras à esquerda e à direita que a acompanham, teríamos a seguinte configuração:

[guitar, NN1, and, CJC, player, NN1, stand, VVB]

Onde: NN1 significa substantivo no singular, CJC conjunção coordenativa, VVB verbo no tempo presente.

Agora se olharmos para as características de co-ocorrência e assumindo que as 12 palavras mais frequentes recolhidas do corpus e que acompanham a palavra *bass* são: *fishing*, *big*, *sound, player, fly, rod, pound, double, runs, playing, guitar, band*. Assumindo uma janela de tamanho 10, ou seja, analisando as 10 palavras (sem ter em conta as suas posições) à volta da palavra-alvo (5 de cada lado), e verificando quais palavras mais frequentes aí se encontram, teremos seguinte vector:

[0,0,0,1,0,0,0,0,0,0,1,0]

As técnicas mais robustas de desambiguação utilizam uma combinação dos dois tipos de características num só vector.

Existem, actualmente, dois tipos de desambiguação por aprendizagem: Supervisionada e Não-Supervisionada; que diferem apenas no processo de treino do classificador de significados. Na aprendizagem supervisionada, o classificador é treinado sobre um conjunto representativo de

exemplos (construídos para cada palavra a ser categorizada com as características anteriormente descritas) previamente desambiguados (ou seja, com o significado expressamente indicado). Como resultado, é obtido um sistema de desambiguação capaz de classificar novas palavras. Os algoritmos de aprendizagem supervisionada mais utilizados para desambiguação são os classificadores de Bayes [Duda e Hart, 1973] e listas de decisão [Rivest, 1987]. Na aprendizagem não-supervisionada não é necessária qualquer classificação prévia sobre os dados para a fase de treino. Isto porque, o processo treino da aprendizagem não-supervisionada para a desambiguação de significados [Schutze, 1998] consiste no agrupamento dos exemplos ainda por desambiguar de acordo com uma medida de similaridade formando um conjunto de *clusters*.

O maior problema encontrado pelos algoritmos de aprendizagem baseados em corpus é o escalonamento, já que grande parte dos trabalhos de investigação que os utilizam têm como conjunto de treino e teste uma pequena amostra de palavras, variando entre 2 a 12 itens lexicais a serem desambiguados.

- **Desambiguação baseada em recursos lexicais**

A necessidade de criar um método de desambiguação para todas as palavras de uma linguagem levou a que um grande número de investigadores apostasse na utilização de bases de conhecimento ou léxicos já construídos (dicionários electrónicos, *thesaurus* ou criados manualmente para o efeito). Esta abordagem é a mais popular para desambiguação do sentido das palavras, tendo como principais fontes de conhecimento recursos disponíveis electronicamente: WordNet [Resnik, 1995; Sussna, 1993; Voorhees, 1993], LDOCE [Guthrie et. al, 1991] ou Rogest's International Thesaurus [Yarowvsky, 1992].

A informação presente em recursos lexicais tem sido utilizada de maneiras diferentes. Por exemplo, Lesk [1986] foi o primeiro a utilizar as definições de cada palavra presentes num dicionário electrónico. Cada uma destas definições era então comparada com todas as definições das palavras presentes no contexto original onde a palavra-alvo aparecia. O significado da palavra a desambiguar com maior sobreposição com as definições observadas das palavras adjacentes era o escolhido. Para exemplificar este raciocínio, Lesk apresentou o seguinte exemplo para desambiguar a palavra *cone* na frase *pine cone*, cujo significado pode ser um dos seguintes:

pine	1. *kinds of evergreen tree with needle-shaped leaves*
	2. *waste away through sorrow or illness*
cone	1. *solid body wich narrows to a point*
	2. *something of this shape whether solid or hollow*
	3. *fruit of certain evergreen trees*

Pelo método de Lesk, o significado escolhido para a palavra *cone* é *cone*#3 já que há uma maior sobreposição (das palavras *evergreen* e *tree*) com os significados de *pine*.

O principal problema encontrado nesta abordagem é o facto de as definições presentes nas entradas dos dicionários serem relativamente curtas e não estarem desambiguadas, não fornecendo assim material suficiente e fiável para a criação de um classificador. Isso pode ser resolvido incluindo informação que está directamente ligada com a palavra a desambiguar. Por exemplo, Mihalcea e Moldovan [2000], além de se basearem nas definições dos synsets (glosses), utilizam a hierarquia dos substantivos para comparar as palavras em comum.

Outro processo de utilização de recursos lexicais consiste em calcular a similaridade[5] entre uma palavra-alvo e todas as outras palavras presentes numa janela textual a sua volta. Esta abordagem de desambiguação de palavras foi proposta primeiramente por Yarowsky [1992]. Neste caso, o texto é tratado como uma lista de palavras (*bag of words*) não ordenadas e cada palavra presente na janela textual (sem considerar a sua posição relativa dentro desta) contribui para a identificação do significado da palavra-alvo. Posteriormente, esta abordagem foi sendo seguida por diversos trabalhos [Agirre e Rigau, 1996; Resnik, 1999; Lytinen, et. al, 2000] variando apenas no modo como é obtida a similaridade entre as palavras e o recurso lexical de base. No Capítulo 4, discutiremos mais detalhadamente esta abordagem, onde será descrita a sua adaptação aos mapas conceptuais.

[5] A similaridade semântica será amplamente abordada no próximo Capítulo (3).

3. Similaridade Semântica

A similaridade é considerada uma medida fundamental numa variedade de modelos computacionais. Os sistemas de categorização classificam novos exemplos baseados no seu grau de similaridade com algum protótipo, abstracção ou exemplo prévio. Algumas teorias da resolução de problemas assumem que novos problemas são resolvidos utilizando problemas similares como exemplo. O raciocínio baseado em casos fundamenta-se na pesquisa de casos numa base de conhecimento através de uma determinada medida de similaridade. Devido à sua importância, a similaridade tem sido o centro de um grande esforço de investigação. Deste modo, neste Capítulo, abordaremos a similaridade semântica desde a comparação entre conceitos (Sub Capítulo 3.1), em maior detalhe, até a sua aplicação directa para a comparação entre estruturas conceptuais mais complexas (Sub Capítulo 3.2), onde iremos focar diferentes áreas onde foi aplicado este mecanismo.

1. Similaridade ou Distância Semântica entre Conceitos

Neste trabalho procuramos estabelecer uma medida de similaridade entre mapas conceptuais. Mas antes disto, é preciso definir o que significa similaridade. Recorrendo a uma definição padrão do dicionário Oxford [Oxford, 2003], encontramos o termo similaridade como sendo "do mesmo tipo, natureza, ou quantidade; possuindo semelhança". Em termos cognitivos, Hahn e Chater [1998] sugerem a seguinte definição sobre similaridade: (1) é uma função das propriedades comuns; (2) é apresentada numa determinada escala de valores; (3) é de valor máximo entre dois objectos idênticos. Nós iremos utilizar o termo similaridade numa granularidade mais fina, no que se aplica aos significados das palavras e conceitos, para posteriormente expandirmos o raciocínio para todo o mapa conceptual.

Como se trata de um processo cognitivo, a avaliação de uma medida de similaridade semântica entre conceitos baseia-se na comparação com os resultados obtidos pelo método utilizado pelo raciocínio humano. Um teste psicológico efectuado em [Miller e Charles, 1991] pode ser utilizado como exemplo, em que uma lista de 30 pares de palavras foi analisada por indivíduos seleccionados com a seguinte tarefa: comparar cada par de palavras de acordo com a similaridade de significado entre elas numa escala de 0 a 4 valores, onde 0 implicaria que as palavras fossem completamente dissimilares, e 4 se fossem sinónimos perfeitos. Obtiveram uma performance (através da média entre as diversas respostas) de 95% de concordância entre os indivíduos.

Definir como duas palavras são consideradas próximas quanto aos seus significados é um problema conhecido como *similaridade semântica* [Ellman, 2000] sendo considerado um mecanismo fundamental para a área de *Information Retrieval* e Integração de Dados, na comparação de objectos que podem ser pesquisados ou integrados a partir de repositórios heterogéneos [Guarino et. al, 1999; Voorhees, 1998; Jiang e Conrath, 1997; Smeaton e Quigley, 1996; Lee et. al, 1993].

A própria definição de "proximidade" não é consensual na medida em que uns investigadores se contentam com uma similaridade de características enquanto que outros, como é o caso de Budanitsky e Hirst [2001], afirmam que o relacionamento (*relatedness*) semântico (ou o seu inverso – distância semântica) é um conceito mais geral que a similaridade. Entidades similares são geralmente relacionadas devido às suas características comuns (como é no caso de *bank – trust company*); mas entidades que não são similares podem estar relacionadas por alguma relação (por exemplo, os merónimos: *car-has-wheel*; os atributos: *beauty-has_atribute-ugly*; ou ainda uma relação de associação frequente: *penguin-Antarctica*). Pela sua maior abrangência e proximidade com o senso comum, os sistemas computacionais necessitam mais do que simples medidas de similaridade, pois esperam relacionar semanticamente duas palavras, mesmo que estas não possuam características comuns mas que, por força do uso, apareçam regularmente relacionadas.

Existem, actualmente, duas abordagens padrão para determinar a similaridade (ou distância) semântica entre conceitos ou palavras: a primeira é baseada no cálculo da distância entre palavras numa hierarquia semântica; enquanto que, a segunda se fundamenta no conteúdo de informação (*information content*) partilhado entre palavras. Mais adiante analisaremos em profundidade cada um destes tipos de abordagens para a atribuição de uma similaridade semântica entre palavras, assim como trabalhos desenvolvidos com base em ambos os tipos.

1.1. Abordagem baseada na Distância Semântica

Tradicionalmente, a comparação entre significados de palavras é conseguida calculando a distância semântica entre as suas definições presentes numa mesma ontologia. Nesta abordagem, a representação da informação é considerada como uma rede semântica, e a distância semântica é, seguindo um raciocínio intuitivo, o número de ligações (arcos) entre os conceitos ao longo do caminho mais curto que os liga [Rada et. al, 1989; Rigau et. al, 1997]. Deste modo, é feita uma analogia ao modelo da memória humana introduzido por Collins e Quillian [1969], onde é

defendida a hipótese[1] de que os indivíduos "armazenam" os conceitos numa estrutura conceptual hierárquica.

Rada e Bicknell [1989] utilizaram o conceito de distância conceptual baseada numa técnica simples de contagem de arcos (além de outras) no WordNet para ordenar *queries* codificadas no *thesaurus* MeSH (*Medical Subject Headings*) que seriam pesquisadas na base de dados bibliográfica MEDLINE. Posteriormente, Resnik [1995] avaliou a performance desta abordagem, reproduzindo os testes padrão de similaridade efectuados por Miller e Charles [1991], onde obteve uma fraca prestação (66% de precisão), comparada com o desempenho humano.

Uma vez que, por esta abordagem, a similaridade semântica é baseada na contagem de arcos do caminho mais curto entre dois conceitos [Rada e Bicknell, 1989], é necessário haver uma igual distribuição de tamanho e valor entre os arcos, ou seja, uma ligação entre dois nós deve ter o mesmo peso que qualquer outra. Se isto não acontecer, deverá ser aplicada uma forma de ponderação para compensar eventuais disparidades, como em [Agirre e Rigau, 1996], onde foi utilizado o conceito de densidade conceptual relativa na hierarquia taxonómica do WordNet para uma tarefa de desambiguação dos sentidos das palavras.

Sussna [1993] aperfeiçoa o método simples de contagem de arcos entre nós ponderando agora os diferentes tipos de relações que compõem o menor caminho entre dois conceitos. A métrica apresentada é válida apenas para substantivos e utiliza como base a rede semântica do WordNet considerando apenas algumas relações: sinónimos, antónimos, hiperónimos, hipónimos, holónimos e merónimos. Além de atribuir pesos diferentes para cada uma das relações que podem constituir o menor caminho entre dois conceitos, Sussna apresenta uma medida de distância semântica que se baseia inclusive no número de relações do mesmo tipo que "saem" de um conceito (*type-specific fanout* – TSF) e a profundidade na hierarquia a que se encontram ambos os conceitos a serem comparados.

A similaridade semântica entre conceitos ou palavras é, geralmente, considerada como inversamente proporcional à distância semântica numa taxonomia tal como é a hierarquia do WordNet [Li et. al, 1995]. Como somente os substantivos e verbos estão organizados hierarquicamente no WordNet, esta hipótese só pode ser válida para estas duas classes gramaticais. Para adjectivos e advérbios, uma medida mais simbólica pode ser adoptada, como é o caso da proposta realizada por [Stetina et. al, 1998] em que advérbios e adjectivos podem ter uma distância quantificada numa escala de 0 a 1. O valor 0 (muito próximos) é atribuído se

[1] Tal como visto na discussão sobre os Mapas Conceptuais (Secção 2.1).

pertencerem ao mesmo synset ou forem antónimos, 0.5 (próximos) se estiverem de alguma forma relacionados, ou 1 para todo os casos restantes. Não foi encontrada qualquer proposta para a determinação da similaridade semântica entre palavras ou conceitos de diferentes classes gramaticais. Este facto pode ser justificado por uma das definições clássicas de similaridade implicar a possibilidade de permuta de dois objectos similares sem alterar o contexto original. No caso de palavras de diferentes classes gramaticais isto pode não ser possível pelo facto de estas assumirem diferentes funções entre si.

O WordNet por si só não apresenta nenhuma informação sobre a distância semântica entre conceitos, mas sim algumas relações semânticas e lexicais entre estes. Hirst e St-Onge [1998] são dos muitos entusiastas do WordNet que têm criticado esta lacuna, e que, além disso, apontam o facto de existirem domínios que estão muito mais discriminados que outros, havendo assim um maior número de ligações entre dois conceitos ou palavras que pertencem a uma mesma família mas que podem representar uma distância física maior. Os autores citam o exemplo de *more stew than steak*, no qual o padrão *more…than…* associa semanticamente duas palavras relacionadas (*stew* e *steak*), enquanto que estão separadas no WordNet por 6 *synsets*, o que não reflecte claramente a real distância semântica entre elas, já que ligações, como por exemplo, entre *public* e *professionals*, estão mais próximas considerando que o número de *synsets* que as separam é menor (4).

Leacock e Chodorow [1998] propuseram uma medida de similaridade semântica que utiliza as relações de generalização/especialização (hiperónimo/hipónimo) presentes no WordNet para encontrar o menor caminho entre dois substantivos a e b de acordo com a fórmula 3.1.

$$sim(x, y) = -\log \frac{len(x, y)}{2D} \qquad \textbf{(3.1)}$$

onde *len(x,y)* é o número de *synsets* que existem no menor caminho que liga os conceitos **x** e **y** na árvore taxonómica do WordNet (só sendo consideradas relações *isa* e *kind_of*); e **D** é a profundidade máxima desta árvore, ou seja, o maior caminho desde um synset mais geral a um mais específico.

Em oposição à maioria das medidas de similaridade propostas, que somente consideram a hierarquia entre os substantivos no WordNet, Hirst e St-Onge [1998] utilizam várias relações semânticas que abrangem todas as classes gramaticais para estabelecer um grau de similaridade entre conceitos. Cada relação no WordNet é classificada quanto à sua direcção como sendo horizontal, vertical ascendente, ou vertical descendente (ver Quadro 3.1). A similaridade entre duas palavras é determinada (e pode ser quantificada como veremos no Capítulo 4) como sendo **extra-forte**, **forte** ou **meio-forte**. Uma similaridade **extra-forte** é obtida somente por uma

palavra e a sua réplica. Já uma similaridade **forte** é aplicável quando duas palavras são ligadas (de acordo com os *synsets* onde ocorrem) por uma relação horizontal. E, nos restantes casos, uma similaridade medianamente forte entre duas palavras ocorre quando existe um caminho possível entre os *synsets* onde estas palavras ocorrem. Devido à grande interligação presente na rede semântica do WordNet, os autores restringiram o conceito de "caminho" à seguinte definição: é uma sequência de 2 a 5 arcos (relações) entre *synsets* e só é considerado válido se forem respeitadas as seguintes regras:

i. Nenhuma outra relação de diferente direcção deve preceder uma relação ascendente.

ii. É permitida no máximo uma mudança de direcção, excepto se for utilizada uma relação horizontal para a transição entre uma ascendente e uma descendente.

Direcção	Relações
horizontal	sinónimo (adjectivos, advérbios, substantivos, verbos), antónimo (adjectivos, advérbios, substantivos, verbos), atributo (adjectivos, substantivos), similar (adjectivos), *also_see* (adjectivos, verbos).
ascendente	hiperónimos (substantivos), holónimos (substantivos).
descendente	hipónimos (substantivos, verbos), merónimo (substantivos), causa (verbos), *entailment* (verbos).

Quadro 3.1: Classificação [Hirst e St-Onge, 98] dada a algumas relações do WordNet.

Ao contrário dos outros graus de similaridade, o valor da meio-forte é ponderada de acordo com a fórmula 3.2.

$$sim(x, y) = c - path(x, y) - k * changes\_of\_direction \quad \textbf{(3.2)}$$

onde C e k são constantes, *path(x,y)* é o número de relações existentes no caminho de x a y, e, finalmente, *changes_of_direction* é o número de mudanças de direcção presentes neste mesmo caminho. Desta forma, a equação é inversamente proporcional ao comprimento do caminho que liga dois conceitos, quanto maior, menor é o grau de similaridade.

1.2. Abordagem baseada no Conteúdo de Informação

Nesta abordagem, a similaridade semântica é analisada em função do caminho que liga dois nós num espaço conceptual hierárquico de modo a encontrar o conceito mais específico que os generalize (que podemos denominar como conceito **generalizador comum mais específico** ou **gcme**). Este conceito será o ponto de convergência entre os dois conceitos analisados, uma vez que é o ponto limite onde a informação é partilhada entre ambos conceitos. Uma vez que neste caso, é focada somente a hierarquia de conceitos de uma rede semântica, esta abordagem requer informação menos detalhada sobre a estrutura da taxonomia.

Outro termo utilizado é o **conteúdo de informação** (*information content*) que um conceito representa introduzido pela Teoria da Informação [Ross, 1976] representado pela fórmula 3.3.

$$CI(c) = -\log P(c) \quad \textbf{(3.3)}$$

onde P(c) é a probabilidade de ocorrência do conceito **c** ou qualquer conceito que seja generalizado por **c**.

Desta forma, quanto maior é a probabilidade de ocorrência de um conceito **c**, mais abstracto é (já que um conceito mais geral, tem a sua probabilidade acrescida de todos os seus mais específicos). E consequentemente, de acordo com a fórmula 3.3, quanto mais abstracto, menor é o conteúdo de informação que um conceito possui. Isto implica que *P(c)* cresce monotonicamente à medida que subimos na hierarquia. No topo da hierarquia, foi adicionado artificialmente um conceito que engloba todos os mais gerais presentes no WordNet cuja a probabilidade de ocorrência é 1, e consequentemente com conteúdo de informação de valor 0.

A similaridade entre os conceitos **a** e **b** pode ser definida então como o conteúdo de informação (fórmula 3.1) do conceito **gcme** de **a** e **b** [Resnik, 1995]:

$$sim(a,b) = -\log P(gcme_{a,b}) \quad \textbf{(3.4)}$$

No caso das heranças múltiplas que podem ocorrer no WordNet, ou seja, um nó ter directamente mais do que um conceito generalizador, a similaridade é determinada utilizando como base o conceito **gcme** mais específico entre as diversas hipóteses.

A utilização de um **gcme** para o cálculo de similaridade semântica entre dois conceitos pode ser fundamentada na aprendizagem baseada em explicação (*explanation-based learning*) [Mitchell et. al, 1986], que apresenta a noção de abdução e generalização, onde, dado um conjunto de exemplos, é necessário encontrar a melhor generalização ou explicação sobre eles. Tendo os exemplos E1 e E2, uma boa generalização é o exemplo E3 que generaliza E1 e E2 da maneira mais económica. Se E e E' são ambos generalizações de E1 e E2, e E é um subtipo de E' então a fórmula lógica $\forall x E(x) \rightarrow E'(x)$ é verdadeira, e portanto, E é uma generalização mais económica uma vez que uma hipótese mais forte admite menos modelos.

Utilizando a hierarquia taxonómica do WordNet (relações *isa*) [Resnik, 1995], um exemplo de determinação do conceito **gcme** é apresentado na Figura 3.1.

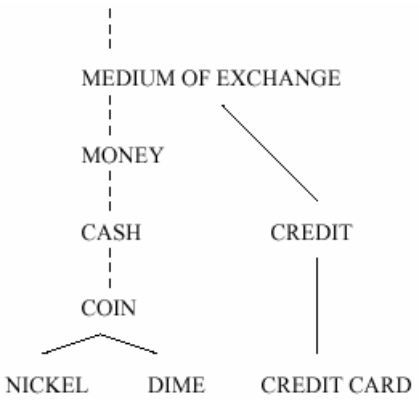


Figura 3.1: O conceito *coin* é considerado **gcme** de *nickel* e *dime* na hierarquia de substantivos do WordNet. Enquanto que *nickel* e *credit_card* são generalizados pelo conceito mais específico *medium_of_*exchange. Exemplo retirado de [Resnik, 1995].

A probabilidade de ocorrência de um dado conceito pode ser definida como a soma das probabilidades de ocorrência das palavras num dado corpus que referenciem este conceito [Resnik, 1995]. O autor utilizou para o efeito, o *Brown Corpus* [Kučera. e Francis, 1967] para contabilizar a frequência e probabilidade de ocorrência dos substantivos presentes no WordNet. No entanto, esta técnica não tem em conta a ambiguidade do significado das palavras, já que uma palavra presente no texto pode referenciar mais do que um conceito. Este aspecto foi de algum modo corrigido por [Richardson e Smeaton, 1995] ao dividir a frequência de uma palavra pelo número de significados que pode possuir (ou seja, o seu grau de polissemia ou homonímia). Por sua vez, esta estimação também poderá ser equívoca uma vez que a distribuição dos significados das palavras no corpus utilizado pode não ser linear.

O método proposto por Resnik pode ser melhor compreendido através do seguinte exemplo: queremos determinar as similaridades entre os conceitos (*car, bicycle)* e *(car, fork)*; a Figura 3.2 representa um extracto da hierarquia de substantivos do WordNet; o número associado a cada nó indica o seu respectivo conteúdo de informação de acordo com a fórmula 3.3. De acordo com os valores apresentados, a similaridade entre *car* e *bicycle* é o conteúdo de informação do seu respectivo conceito **gcme** *vehicle* de valor 8.30; enquanto que a similaridade entre *car* e *fork* é de 3.53. Estes resultados aproximam-se da nossa intuição de que um carro é mais similar a uma bicicleta que a um garfo.

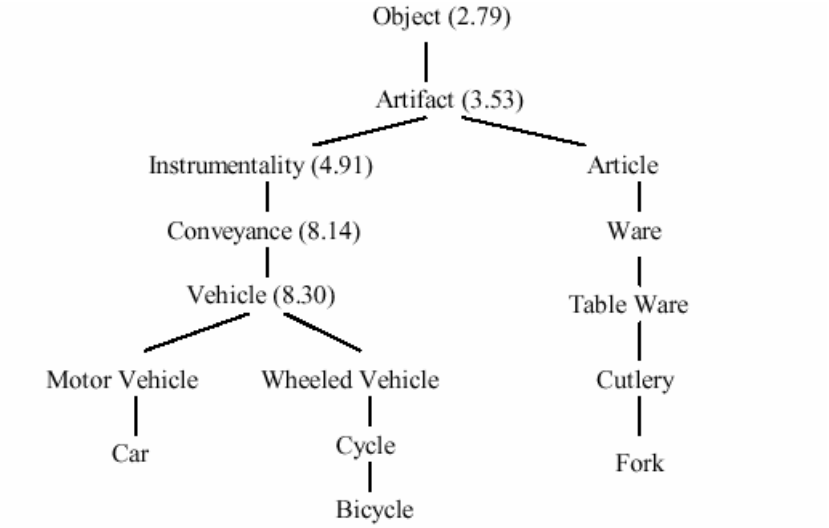


Figura 3.2: Extracto da hierarquia de substantivos do WordNet com conteúdo de informação associado. Exemplo retirado de [Jiang e Conrath, 1997].

É importante notar que, pela abordagem seguida por Resnik, não há distinção na similaridade semântica entre qualquer par de conceitos desde que se mantenha o mesmo conceito **gcme**. No exemplo ilustrado, a similaridade semântica entre *bicycle* e *fork* é a mesma de entre *bicycle* e *table_ware*. Em resposta a este problema, Jiang e Conrath [1997] propuseram uma abordagem híbrida para a similaridade semântica que combina o conteúdo da informação com o método da distância semântica (abordado na Secção anterior) através da seguinte métrica:

$$dist(x, y) = CI(x) + CI(y) - 2 * CI(gcme_{x,y}) \quad \textbf{(3.5)}$$

onde a equação para o cálculo do conteúdo de informação está definida em (3.3).

Os autores fazem a analogia da aplicação desta métrica a um espaço multidimensional semântico (considerando a hierarquia de substantivos do WordNet) onde cada nó (conceito) encontra-se num eixo específico e tem uma massa (o seu conteúdo de informação), e a distância semântica entre dois nós que estão no mesmo eixo (ou seja, um é a generalização do outro) é apenas a diferença da sua massa semântica; enquanto que se estiverem em eixos separados (é preciso encontrar um **gcme**), é a soma das duas distâncias em relação a um conceito comum. Neste trabalho, os autores realizaram experiências utilizando como corpus de base para o cálculo da informação de conteúdo o *SemCor*[2] e observaram que, na altura, este corpus só incluía metade das palavras presentes no WordNet, diminuindo a abrangência da medida formulada. Apesar desta limitação, eles afirmaram que obtiveram uma precisão de 82.8% na comparação de 30 pares de substantivos. Como veremos na Secção 4.3, uma adaptação desta medida será utilizada no cálculo de similaridade entre conceitos.

[2] Este corpus está descrito no Capítulo (2) de discussão do WordNet (Sub Secção 2.1.3).

Em [Lin, 1998], o autor propõe uma análise teórica do problema da similaridade, e chega a uma medida genérica (3.6) que afirma ter uma abrangência, ao contrário das outras medidas até aqui propostas, aplicável a diferentes domínios que sigam um modelo probabilístico, tais como: valores ordinais, vectores de características, palavras, conceitos numa taxonomia, entre outros. Em relação ao domínio de uma taxonomia, a sua medida é adaptada (3.7) para a utilização de um conceito **gcme** e mostra ser um ligeira variação de (3.5), obtendo, segundo o autor, o desempenho de 83% de acordo com experiências sobre 28 pares de conceitos.

$$sim(x,y) = \frac{CI(common(x,y))}{CI(description(x,y))} \quad \textbf{(3.6)}$$

onde *common(x,y)* é a quantidade de informação necessária para descrever a semelhança entre x e y, e *description(x,y)* a necessária para completamente descrever o que são x e y.

$$sim(x,y) = \frac{2 * CI(gcme_{x,y})}{CI(x) + CI(y)} \quad \textbf{(3.7)}$$

1.3. Resumo das Medidas Apresentadas

No Quadro 3.2 é apresentado um resumo das medidas referidas nas secções anteriores. Algumas destas medidas serão testadas e exemplificadas em pormenor no próximo Capítulo.

Ao analisarmos a similaridade semântica entre conceitos (e eventualmente entre relações), adquirimos conhecimento de base para analisar o mapa conceptual (formado por conceitos e relações) como um todo. Mas será esta uma tarefa linear, na medida em que comparando as partes conseguimos no fim relacionarmos o todo? É a esta questão que procuraremos responder no próximo Sub Capítulo, tentando em primeiro lugar descobrir como é o processo humano de comparação, para em seguida vislumbrar que metodologias já existem aplicadas a estruturas conceptuais (comos os grafos conceptuais).

Principais Medidas de Similaridade entre Conceitos		
Base da Abordagem	**Autores**	**Resumo**
Distância Semântica	Rada e Bicknell (1989)	Contagem simples de arcos do caminho mais curto entre dois conceitos na árvore hierárquica de substantivos do WordNet
	Sussna (1993)	Variação do método anterior. Considera, além das relações hierárquicas entre substantivos, as relações de antónimos, holónimos e merónimos.
	Agirre e Rigau (1996)	Contagem de arcos de forma ponderada de acordo com o peso das relações hierárquicas entre substantivos (densidade conceptual).
	Hirst e St. Onge (1998)	Considera a direcção e o comprimento de diferentes relações semânticas que liguem dois conceitos. Propõem uma medida que relaciona adjectivos, advérbios, substantivos e verbos entre si.
	Leacock e Chodorow (1998)	Variação do método original de contagem de arcos que utiliza uma função inversa logarítmica e normaliza os comprimentos dos caminhos entre substantivos.
Conteúdo da Informação	Resnik (1995)	Localiza o conceito generalizador comum mais específico (**gcme**) entre dois substantivos na árvore hierárquica do WordNet, quantificando, de seguida, o seu Conteúdo de Informação.
	Richard e Smeaton (1995)	Variação do método anterior que quantifica também a ambiguidade das palavras alvo de comparação.
Híbridas	Jiang e Conrath (1997)	Contabiliza a distância semântica entre dois substantivos através do seu **gmce** utilizando o conteúdo da informação de cada um para quantificar o seu peso no resultado final.
	Lin (1998)	Variação da medida anterior que normaliza a distância semântica entre substantivos.

Quadro 3.2: Resumo das principais medidas de similaridade entre conceitos apresentadas no Sub Capítulo 3.1 (Similaridade entre Conceitos)

2. Similaridade entre estruturas conceptuais: Aplicações em diferentes áreas

Em diversas áreas, a similaridade entre objectos, sejam eles simples ou estruturados, é uma componente essencial para sistemas computacionais de pesquisa, classificação, reconhecimento e validação. E, até que ponto a similaridade pode ser um fim ou medida absoluta, mas sim um método de comparação relativa entre objectos pertencentes a um conjunto? Diversas abordagens foram propostas para o problema, umas centradas no conteúdo da informação que os objectos representam, outras mais voltadas para a estrutura na qual os objectos se organizam, e finalmente, abordagens híbridas que combinam o conteúdo e forma dos objectos. Serão apresentadas as visões e abordagens sobre a similaridade para algumas destas áreas de estudo, sendo analisadas para cada abordagem, a sua aplicabilidade para os mapas conceptuais e complexidade computacional.

2.1. A Similaridade como um Processo Cognitivo: Contraste de Características vs. Alinhamento Estrutural

Como base de todo o processo cognitivo, o processo de comparação da similaridade entre objectos, segundo os psicólogos, é crucial para o raciocínio analógico e reconhecimento de padrões [Love, 2000]. Existem duas correntes dentro da psicologia que se propõem estabelecer o que torna dois objectos similares: a tese de que os objectos são descritos por um conjunto de características, e que a similaridade é obtida através do **contraste de**stas **características** (comuns faces às diferentes) foi proposta inicialmente por [Tversky, 1977]; e posteriormente, surgiu a ideia defendida por [Markman e Gentner, 1993], que se baseia num modelo de similaridade estruturado tripartido em diferentes tipos de características: semelhanças comparáveis (*alignable similarities*), diferenças comparáveis (*alignable differences*) e diferenças não comparáveis (*non-alignable differences*), denominado **alinhamento estrutural**. A seguir, ambas as correntes serão exemplificadas e analisadas em detalhe.

2.1.1. Contraste de Características (*Feature-Constrast Model*)

Esta abordagem [Tversky, 1977] formaliza o raciocínio intuitivo de que a similaridade aumenta com as semelhanças e decresce com as diferenças. Cada entidade a ser comparada é representada por um conjunto de características seleccionadas de um conjunto universal, sendo consideradas características comuns: aquelas que pertencem à intersecção entre os conjuntos de características próprias; e características distintivas: as características que somente pertencem a um dos conjuntos. Para elucidar este raciocínio, *tomate* e *cereja* podem ser consideradas similares devido às suas características comuns: redondos, frutos, suculentos, pele lisa, vermelhos, etc. Da mesma forma esta similaridade é diminuída pelas diferenças: tamanho, presença de caroço, entre outras.

O autor realizou uma série de experiências onde concluiu que as características comuns são mais determinantes para a nossa avaliação cognitiva de similaridade face às características distintivas. Este facto é justificado pelo facto de que quando estamos a comparar a similaridade de objectos, atribuímos, geralmente, mais importância às semelhanças face às diferenças.

A similaridade entre objectos defendida por Tversky é definida pela seguinte equação:

$$sim(x, y) = \alpha|X \cap Y| - \beta|X - Y| - \gamma|Y - X| \qquad \textbf{(3.8)}$$

onde $\alpha, \beta, \gamma \geq 0$.

Esta equação (3.8), em que X e Y são os conjuntos de características respectivamente dos objectos x e y, não é simétrica, na medida em que *sim(x,y)* pode ser diferente de *sim(y,x)*. A razão desta assimetria é explicada pelo autor devido ao facto de quando comparamos dois

objectos, por exemplo em *sim(x,y)*, estamos a focar um x, denominado *focus*, em detrimento de um y, a *base* da comparação. Tversky utiliza como exemplo a similaridade entre "*Coreia do Norte"* e *"China"*, que nesta ordem aparenta ser maior do que o contrário, isto porque conhecemos muito mais características sobre o segundo país. Esta assimetria pode ser indicada na atribuição de pesos aos diversos termos da equação (3.8) através de $\beta>\gamma$.

Tversky defende claramente que não propõe uma métrica, uma vez que a equação apresentada não respeita nenhuma das seguintes propriedades geométricas:

- simetria: *sim(x,y)* pode ser diferente de *sim(y,x)*;
- minimalidade: nem todos os objectos idênticos são vistos como igualmente similares; os objectos mais complexos (e por sua vez descritos por mais características) que são idênticos (dois gémeos) são considerados mais similares que objectos idênticos mais simples (dois quadrados);
- desigualdade triangular: esta pode ser violada quando dois objectos, A e B, ao partilharem uma característica comum e, B e C partilharem também uma outra característica, não significar que A e C tenham qualquer semelhança. Por exemplo [Goldstone, 1999], sabendo que "*lâmpada*" e "*lua*" são ambas brilhantes e que *"lua"* e "*bola*" são redondas, não implica que *"lâmpada"* e *"bola"* sejam similares.

Outras medidas baseadas em operações sobre conjuntos evoluíram da noção que a similaridade entre duas entidades é baseada nas características em comum que possuem (nomeadamente X intersecção Y $\rightarrow |X \cap Y|$), dentre estas [Manning e Schutze, 1999]:

- O coeficiente Dice (3.9) normaliza a similaridade entre diferentes entidades tendo em conta o tamanho do conjunto de características de cada entidade.

$$sim(x, y) = \frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid} \qquad \textbf{(3.9)}$$

- Por sua vez, o coeficiente Jaccard (3.10) penaliza um pequeno número de características partilhadas entre as entidades de uma maneira mais explícita que a medida anterior. Ambas as medidas variam entre 0.0 (completa dissimilaridade) a 1.0 (total similaridade), no entanto, o coeficiente Jaccard atribui um valor de similaridade menor para os casos em que há poucas características comuns entre as entidades comparadas.

$$sim(x, y) = \frac{\mid X \cap Y \mid}{\mid X \cup Y \mid} \qquad \textbf{(3.10)}$$

- O coeficiente de sobreposição (*overlap*) (3.11) também é conhecido como uma medida de inclusão, pois atribui um valor de 1.0 caso uma das entidades possua todas as características da outra (ou seja, $X \subseteq Y$ ou $Y \subseteq X$).

$$sim(x, y) = \frac{| X \cap Y |}{\min(| X |, | Y |)} \quad \textbf{(3.11)}$$

- O coeficiente Cosseno (3.12) é idêntico ao coeficiente Dice no valor atribuído a conjuntos com o mesmo tamanho, mas penaliza menos do que este último (atribuindo um grau de similaridade maior), para o caso de compararmos conjuntos de características com tamanhos muito diferentes entre si. Esta particularidade do coeficiente de cosseno permite comparar entidades cuja quantidade de informação é diferente sobre cada uma, em que o facto deste conhecimento desproporcional existir não irá afectar a similaridade entre ambas.

$$sim(x, y) = \frac{| X \cap Y |}{\sqrt{| X | x | Y |}} \quad \textbf{(3.12)}$$

2.1.2. Alinhamento Estrutural (*Structural Alignment*)

Em oposição à abordagem anterior, Markman e Gentner [1993] ressaltam a importância da estrutura durante o processo de comparação de similaridade. Neste caso, a similaridade entre estruturas conceptuais é baseada num alinhamento lógico, ou mapeamento estrutural, entre as representações das estruturas. Para uma característica ser considerada comum, deve existir em ambas as estruturas, e além disso devem ocupar a mesma importância ou posição relativa (ou seja, as características devem ser mapeadas segundo uma função de mapeamento, como no isomorfismo entre grafos). As características partilhadas que não ocupam a mesma posição relativa em cada estrutura conceptual não são avaliadas no processo de comparação (*non-alignable*). Para exemplificar este raciocínio, os autores apresentam a comparação de duas universidades (X e Y). Ambas possuem departamentos de Ciências da Computação que dispõem de super computadores de última geração, que neste aspecto são consideradas características comuns. Mas ao contrapormos o sucesso da universidade X na expansão dos seus programas de investigação, face à universidade Y, que se encontra estagnada nesta vertente, a investigação constitui uma diferença que é pesada (*alignable difference*) no processo de comparação. Uma diferença que não deve ser levada em conta (*non-alignable differences*) na comparação entre as duas universidades pode ser, por exemplo, o facto de, na universidade X haver um curso que não exista na Y.

De acordo com as suas experiências, Markman e Gentner concluíram que a maioria dos indivíduos consultados considera as diferenças comparáveis (*alignable difference*) mais

importantes do que as não-comparáveis (*non-aliegnable differences*) no processo de comparação de similaridade. Isto talvez possa ser explicado pela capacidade intuitiva de conseguirmos, mais facilmente, listar as principais diferenças entre conceitos muito similares do que entre conceitos que nada têm em comum. Por exemplo, é mais fácil contabilizar as diferenças entre os conceitos de "hotel" e "motel" do que entre "hotel" e "motocicleta".

Esta abordagem, que é a base da principal investigação sobre o raciocínio analógico, vem propor uma visão da similaridade adaptada à comparação entre objectos estruturados (que podem estar organizados proposicional e/ou hierarquicamente) face aos modelos geométricos ou baseados em características, na medida que considera a estrutura interna deste mesmos objectos e não apenas o seu conjunto de coordenadas ou características. A comparação entre objectos evolui de uma simples verificação de características comuns para, em primeiro lugar, determinar que elementos (ou características) podem ser comparados (ou alinhados) num mapeamento estrutural. Esta possível equivalência entre características é obtida através da análise da sua função no objecto a que se refere. Caso duas características (cada uma pertencente a um objecto diferente) tenham funções similares, podem ser comparadas na determinação da similaridade entre os objectos. Por exemplo: um carro que possua uma roda verde quando comparado a um camião de capota verde, não lhe é mais similar por partilhar o atributo verde, já que as peças referidas não são equivalentes. Esta particularidade é demonstrada pelos autores que afirmam a influência das características comuns na similaridade somente se estas pertencerem a objectos que possam ser substituídos por serem equivalentes na sua funcionalidade.

2.2. Avaliação da Aprendizagem por Mapas Conceptuais

A área das Ciências da Educação é uma das principais que, actualmente, estuda e utiliza os mapas conceptuais como uma ferramenta de **ensino** (pela possibilidade do professor expor uma matéria através dos seus conceitos-chave e os interligar), **aprendizagem** (os estudantes são incentivados a construir os seus próprios mapas conceptuais da matéria leccionada) e **avaliação** (os mapas conceptuais permitem saber se e como está a evoluir a aprendizagem dos alunos) [Zimmaro e Cawley, 1998]. Esta última tarefa de avaliação dos mapas conceptuais tem sido alvo de grande discussão quanto à natureza da abordagem: os mapas devem ser avaliados quantitativamente ou qualitativamente em relação à sua coerência com o domínio representado? De seguida, são apresentadas as diferentes abordagens a esta temática e até que ponto é possível automatizar este processo cognitivo desempenhado pelo professor.

Em [Hibberd et. al, 2002], é feito um estudo comparativo de diversas abordagens para a validação, por parte de um professor, de mapas conceptuais construídos por alunos para análise

da aquisição de conhecimentos. Em alguns trabalhos analisados tem sido apontada a utilização de uma grelha de avaliação. Estes sistemas de avaliação propostos apresentam alguma arbitrariedade de valores escolhidos para diferentes componentes de um mapa, não havendo uma harmonização entre eles da classificação escolhida. Deste modo, a avaliação de mapas conceptuais passa a ser um processo subjectivo, na medida em que cada ponto de vista considera diferentes componentes como sendo os principais indicadores da validade do mapa. Por exemplo, enquanto que algumas propostas de validação valorizam apenas as proposições correctas (formada pelas ligações entre conceitos) existente num mapa conceptual, sem ter em conta as proposições mal-formadas [McClure e Bell, 1990; Lomask et. al, 1992; Novak e Gowin, 1984], outras propõem diminuir a pontuação de uma mapa se este apresentar proposições incorrectas [Hoz et. al, 1990].

Outra abordagem utilizada consiste no cálculo da frequência de componentes específicos utilizados na construção dos mapas conceptuais, permitindo comparar os mapas conceptuais ao longo de diferentes dimensões (conceitos, relações, proposições, etc.). Mas o facto de apenas contabilizar a frequência, não tendo em conta a localização dos componentes no mapa e a sua interligação com outros elementos do mapa, não possibilita avaliar a relação intrínseca existente entre o conteúdo do mapa e a sua estrutura.

GoldSmith e Davenport [1990] propõem uma análise quantitativa de proximidade (*closeness analysis*) que tem em conta o conteúdo e estrutura na comparação entre mapas conceptuais. Esta abordagem fornece uma indicação da qualidade e do grau de relacionamento entre mapas conceptuais que sejam constituídos por um conjunto de nós comuns apenas distinguindo-se por possuírem diferentes interligações entre os nós. Este facto vem limitar a abrangência da aplicação da abordagem, uma vez que somente mapas com o mesmo número e mesmos nós podem ser comparados. Isto obriga a que os alunos utilizem sempre todos os conceitos de um conjunto fornecido previamente à construção do mapa, o que leva à uma uniformização indesejável dos mapas construídos.

A análise qualitativa rejeita qualquer hipótese de grelha genérica para a pontuação de mapas conceptuais. O autor sugere a análise detalhada e específica a cada domínio de estudo, para ser possível caracterizar simultaneamente o conteúdo e estrutura dos mapas conceptuais. Apesar de no momento ser esta a abordagem com maior adesão, não existem ainda técnicas concretas para a análise qualitativa de mapas conceptuais.

Apesar de inicialmente ter sido defendido um ponto de vista quantitativo para a comparação entre mapas conceptuais onde pontuava-se cada conceito ou proposição correctamente definidos (ou por outro lado, penalizava-se assumpções incorrectas apresentadas

sobre o domínio em causa), actualmente, é mais aceite na área das Ciências da Educação como uma visão qualitativa onde o mapa conceptual é analisado pelo professor [Hibberd et. al, 2002] comparando-o (processo cognitivo) a um mapa por si construído que representa o modelo a seguir. Neste processo, é claramente respeitado pelo professor o facto de não haver uma única representação do domínio debatido e o grau de especificidade ou detalhe poder variar de mapa para mapa.

2.3. Modelo Computacional: Comparação de Grafos Conceptuais

Montes-y-Gómes e López-López [2000] propuseram uma abordagem híbrida de comparação de grafos conceptuais[3] que tem em conta a estrutura e o conteúdo dos grafos. A sua proposta baseia-se num algoritmo de comparação parcial de grafos conceptuais [Myaeng e López-López, 1992] e no coeficiente de Dice[4] consistindo basicamente de duas etapas:

- Definir a sobreposição existente entre dois grafos conceptuais G_1 e G_2, através da construção do grafo de sobreposição $G_c = G_1 \cap G_2$;
- Medir a similaridade relativa à sobreposição existente entre os dois grafos.

O grafo de sobreposição G_c é constituído por todos os conceitos comuns e relações que interligam os mesmos conceitos em ambos os grafos G_1 e G_2. Este grafo pode ser directamente construído com base no algoritmo de comparação genérica de grafos por Myaeng e López-López [1992]. Neste trabalho, os autores apresentam um algoritmo flexível para encontrar os máximos subgrafos comuns entre dois grafos G_1 e G_2. A flexibilidade do algoritmo é apontada pelo facto deste permitir a adaptação da comparação entre grafos para considerar:

- todos os conceitos e relações entre os grafos G1 e G2;
- apenas os conceitos entre os grafos G1 e G2;
- apenas as relações entre os grafos G1 e G2;
- apenas as direcções das relações entre os grafos G1 e G2 (*unconstrained matching*).

Um exemplo de grafo de sobreposição de dois grafos é apresentado na Figura 3.3. Neste exemplo, apenas os conceitos A, B e C pertencem a ambos os grafos G_1 e G_2. Embora existam os seguintes arcos que unem os conceitos A, B e C: A—B, A— C e B — C, somente os dois primeiros são comuns aos dois grafos. Apesar de não estar explícito na Figura 3.3, um arco para ser considerado parte do grafo de sobreposição de dois grafos, deve possuir a mesma designação (etiqueta) e direcção em ambos os grafos.

[3] Tal como visto no Sub Capítulo 2.1 (Mapas Conceptuais), os grafos conceptuais são considerados uma versão dos mapas conceptuais com um maior formalismo de sintaxe e semântica na sua organização.
[4] Apresentado na Sub Secção 3.2.1.1 (Contraste de Características).

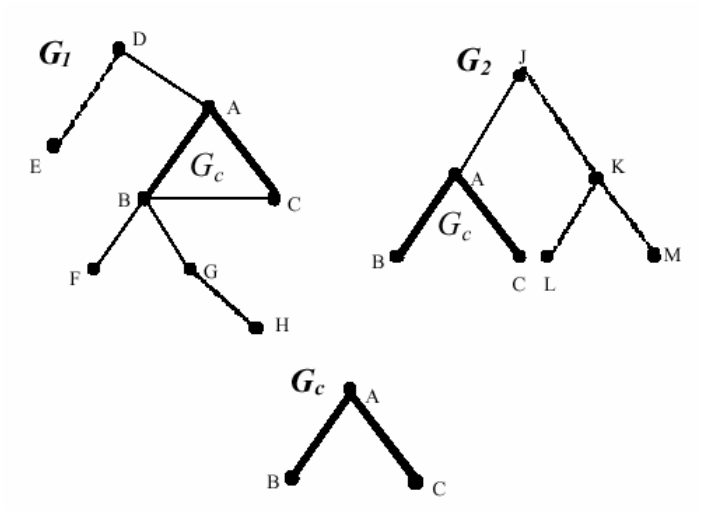


Figura 3.3: Grafo de sobreposição G_c dos grafos G_1 e G_2.
Extraído de [Montes-y-Gómes e López-López, 2000].

Para medir a similaridade baseada na intersecção entre os dois grafos é proposta uma medida (com escala de [0,1], onde o valor 0 representa completa dissimilaridade, e o valor 1 a equivalência total ente dois grafos) que tem em consideração os dois tipos de componentes de um grafo conceptual: conceitos e relações. Deste modo, a medida de similaridade (3.13) é uma função cumulativa de duas componentes: a similaridade conceptual e a similaridade relacional, respectivamente. Devido à natureza secundária da similaridade relacional, já que depende da existência de conceitos comuns, a similaridade final não deve ser nula quando não existirem relações comuns aos dois grafos comparados, o que explica o facto dos coeficientes introduzidos.

$$sim(G_1, G_2) = sim_{conc}(G_1, G_2) \times (a + b \times sim_{rel}(G_1, G_2)) \quad \textbf{(3.13)}$$

onde sim_{conc} e sim_{rel} representam respectivamente a similaridade conceptual e relacional entre dois grafos.

A similaridade conceptual sim_{conc} (3.14) representa o número de conceitos comuns entre dois grafos e é baseada no coeficiente de Dice:

$$sim_{conc}(G_1, G_2) = \frac{2 \mid G_c \mid}{\mid G_1 \mid + \mid G_2 \mid} \quad \textbf{(3.14)}$$

que varia de 0 (quando não existem conceitos em comum entre os dois grafos) e 1 (os dois grafos possuem o mesmo conjunto de conceitos).

A similaridade relacional sim_{rel} (3.15) indica a similaridade do contexto dos conceitos apresentados em ambos os grafos. Esta componente mede a proporção existente entre o grau de ligação dos conceitos em G_c e o grau de ligação dos conceitos nos grafos G_1 e G_2

respectivamente. Deste modo, segundo os autores, uma relação (arco) entre dois conceitos (nós) possui menos informação se estes são fortemente conexos[5] do que se forem fracamente conexos.

$$sim_{rel}(G_1, G_2) = \frac{2m(G_c)}{m_{G_c}(G_1) + m_{G_c}(G_2)} \quad \textbf{(3.15)}$$

onde $m(G_c)$ é número de arcos em G_c e $m_{G_c}(G)$ é o número de arcos em G coincidentes com os arcos de G_c mais os arcos imediatamente vizinhos a estes. Em termos mais formais é a vizinhança de $G_c \subseteq G$ que consiste nos arcos de G com pelo menos um dos extremos pertencente a G_c. Um exemplo de cálculo da similaridade relacional é apresentado na Figura 3.4. Nesta Figura, as linhas a negrito representam os arcos comuns aos dois grafos. Os arcos marcados com ✓ representam os vizinhos imediatos do grafo G_c.

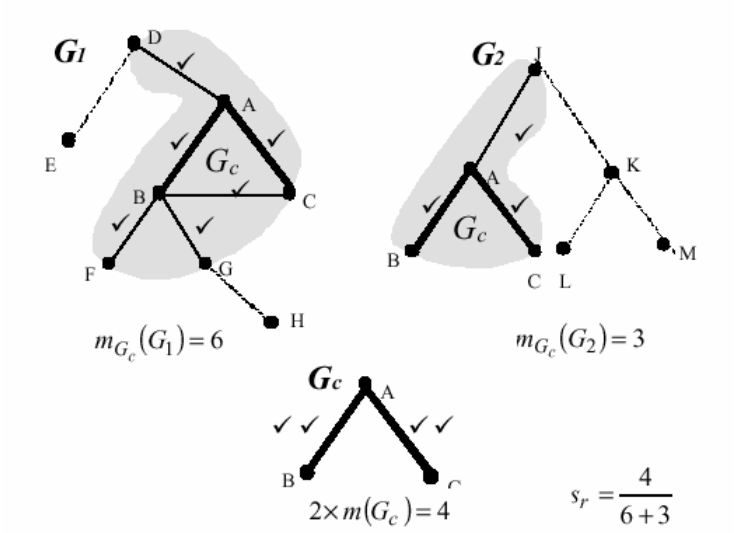


Figura 3.4: Cálculo da similaridade relacional entre os grafos G_1 e G_2.
Extraído de [Montes-y-Gómes e López-López, 2000].

Em todo o processo, é feita uma comparação directa entre conceitos e relações de dois grafos distintos, ou seja, se duas relações ou conceitos não são exactamente o mesmo, não são considerados para a comparação, o que reduz drasticamente o tempo de processamento. No nosso trabalho estamos interessados em algo mais flexível, que tenha o objectivo de estabelecer a comparação entre grafos ao nível da proximidade entre os seus conceitos (e possivelmente entre as relações que os interligam), que apesar de poderem não ser exactamente iguais, podem estar semanticamente próximos numa taxonomia utilizada como base de conhecimento. É com este princípio que abordaremos no próximo Capítulo um conjunto de alternativas que foram implementadas e testadas para comparar semanticamente mapas conceptuais.

[5] A classificação de um nó quanto ao seu grau de ligação aos outros nós de um grafo (fortemente conexo, fracamente conexo) é apresentada em anexo.

No seu trabalho, Montes-y-Gómes e López-López apenas ilustram 3 exemplos de aplicação do seu algoritmo a alguns pequenos grafos conceptuais (que em média não ultrapassam 6 nós nem 5 arcos) não sendo possível prever a sua viabilidade em grafos de maiores dimensões, tanto a nível de tempo de execução como de complexidade algorítmica. No entanto, uma vez que o algoritmo de comparação de grafos conceptuais é baseado e depende fortemente do algoritmo de detecção dos máximos subgrafos comuns entre dois grafos proposto por Myaeng e López-López em 1992, iremo-nos debruçar sobre a performance deste último. Neste caso, foi feito um estudo sobre o efeito do tamanho dos grafos comparados, tanto ao nível do número de conceitos como de relações, e tipo de comparação (de conceitos e relações, apenas de conceitos, apenas de relações ou apenas das direcções das relações entre os dois grafos) na performance final. Os autores utilizaram diferentes tamanhos de grafos com diferentes níveis de interligação (mais ou menos relações entre os conceitos). Concluíram que a comparação entre grafos que tem em conta apenas as direcções das relações (*unconstrained matching)* geram um número muito elevado de grafos de associação tendo assim um tempo de execução que aumenta exponencialmente consoante o número de conceitos e relações envolvidas. Deste modo, decidiram restringir a procura criando apenas grafos de associação entre dois grafos que possuam conceitos ou relações comuns. Esta adaptação, como consequência óbvia, diminui consideravelmente o tempo necessário de execução absoluto em relação ao primeiro tipo de comparação, mas que ainda é fortemente dependente do número de conceitos ou relações envolvidas.

Uma vez contextualizada a similaridade semântica através de trabalhos actuais, no próximo Capítulo, iremos aplicar a metodologia de Contraste de Característica aos Mapas Conceptuais utilizando uma variação quanto à determinação da equivalência entre dois conceitos de mapas conceptuais distintos. Utilizaremos, tal como alguns especialistas da área de Ciências de Educação inicialmente propunham, uma análise quantitativa para a criação de um modelo computacional de comparação de mapas conceptuais. Iremos também propor outros métodos de comparação que consideram além da semântica dos mapas conceptuais, a importância que os diferentes conceitos representam para a descrição do domínio apresentado. Deste modo, não centramos a comparação semântica num isomorfismo ou alinhamento estrutural entre mapas conceptuais, na medida em que não consideramos o significado ou tipo das relações que interligam os conceitos. Esta opção deveu-se essencialmente ao facto de querermos essencialmente uma medida sem grandes custos computacionais que permita comparar os significados de mapas conceptuais, como veremos a seguir.

4. Propostas de Similaridade Semântica entre Mapas Conceptuais: Uma abordagem experimental

Os mapas conceptuais são uma forma de representação do conhecimento num dado domínio que, ao contrário dos grafos conceptuais, não possuem uma semântica explicitamente associada, sendo assim pouco eficaz a sua aplicação computacional directa [Kremer, 1994]. Ao procurarmos o significado de um mapa conceptual como um todo, decidimo-nos, em primeiro lugar, aumentar o nível de granularidade e centrarmo-nos nos elementos fundamentais que constituem o mapa conceptual: nós (conceitos) e arcos (relações). Posteriormente, descoberto o significado de cada elemento do mapa, será possível descobrir o domínio genérico que o mapa representa como um todo. E só assim, no nosso entender, depois de contextualizar cada elemento representado num mapa conceptual, nos parece possível dizer quão similar é um mapa em relação a outro.

Para descobrir de um forma automática a proximidade semântica entre dois mapas conceptuais através de uma abordagem quantitativa, procurámos testar um conjunto de métricas que consideram não só o conteúdo apresentado pelo mapa conceptual mas também pela forma como este está organizado e interligado no domínio em questão.

De seguida, é descrito em detalhe todo o processo de contextualização e a avaliação semântica de um mapa conceptual e os recursos utilizados, assim como os estudos comparativos que foram realizados para tentar encontrar a melhor forma de medir semanticamente a distância entre duas representações sobre um mesmo ou diferentes domínios.

1. Esquema Geral

Neste trabalho, propusemo-nos criar uma abordagem para a avaliação da distância semântica entre mapas conceptuais no seguimento do trabalho desenvolvido em [Pereira, 2000] e [Alves et. al, 2001]. O nosso principal objectivo era formular uma medida de similaridade semântica que se baseasse em tão pouca informação quanto fosse possível de forma a minimizar a sua complexidade computacional. Esta informação deverá consistir, basicamente, como veremos adiante: (1) na utilização da ferramenta léxico-semântica WordNet, ao nível da sua base de conceitos e algumas das relações semânticas nela disponibilizadas; e (2) de informação estatística sobre a utilização das palavras no quotidiano (com base no corpus *SemCor*). Como compromisso de performance, procurámos restringir-nos apenas ao domínio descrito no mapa conceptual, evitando possíveis interligações inter-tópicos que poderiam eventualmente tornar o problema em causa exponencialmente complexo.

A funcionalidade principal pretendida pelo nosso trabalho visa comparar semanticamente mapas conceptuais não de uma forma absoluta, o que poderia ser equívoco e pouco consistente, como veremos adiante, mas sim tendo a relatividade como ponto fundamental. Um mapa conceptual é mais próximo semanticamente de outro em relação a um terceiro, criando-se, desta forma, uma lista dos mapas conceptuais mais similares de um mapa alvo para análise. Sendo assim, a nossa base de trabalho é um conjunto de mapas conceptuais e um mapa *target*, e o objectivo final é conseguir criar uma ordem semântica do conjunto de mapas de forma automática, com uma eventual validação do utilizador, durante uma fase de treino inicial.

Visto não ser uma tarefa fácil (nem mesmo por vezes para os humanos), a comparação semântica de mapas conceptuais acarreta uma série de dificuldades:

- Uma vez que não existe uma forma padrão de criação dos mapas conceptuais, dependendo muitas vezes da profundidade do conhecimento que o autor tem sobre o domínio em questão, um conceito ou relação pode ser descrito de variadas maneiras. Como relacionar mapas conceptuais que representem um mesmo raciocínio cognitivo mas que apenas diferem no seu grau de detalhe?
- Ao descrever um determinado domínio, o contexto é, geralmente, subentendido como sendo o mundo real. Como podemos modelar o conhecimento sobre factos do dia-a-dia de uma forma linear e genérica?
- À primeira vista, um mapa conceptual é composto por elementos que são descritos por palavras. E como a língua está constantemente sujeita a novas palavras e regras, é necessário ter uma base lexical e semântica o mais actualizada possível para um completo entendimento do conhecimento que é representado.

Por não haver actualmente nenhuma medida amplamente utilizada como padrão, ou que se aplique concretamente a mapas conceptuais, ou a qualquer representação de conhecimento estruturada, procuramos dividir a tarefa final de comparação semântica de conhecimento em alguns módulos como mostra a Figura 4.1.

Descreveremos em seguida o trajecto que os mapas conceptuais percorrem durante o processo de comparação semântica:

i. Numa primeira fase, é identificado o significado de cada elemento (apenas conceitos) que compõe um mapa conceptual recorrendo-se a recursos lexico-semânticos e factuais.

ii. Depois de contextualizados, os mapas podem ser analisados por diferentes métodos de atribuição de similaridade semântica que foram implementados e testados num estudo comparativo discutido adiante.

iii. É feita uma ordenação em relação à similaridade dos mapas conceptuais a um mapa alvo de comparação, obtendo-se assim uma medida relativa de distância semântica no universo do conjunto de mapas apresentados.

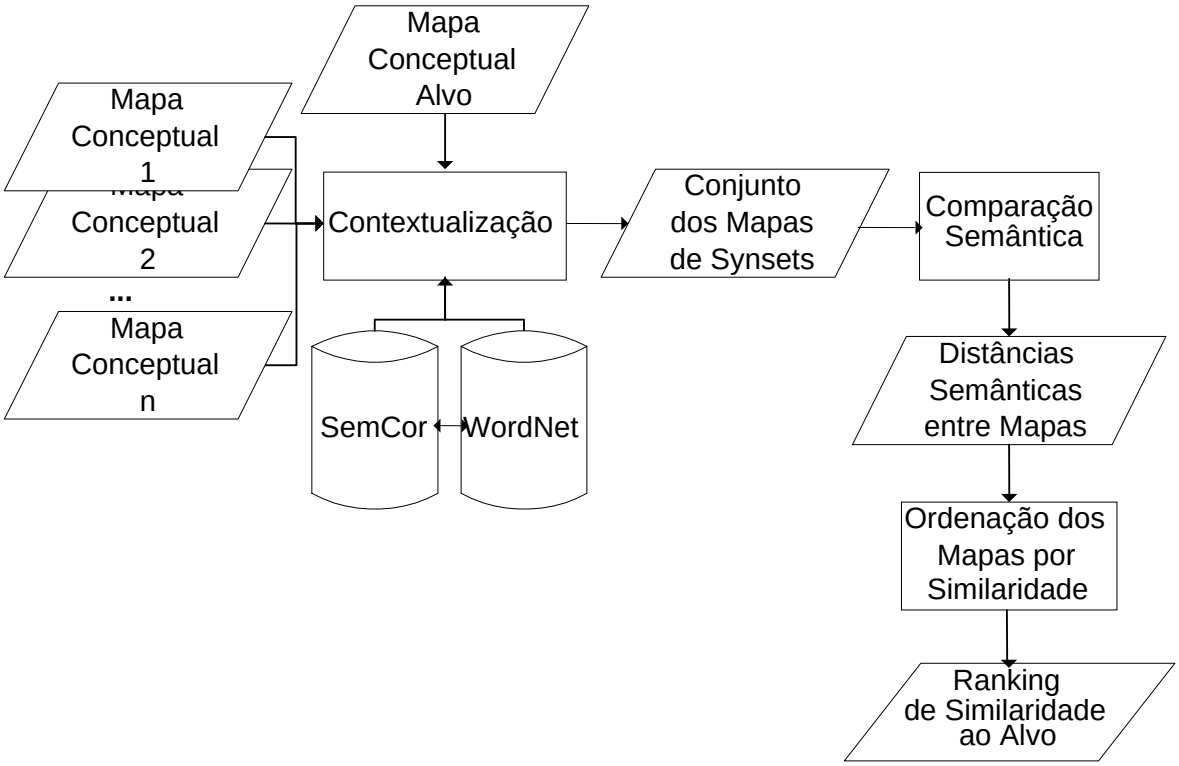


Figura 4.1: Esquema geral para a comparação semântica de mapas conceptuais

Os módulos da Figura 4.1 serão, de seguida, apresentados mais detalhadamente, bem como as respostas às dificuldades sublinhadas anteriormente.

2. Contextualização: Semântica de um Mapa Conceptual

Por ser uma linguagem visual informal, ao construirmos um mapa conceptual não há qualquer referência sobre o significado de cada elemento ou mesmo do mapa em si, deixando, para o leitor do mapa, a tarefa cognitiva de associar uma semântica própria através do contexto representado pela interligação de conceitos por relações.

Nesta Sub Capítulo, propomo-nos reproduzir este processo cognitivo aplicando uma abordagem de *Word Sense Disambiguation* adaptada aos mapas conceptuais. Cada conceito (ou relação) é associado a um significado numa taxonomia de base, no nosso caso o WordNet, para posterior processamento computacional do mapa. Esta associação não é unívoca, na medida em que uma palavra pode descrever diferentes significados. É necessário tentar encontrar o

significado real representado pelo elemento em foco tendo em atenção todos os outros que lhe são adjacentes no mapa.

Uma palavra, quando vista isoladamente, pode descrever diversos significados dependendo da interpretação do leitor. Só é possível determinar o real significado expresso por uma palavra, numa utilização em particular, se examinarmos o seu contexto envolvente.

Sendo os mapas conceptuais constituídos por conceitos e relações descritos por palavras, será que, se conseguirmos descobrir o significado de cada palavra presente num mapa conceptual, de uma forma automática, estaremos mais próximos da sua interpretação e avaliação por um sistema computacional? À esta pergunta iremos responder com alguns resultados.

Para descobrir o significado de cada conceito de um mapa conceptual utilizaremos técnicas de Desambiguação dos Sentidos de Palavras adaptadas a esta nova representação de conhecimento, o mapa conceptual, ao invés de texto livre.

Actualmente, algumas aplicações de processamento da língua natural são disponibilizadas publicamente prontas a serem utilizadas, como é o caso, por exemplo, de algumas que fazem classificação gramatical (*POS taggers*). O mesmo não acontece, ainda, com a desambiguação do significado de palavras. Isto significa que, mesmo para testar diferentes "desambiguadores" é necessário implementar de raiz o seu algoritmo, o que não é claramente o objectivo principal do nosso trabalho. Procurámos então, com base no que está publicado, escolher e adaptar, de entre as abordagens estudadas[1], um método que atendesse a três factores: precisão razoável, disponibilidade e aplicabilidade em larga escala. Apesar da precisão de um método de desambiguação demonstrar o seu sucesso, esta medida não pode ser considerada de forma absoluta, já que diversos trabalhos são testados em pequenos conjuntos palavras, e não em grandes conjuntos de textos (aplicabilidade). Chegámos, assim, à conclusão de que o melhor seria utilizar uma abordagem baseada num recurso lexical disponível, o WordNet, pelas seguintes razões:

- Esta subárea do Processamento da Língua Natural ainda se encontra numa fase de forte investigação, sem nenhum método estabelecido como o melhor ou mais eficiente;
- Não existe actualmente um *corpus* de mapas conceptuais, sendo impossível obter quaisquer dados estatísticos sobre a utilização de palavras neste tipo de representação;
- A taxonomia escolhida, o WordNet, tem sido amplamente utilizada na comunidade científica para manipulação de língua natural e processamento semântico pela sua disponibilidade,

[1] Ver discussão do tema na Sub Secção 2.2.2 (Desambiguação dos Significadso de Palavras).

constante actualização e organização; trata-se de um recurso já considerado padrão, apesar de ser desejável que algumas limitações que apresenta venham a ser ultrapassadas;

- A utilização de restrição seleccionadora, um método simbólico, não seria aplicável, já que a generalidade pretendida de uma métrica conceptual para comparar quaisquer domínios tornaria a enumeração das diversas regras de restrição um trabalho laborioso;
- Por fim, a aplicação de uma abordagem que utiliza já uma base de conhecimento comum a outras trabalhos da área permite comparar a performance e precisão conseguidas.

Entre os trabalhos estudados, o realizado por Agirre e Rigau [1996] serviu de inspiração para o módulo de contextualização, devido à base de conhecimento escolhida (o WordNet), simplicidade do método empregue (que pode ser facilmente adaptada aos mapas conceptuais) e precisão considerável (acima dos 80%). Este método consiste em comparar cada significado (na presença de vários) de uma palavra presente no texto com todos os significados que se encontram a sua volta. Para cada palavra do texto a desambiguar, os autores delimitam uma janela de comprimento constante centrada nesta mesma palavra. A comparação feita entre os significados dos pares de palavras presentes na janela é feita através de uma medida de similaridade[2] baseada na distância semântica[3] que consiste na contabilização dos arcos da taxonomia hierárquica (aplicável apenas aos substantivos) do WordNet. O significado a escolher para a palavra central da janela textual será o que obtiver maior proximidade ao conjunto de palavras envolventes. Adaptámos este método a fim de desambiguar mapas conceptuais da seguinte forma:

- a janela textual originalmente proposta passou a ser o subgrafo imediatamente à volta do conceito a desambiguar;
- tal como no método original, iremo-nos debruçar apenas sobre os substantivos (conceitos ou nós dos mapas conceptuais). Os verbos (relações ou arcos) não serão considerados, já que observamos poucos métodos aplicáveis a esta classe gramatical devido à sua alta mutabilidade, ou seja, os significados dos verbos dependem fortemente das palavras (nesta caso os substantivos) que os acompanham[4];
- Ao contrário da medida de similaridade proposta inicialmente por Agirre e Rigau [1996], decidimos investir numa medida baseada no conteúdo da informação de cada conceito explorado.

[2] A similaridade entre conceitos aplicada neste trabalho será amplamente discutida na próxima Secção (4.3).
[3] Ver as abordagens existents sobre a similaridade descritas no Sub Capítulo 3.1 (Similaridade entre Conceitos).
[4] Tal como foi defendido na Sub Secção 2.2.1.

No nosso ponto de vista, a desambiguação de um mapa conceptual consistirá basicamente em seleccionar o significado mais apropriado a cada palavra presente no mapa, entre os diversos possíveis que existam no WordNet. A desambiguação das palavras que representam os conceitos (nós) do mapa conceptual será efectuada de acordo com o algoritmo apresentado no Quadro 4.1.

```
Contextualização(CM)
    CM_desambiguado ← ∅
    para cada c ∈ CM
    fazer
(1)   S ← SignificadosWordNet(c)
      definir Score[|S|]
      i ← 0
      enquanto i < |S|
      fazer
            Score[i] ← 0
            i ← i + 1
      fim enquanto
(2)   A ← ConceitosAdjacentes(c, CM, CM_desambiguado)
      i ← 0
      para cada s ∈ S
      fazer
            para cada a ∈ A
            fazer
(3)               Score[i] ← Score[i] + sim(s,a)
            fim para
            i ← i + 1
      fim para
(4)   conceitofinal ← MaiorScore(Score[], S)
      adiciona conceitofinal CM_desambiguado
    fim para
    devolve CM_desambiguado
```

Quadro 4.1: Contextualização de um mapa conceptual. As linhas numeradas serão detalhadamente explicadas através da aplicação a um exemplo prático.

Para uma melhor ilustração de todo o processo, acompanharemos a descrição das principais etapas do algoritmo (ver linhas numeradas do Quadro 4.1) através da utilização de um exemplo simples de um mapa conceptual apresentado na Figura 4.2.

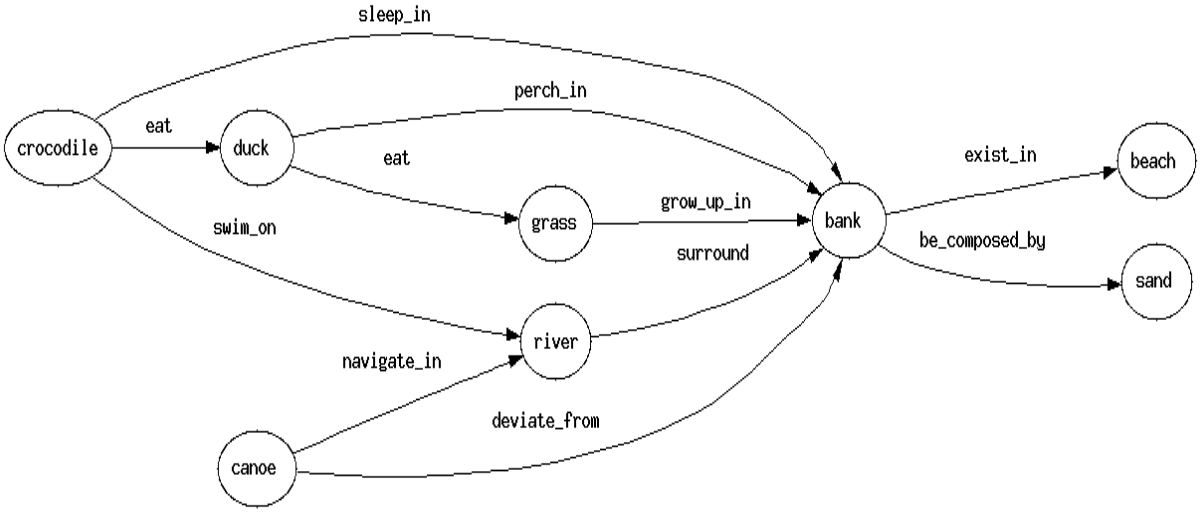
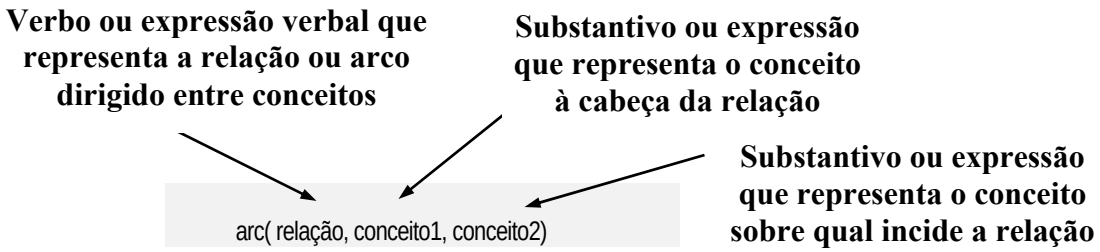


Figura 4.2: Exemplo simples de uma mapa conceptual.

Etapa 1. Dado um Mapa Conceptual CM, o objectivo do algoritmo de Contextualização é construir um Mapa Conceptual Desambiguado, $CM_{desambiguado}$, que represente os reais significados dos conceitos presentes no primeiro. Neste ponto, é feita a identificação dos significados possíveis de um conceito (ou relação) no WordNet. A representação interna de cada mapa conceptual é feita através de predicados ternários associados a cada interligação (arco dirigido) presente no grafo, através da seguinte sintaxe:



No nosso exemplo da Figura 4.2, o mapa conceptual possuirá a seguinte codificação:

```
arc(exist_in, bank, beach).
arc(eat, duck, grass).
arc(grow_up_in, grass, bank).
arc(be_composed_by, bank, sand).
arc(deviate_from, canoe, bank).
arc(navigate_in, canoe, river).
arc(surround, river, bank).
arc(swim_on, crocodile, river).
arc(eat, crocodile, duck).
arc(sleep_in, crocodile, bank).
arc(perch_in, duck, bank).
```

Para cada conceito (ou relação), representado neste caso por uma palavra ou expressão, é extraída do WordNet uma lista dos seus significados (*synsets*) possíveis na taxonomia como podemos ver na Tabela 4.1.

	Significados	
bank	1	a financial institution that accepts deposits and channels the money into lending activities.
	2	sloping land especially the slope beside a body of water.
	3	a supply or stock held in reserve for future use especially in emergencies.
	4	a building in which commercial banking is transacted.
	5	an arrangement of similar objects in a row or in tiers.
	6	a container usually with a slot in the top for keeping money at home.
	7	a long ridge or pile.
	8	the funds held by a gambling house or the dealer in some gambling games.
	9	a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force.
	10	a flight maneuver; aircraft tips laterally about its longitudinal axis especially in turning.
beach	1	an area of sand sloping down to the water of a sea or lake.
canoe	1	small and light boat; pointed at both ends; propelled with a paddle.
crocodile	1	large voracious aquatic reptile having a long snout with massive jaws and a body covered with bony plates; of sluggish tropical waters.
duck	1	small wild or domesticated web-footed broad-billed swimming bird usually having a depressed body and short legs.
	2	in cricket a score of nothing by a batsman.
	3	flesh of a duck domestic or wild.
	4	a heavy cotton fabric of plain weave; used for clothing and tents.
grass	1	narrow-leaved green herbage: grown as lawns; used as pasture for grazing animals; cut and dried as hay.
	2	a strong-smelling plant from whose dried leaves a number of euphoriant and hallucinogenic drugs are prepared.
	3	German writer of novels and poetry and plays born 1927.
	4	animal food for browsing or grazing.
	5	a soft drug consisting of the dried leaves of the hemp plant; smoked or chewed for euphoric effect.
river	1	a large natural stream of water larger than a creek.
sand	1	a loose material consisting of grains of rock or coral
	2	French writer known for works concerning women's rights and independence 1804-1876.
	3	informal fortitude and determination

Tabela 4.1: Significados de cada palavra expressas no nós do mapa conceptual da Figura 4.2.

Os conceitos que possuam apenas um significado possível no WordNet são considerados já desambiguados.

Etapa 2. Dado um conceito, identificar o contexto presente no mapa conceptual à sua volta através da função. Para um dado conceito, um contexto será entendido como todos os

significados dos conceitos que lhe são adjacentes, ou seja, que possua uma interligação (convergente ou divergente) ao conceito em questão, no grafo que é constituído pelo mapa conceptual. No nosso exemplo teremos os seguintes contextos[5] para os seguintes nós:

Conceito	Contexto
bank	beach canoe crocodile duck#1 duck#2 duck#3 duck#4 grass#1 grass#2 grass#3 grass#4 grass#5 river sand#1 sand#2 sand#3
beach	bank#1 bank#2 bank#3 bank#4 bank#5 bank#6 bank#/ bank#8 bank#9 bank#10
canoe	bank#1 bank#2 bank#3 bank#4 bank#5 bank#6 bank#/ bank#8 bank#9 bank#10 river
crocodile	bank#1 bank#2 bank#3 bank#4 bank#5 bank#6 bank#/ bank#8 bank#9 bank#10 duck#1 duck#2 duck#3 duck#4 river
duck	bank#1 bank#2 bank#3 bank#4 bank#5 bank#6 bank#/ bank#8 bank#9 bank#10 crocodile grass#1 grass#2 grass#3 grass#4 grass#5
grass	bank#1 bank#2 bank#3 bank#4 bank#5 bank#6 bank#/ bank#8 bank#9 bank#10 duck#1 duck#2 duck#3 duck#4
river	bank#1 bank#2 bank#3 bank#4 bank#5 bank#6 bank#/ bank#8 bank#9 bank#10 canoe crocodile
sand	bank#1 bank#2 bank#3 bank#4 bank#5 bank#6 bank#/ bank#8 bank#9 bank#10

Caso um conceito adjacente já tenha sido desambiguado é considerado apenas o seu significado final escolhido para fazer parte do contexto.

Etapa 3. Calcular a similaridade entre os seus significados e os dos outros conceitos presentes no contexto (uma vez que estes também são ambíguos): A similaridade entre dois significados (*synsets* do WordNet) **x** e **y** será calculada com base na fórmula 3.5 de distância semântica proposta por Jiang e Conrath[6] [1997] apresentada no Sub Capítulo 3.1:

$$dist(x, y) = CI(x) + CI(y) - 2 * CI(gcme_{x,y}) \quad \textbf{(3.5)}$$

sendo $gcme_{x,y}$ o conceito mais específico na taxonomia hierárquica de substantivos do WordNet que generalize ambos os conceitos **x** e **y**, e *CI(x)* o conteúdo de informação definido pela fórmula 3.4:

$$CI(c) = -\log P(c) \quad \textbf{(3.4)}$$

onde *P(c)* é a probabilidade de utilização do conceito **c**. No seu trabalho, Jiang e Conrath não especificaram explicitamente um valor máximo para a sua medida de distância semântica entre dois conceitos, apenas o valor mínimo 0 de distância, representando o caso de conceito de significados idênticos, ou seja, a distância mínima possível. Ao analisar a fórmula 3.5,

[5] Onde palavra#n indica o significado n da palavra indicada caso a palavra possua mais do que um significado possível no WordNet.
[6] O porquê desta escolha será fundamentado no próximo Sub Capítulo (Comparação Semântica) e Capítulo 5 (Testes Efectuados).

verificamos que a distância máxima obtida depende claramente da abrangência do *corpus*, neste caso o *SemCor*, e as probabilidades dos conceitos nele calculadas.

Como o nosso objectivo era obter uma medida de similaridade entre dois significados no intervalo [0,1], onde o valor 0 indicaria significados completamente dissimilares e o valor 1 significados idênticos, utilizamos o valor inverso da distância semântica da fórmula (3.5) originalmente proposta:

$$sim(x, y) = 1 - \frac{dist(x, y)}{\max(dist(w, z))} \quad \textbf{(4.1)}$$

onde *max(dist(w,z))* representa a maior distância possível entre dois conceitos presentes no WordNet. Este limite será calculado e discutido em maior detalhe no próximo Sub Capítulo. Sendo assim, foi calculada a similaridade entre cada par de significados das palavras num mesmo contexto para descobrir o significado de cada nó do mapa como pode ser observado na Tabela 4.2 (um exemplo de desambiguação da palavra *bank*). O processo é iniciado no nó do mapa conceptual de maior grau (grau de entrada somado ao grau de saída) de ligação, ou seja, aquele que possua um maior contexto à sua volta.

Conceito a desambiguar	Contexto																Total por Significado
	beach	canoe	crocodile	duck#1	duck#2	duck#3	duck#4	grass#1	grass#2	grass#3	grass#4	grass#5	river	sand#1	sand#2	sand#3	
bank#1	0,33	0,28	0,24	0,29	0,24	0,24	0,24	0,38	0,24	0,24	0,26	0,28	0,41	0,32	0,24	0,29	4,52
bank#2	**0,76**	**0,29**	**0,25**	**0,31**	**0,16**	**0,21**	**0,25**	**0,40**	**0,25**	**0,21**	**0,24**	**0,25**	**0,38**	**0,30**	**0,21**	**0,21**	**4,68**
bank#3	0,25	0,19	0,15	0,21	0,15	0,15	0,15	0,30	0,16	0,16	0,18	0,19	0,32	0,14	0,15	0,20	3,05
bank#4	0,32	0,36	0,23	0,28	0,14	0,19	0,33	0,38	0,23	0,19	0,21	0,23	0,36	0,27	0,19	0,19	4,10
bank#5	0,17	0,11	0,07	0,13	0,07	0,07	0,07	0,22	0,07	0,08	0,10	0,11	0,24	0,16	0,07	0,12	1,86
bank#6	0,25	0,38	0,16	0,21	0,07	0,12	0,25	0,30	0,16	0,12	0,14	0,16	0,19	0,20	0,12	0,12	2,95
bank#7	0,69	0,22	0,18	0,14	0,09	0,14	0,18	0,23	0,18	0,15	0,17	0,18	0,31	0,23	0,14	0,14	3,37
bank#8	0,12	0,06	0,02	0,08	0,02	0,02	0,02	0,17	0,02	0,02	0,05	0,06	0,19	0,11	0,02	0,07	1,05
bank#9	0,62	0,15	0,11	0,17	0,02	0,07	0,11	0,26	0,11	0,07	0,10	0,11	0,24	0,16	0,07	0,07	2,44
bank#10	0,09	0,04	0,00	0,06	0,00	0,00	0,00	0,15	0,00	0,00	0,02	0,04	0,17	0,08	0,00	0,05	0,70

Tabela 4.2: Cálculo de similaridade para cada par de significados das palavras do contexto à volta da palavra *bank* no mapa conceptual da Figura 4.2.

Etapa 4. Escolher o significado com melhor pontuação, ou seja, maior proximidade com todos os significados dos outros conceitos do contexto. Uma vez descoberto o significado de um conceito, este será considerado nos próximos cálculos de similaridade com os significados dos outros conceitos que ainda permanecem ambíguos. No nosso exemplo, o conceito representado

pela palavra *bank* foi o primeiro a ser contextualizado com o seguinte significado: "*sloping land especially the slope beside a body of water*".

Continuando o processo de desambiguação, na Tabela 4.3 é apresentada a identificação dos significados dos restantes conceitos ambíguos pela seguinte ordem crescente de grau de ligação: *duck*, *grass* e *sand*.

Conceito a desambi-guar	Contexto																	Total por Significado
	bank#2	beach	canoe	crocodile	duck#1	duck#2	duck#3	duck#4	grass#1	grass#2	grass#3	grass#4	grass#5	river	sand#1	sand#2	sand#3	
duck#1	**0,31**	X	X	**0,46**					**0,40**	**0,25**	**0,25**	**0,13**	**0,14**	X	X	X	X	**1,94**
duck#2	0,16	X	X	0,00					0,15	0,00	0,00	0,02	0,04	X	X	X	X	0,37
duck#3	0,21	X	X	0,05					0,20	0,05	0,05	0,29	0,09	X	X	X	X	0,94
duck#4	0,25	X	X	0,09					0,14	0,09	0,05	0,07	0,09	X	X	X	X	0,78
grass#1	**0,40**	X	X	X	**0,40**	—	—	—						X	X	X	X	**0,80**
grass#le2	0,25	X	X	X	0,25	—	—	—						X	X	X	X	0,50
grass#3	0,21	X	X	X	0,25	—	—	—						X	X	X	X	0,46
grass#4	0,24	X	X	X	0,13	—	—	—						X	X	X	X	0,37
grass#5	0,25	X	X	X	0,14	—	—	—						X	X	X	X	0,39
sand#1	**0,30**	X	X	X	X	X	X	X	X	X	X	X	X	X				**0,30**
sand#2	0,21	X	X	X	X	X	X	X	X	X	X	X	X	X				0,21
sand#3	0,21	X	X	X	X	X	X	X	X	X	X	X	X	X				0,21

Tabela 4.3: Cálculo de similaridade para cada par de significados das palavras do contexto à volta das palavras *duck*, *grass* e *sand* no mapa conceptual da Figura 4.2. O símbolo X representa os significados que não serão considerados para o cálculo por não aparecerem no contexto da palavra em causa. O símbolo — representa os significados que já não serão considerados por já não indicarem o significado correcto do conceito que representam.

No final do processo de contextualização, obtivemos o resultado apresentado na Tabela 4.4 que contém os significados escolhidos para cada conceito presente no mapa conceptual da Figura 4.2.

Conceito	Significado escolhido
bank#2	sloping land especially the slope beside a body of water
beach	an area of sand sloping down to the water of a sea or lake
canoe	small and light boat; pointed at both ends; propelled with a paddle
crocodile	large voracious aquatic reptile having a long snout with massive jaws and a body covered with bony plates; of sluggish tropical waters
duck#1	small wild or domesticated web-footed broad-billed swimming bird usually having a depressed body and short legs
grass#1	narrow-leaved green herbage: grown as lawns; used as pasture for grazing animals; cut and dried as hay
river	a large natural stream of water larger than a creek
sand#1	a loose material consisting of grains of rock or coral

Tabela 4.4: Resultado do processo de contextualização (ou desambiguação) das palavras do mapa conceptual da Figura 4.2.

3. Comparação Semântica: Similaridade entre Mapas Conceptuais

Nesta fase, uma vez que os mapas conceptuais já se encontram contextualizados na base léxico-semântica WordNet, é feita uma comparação quanto à similaridade semântica entre um mapa-alvo e um conjunto de mapas a comparar. Esta comparação será sobretudo relativa ao conjunto de mapas que temos em cada momento, uma vez que o conceito de similaridade não é constante, na medida em que o valor de comparação entre dois mapas conceptuais pode variar dependendo do contexto apresentado (os outros mapas que fazem parte do conjunto).

Como comparar mapas conceptuais de forma global? Correndo o risco da não-composicionalidade (o que se aplica às partes pode não ser generalizado ao todo), apresentamos a comparação global e semântica de mapas conceptuais como um estudo quantitativo dos conceitos que os constituem. Este compromisso foi assumido, principalmente, devido à necessidade de evitar uma complexidade da análise estrutural (uma vez que o problema da detecção de subgrafos ainda é conhecido como NP-completo) de modo a facilitar uma possível automatização do processo de comparação semântica por um modelo computacional.

Mesmo uma medida relativa de similaridade que permita ordenar mapas conceptuais em relação a um mapa-alvo, necessita de valores numéricos, mesmo que aplicáveis somente ao contexto apresentado em cada momento, para criar uma escala de comparação. Face a esta realidade, como estabelecer um valor que quantifique a similaridade semântica entre duas estruturas conceptuais? Debruçámo-nos então sobre a possibilidade de utilizar o modelo de similaridade entre conceitos (elemento constituinte dos mapas conceptuais) para a comparação dos próprios mapas conceptuais. De seguida, na Secção 4.3.1, é referido que medidas foram

consideradas para o estudo da similaridade semântica entre conceitos, para depois, na Secção 4.3.2, visualizarmos o processo a nível mais lato alargado aos mapas conceptuais

3.1. Similaridade entre Conceitos

Depois de estudadas as diferentes abordagens e técnicas que existem actualmente para o cálculo da similaridade entre conceitos no Sub Capítulo 3.1, surge-nos o seguinte pensamento: alguma destas medidas já responderá satisfatoriamente ao problema da comparação semântica entre conceitos? E em caso negativo, que melhoramentos ou novas abordagens podem ser propostas? Deste modo, no âmbito do problema da desambiguação de significados, testámos[7] algumas destas medidas e a sua aplicabilidade aos mapas conceptuais, no que se refere à comparação entre dois nós distintos. De seguida, veremos em detalhe as medidas utilizadas e problemas encontrados na sua aplicação.

Numa primeira abordagem, a distância semântica (ou o seu inverso, a similaridade) entre conceitos (*synsets*) do WordNet era simplesmente a soma do número de arcos que constituíam o caminho mais curto possível entre dois conceitos na taxonomia do WordNet (relações *is-a*) tal como seguido em [Rada et. al, 1989]. Utilizámos, para tal, a definição de conceito **generalizador comum mais específico** ou **gcme**[8], ou seja, aquele conceito mais geral que está ligado a ambos os conceitos, **a** e **b,** através de relações hierárquicas (*is-a*) e se encontra a um nível mais específico (já que o seu **hiperónimo** também obedece à primeira regra, generalizando ambos os conceitos). Deste modo, o comprimento do caminho entre dois conceitos, **a** e **b**, será o número de arcos (relações *is-a*) entre o conceito **a** e o seu generalizador comum mais específico (**gcme**), somado aos que estão entre o **gcme** e o conceito **b.**

Posteriormente, visto que os valores apresentados eram absolutos e necessitávamos de uma similaridade relativa, utilizámos uma variação desta medida proposta inicialmente por Leacock e Chodorow (3.1) que estabelece uma razão entre o comprimento (em número de synsets) do caminho entre dois conceitos, e o comprimento do maior caminho possível entre dois conceitos na árvore taxonómica de substantivos. Na versão 1.7.1 do WordNet esta profundidade máxima é de 18 *synsets*, que como exemplo, é o comprimento do caminho do synset mais específico {*rock_hind*, *epinephelus_adscensionis*} até ao mais geral {*entity*, *physical_thing*}. Todos os substantivos mais gerais foram ligados através da criação de um conceito abstracto (*root*) que os generaliza a todos, aumentando, deste modo, a profundidade máxima da árvore para 19 *synsets*.

[7] Estes testes serão apresentados no Capítulo 5 (Testes Efectuados).
[8] Tal como definido no Sub Capítulo 3.1 (Similaridade entre Conceitos).

$$sim(x,y) = -\log \frac{len(x,y)}{2*19} \quad \textbf{(3.1)}$$

O intervalo de valores de similaridade obtidos pela fórmula (3.1) está compreendido no intervalo [0,1.58], já que o valor máximo de similaridade é obtido entre conceitos iguais:

$$sim(x,x) = -\log \frac{len(x,x)}{2*19} = -\log \frac{1}{2*19} \cong 1.58$$

Utilizámos o valor 1.58 como factor normalizador final.

Como exemplo de aplicação prática das medidas de similaridade à nossa realidade, utilizaremos o seguinte subconjunto de conceitos referido no módulo de contextualização (Sub Capítulo 4.2):

Conceito	Significado no WordNet
bank#1	a financial institution that accepts deposits and channels the money into lending activities
bank#2	sloping land especially the slope beside a body of water
bank#6	a container usually with a slot in the top for keeping money at home
canoe	small and light boat; pointed at both ends; propelled with a paddle
crocodile	large voracious aquatic reptile having a long snout with massive jaws and a body covered with bony plates; of sluggish tropical waters.
duck#1	small wild or domesticated web-footed broad-billed swimming bird usually having a depressed body and short legs
duck#4	a heavy cotton fabric of plain weave; used for clothing and tents
sand#1	a loose material consisting of grains of rock or coral

A localização hierárquica relativa a cada um destes significados na taxonomia de conceitos do WordNet é ilustrada na Figura 4.3.

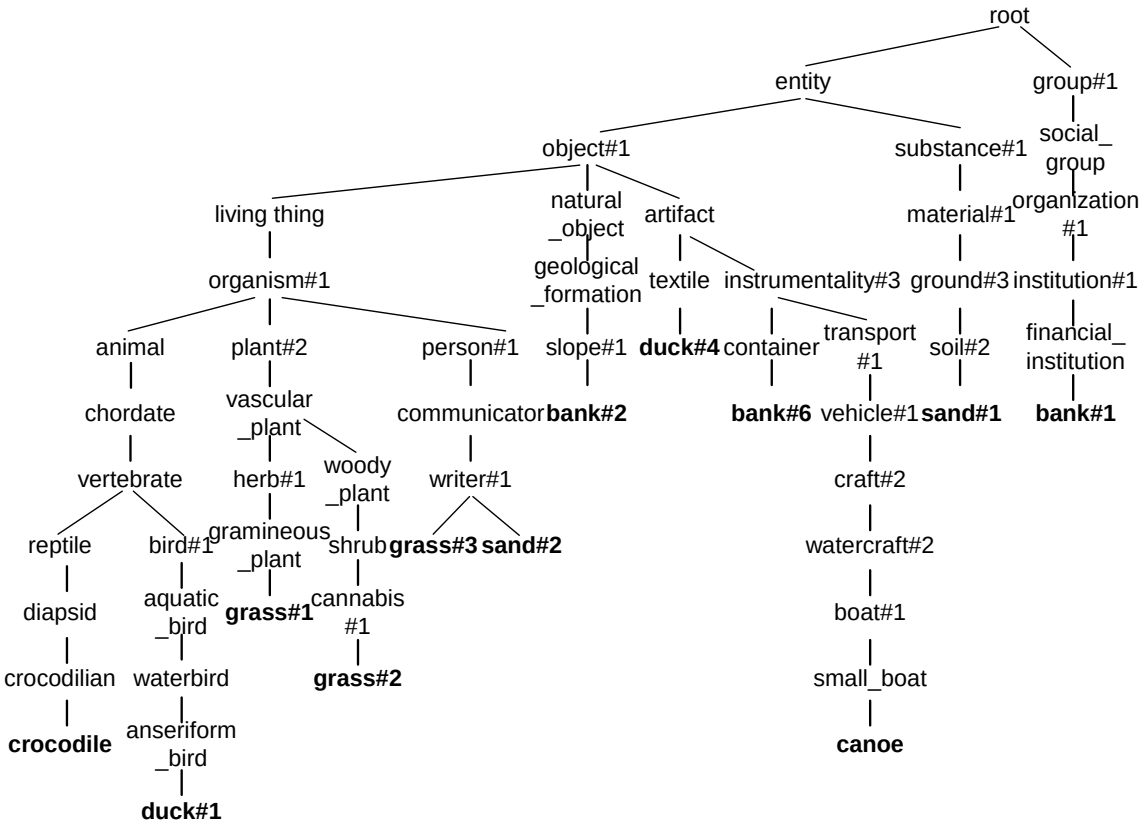


Figura 4.3: Extracto da taxonomia hierárquica (relações is*a*) do WordNet 1.7.1. Onde palavra#n significa o enésimo significado da palavra se possuir mais do que uma definição no WordNet.

Ao compararmos alguns destes conceitos, de acordo com a fórmula de Leacock e Chodorow (3.1) normalizada, obtemos os valores de similaridade apresentados na Tabela 4.5.

Um problema constatado neste simples método foi o facto de considerar que as ligações hierárquicas (*is-a*) na taxonomia do WordNet representam uma medida uniforme de distância. Isto não parece representar a realidade, pois existem categorias muito mais detalhadas (com várias subcategorias) do que outras. Por exemplo, os conceitos ***duck#1*** *is-a* ***anseriform_bird*** parecem ser muito mais próximos do que ***bank#6*** *is-a* ***container***, e no entanto, ambos os pares possuem igualmente uma relação (arco) *isa* a ligá-los.

Outro problema é o facto de considerarmos que a distância semântica entre dois conceitos mais genéricos seja equivalente à distância semântica de dois conceitos mais específicos não tendo em conta a sua profundidade (ou, especificidade) na taxonomia. Por exemplo, os conceitos ***reptile*** e ***bird#1*** parecem ser semanticamente muito mais próximos do que os conceitos ***substance*** e ***object***, e no entanto, a similaridade nos dois casos é a mesma se considerarmos esta medida.

Uma vez que, através desta medida, estamos apenas a considerar as relações hierárquicas do WordNet (taxonomia), a similaridade entre ***sand#1*** e ***bank#2*** tem o mesmo valor da

similaridade entre ***sand#1*** e ***bank#6***, apesar de termos consciência de que, na realidade, o primeiro par é mais utilizado em conjunto. Isto acontece, como podemos observar na Figura 4.3, porque a classificação (*isa*) do conceito "areia" não se intersecta com a do conceito "banco de areia". Em vez disso, existe um outro tipo de relação semântica presente no WordNet que poderia aproximar estes dois conceitos (*is_part_of/has_part*) como veremos a seguir.

conceitos		len(x,y)	$similaridade_{Leacock\&Chodorow}$
canoe	bank#1	13	-log(13/38)/1,58 ≅ **0,29**
	bank#2	14	-log(14/38)/1,58 ≅ **0,27**
	bank#6	10	-log(10/38)/1,58 ≅ **0,37**
crocodile	duck#1	10	-log(10/38)/1,58 ≅ **0,37**
	duck#4	13	-log(13/38)/1,58 ≅ **0,29**
duck#1	bank#1	19	-log(19/38)/1,58 ≅ **0,19**
	bank#2	15	-log(15/38)/1,58 ≅ **0,26**
	bank#6	15	-log(15/38)/1,58 ≅ **0,26**
duck#4	bank#1	12	-log(15/38)/1,58 ≅ **0,32**
	bank#2	8	-log(8/38)/1,58 ≅ **0,43**
	bank#6	6	-log(6/38)/1,58 ≅ **0,51**
sand#1	bank#1	13	-log(13/38)/1,58 ≅ **0,29**
	bank#2	11	-log(11/38)/1,58 ≅ **0,34**
	bank#6	11	-log(11/38)/1,58 ≅ **0,34**

Tabela 4.5: Cálculo de similaridade entre alguns pares de significados (**x** e **y**) das palavras presentes no mapa conceptual da Figura 4.2.

Actualmente a maioria das medidas semânticas propostas com base no WordNet são aplicáveis somente a substantivos, devido à sua forte organização hierárquica. Por sua vez, Hirst e St. Onge[1998] apresentam uma medida em que consideram um grande número de relações possíveis entre palavras que podem pertencer a qualquer uma das quatro classes gramaticais (adjectivo, advérbio, substantivo ou verbo).

Tal como apresentado na Secção sobre Abordagem de Similaridade Baseada na Distância Semântica (3.1.1), existem 3 tipos de relações (exclusivas) entre duas palavras: **extra-forte (**que os autores pontuam com uma cotação de 24 valores); **forte** (16 valores) e **média-forte** que obedece a seguinte fórmula:

$$sim(x,y) = 8 - path(x,y) - changes\_of\_direction \qquad \textbf{(3.2)}$$

onde *path(x,y)* é o número de relações (ver Quadro 3.1) existentes no caminho de x a y, e finalmente, *changes_of_direction* é o número de mudanças de direcção presentes neste mesmo caminho. Como o valor de maior similaridade atribuível por esta medida é 24, este será o factor normalizador para obtermos uma medida no intervalo pretendido [0,1].

Aplicando a medida de similaridade de Hirst e St-Onge às combinações entre conceitos da Tabela 4.5, devido à grande dimensão da árvore de pesquisa ao longo de diferentes relações (entre substantivos: antónimos, atributos, hiperónimos, hipónimos, holónimos e merónimos) à volta dos conceitos (ver Figura 4.4 e Figura 4.5), decidimos analisar mais detalhadamente apenas a similaridade entre três pares de conceitos: ***sand#1*** e ***bank#2***; ***sand#1*** e ***bank#6***; ***crocodile#1*** e ***duck#1***.

Intuitivamente, a similaridade semântica entre o primeiro par (***sand#1*** e ***bank#2***) parece-nos maior que no segundo caso (***sand#1*** e ***bank#6***). Este facto é comprovado por Hirst e St-Onge considerarem, ao contrário de Leacock e Chodorow, outras relações semânticas do WordNet além das hierárquicas entre substantivos (e outras classes gramaticais). Deste modo, enquanto que a similaridade entre os conceitos "areia" e "cofre" é considerada nula por não existir um caminho válido até 5 relações (como estipulado pelos autores), o mesmo não se verifica no cálculo da similaridade entre "areia" e "banco de areia":

$$sim(sand\#1, bank\#2) = \frac{8 - path\_length - changes\_of\_direction}{24} = \frac{8 - 4 - 1}{24} = 0{,}125$$

Já no cálculo da similaridade entre os conceitos "crocodilo" e "pato", como podemos observar na Figura 4.5, não é possível estabelecer nenhum caminho válido ao longo das várias relações semânticas do WordNet, resultando numa similaridade nula entre estes dois conceitos de acordo com os critérios estabelecidos pelos autores. Esta inflexibilidade é um dos principais factores que prejudicam a aplicação em larga escala desta medida. Para se ter uma ideia mais geral da sua rigidez, esta medida apenas produziu dois valores de similaridade não-nulos para os pares de conceitos presentes na Tabela 4.5: entre ***duck#4*** e ***bank#6*** (≅ 0,083) e entre ***sand#1*** e ***bank#2***.

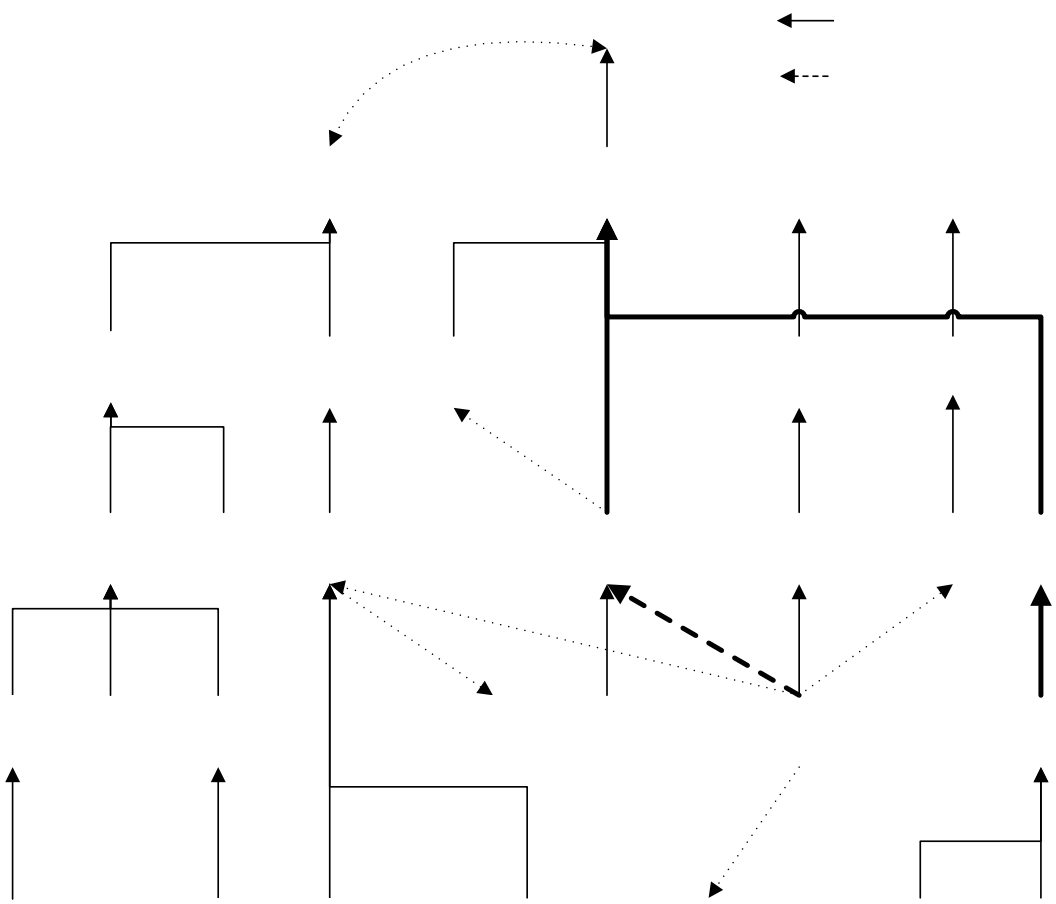


Figura 4.4: Extracto da árvore de pesquisa (WordNet 1.7.1) para estabelecer a medida proposta em 3.2 entre os conceitos ***sand#1*** e ***bank#2***, e entre ***sand#1*** e ***bank#6***.

A principal particularidade da medida de Hirst e St-Onge é a contabilização de diferentes relações semânticas do WordNet que, em alguns casos, como vimos no exemplo, provou ser mais eficaz que apenas a comparação taxonómica (relações *isa* que formam a árvore hierárquica dos substantivos). Esta flexibilidade é a sua mais-valia, enriquecendo a análise de similaridade semântica, mas que por outro lado, aumenta a complexidade e tempo gasto na pesquisa dos caminhos possíveis entre conceitos.

instrumentality

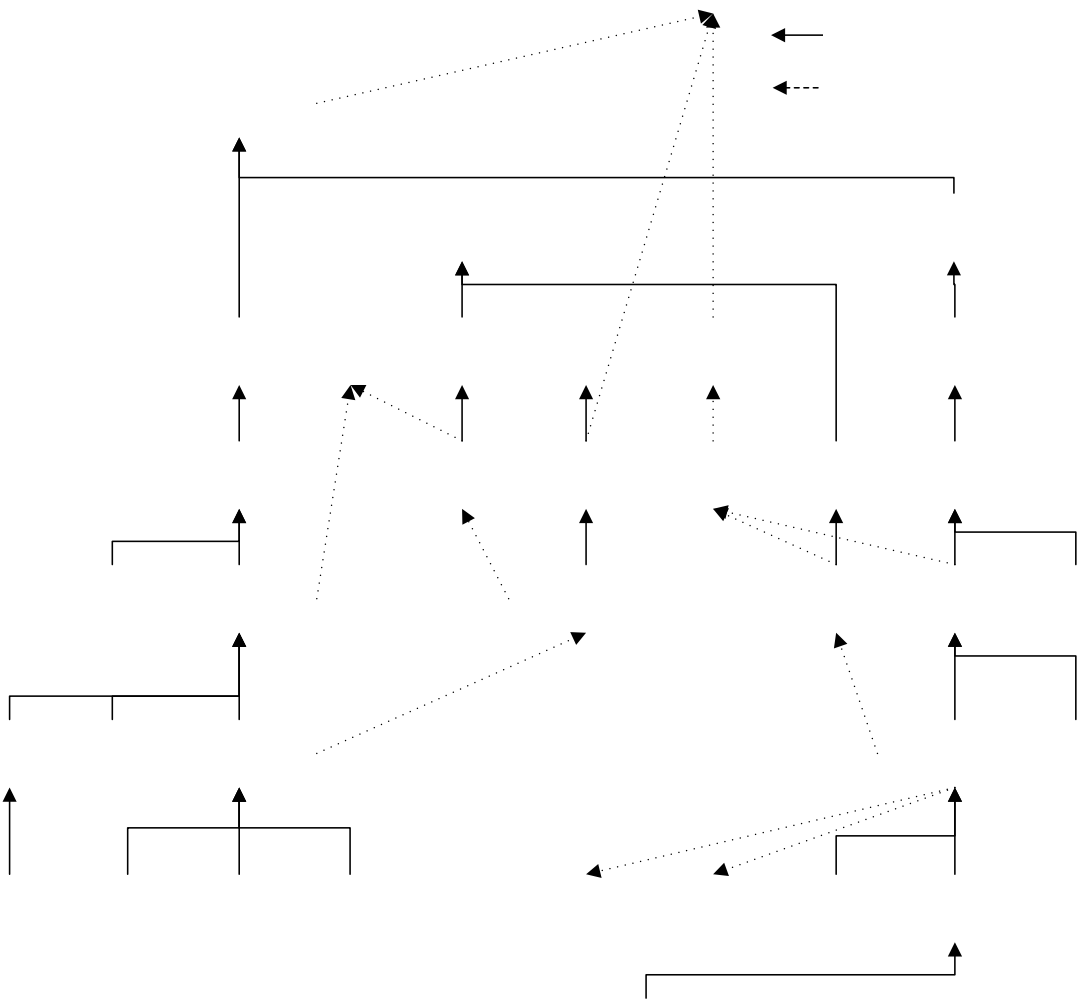


Figura 4.5: Extracto da árvore de pesquisa (WordNet 1.7.1) para estabelecer a medida proposta em 3.2 entre os conceitos ***crocodile*** e ***duck#1***.

Tal como a abordagem anterior (proposta por Resnik), nesta medida não há qualquer diferenciação na similaridade entre conceitos genéricos e conceitos mais específicos, por exemplo: a similaridade entre ***object#1*** e ***substance#1***, duas classes de conceitos que estão no topo da hierarquia de substantivos, é igual a similaridade entre ***Grass#3*** e ***Sand#2***, dois escritores de nacionalidades diferentes. Face a este problema, passámos a ter em conta a profundidade a que se encontra o conceito mais específico que generalize dois conceitos na hierarquia dos substantivos (**gcme)**, ou seja, dois conceitos com um **gcme** muito específico são considerados mais próximos que dois conceitos com um **gcme** mais geral. Esta característica, definida como ***Information Content*** do **gcme**, foi inicialmente defendida por Resnik [1995]. Na medida apresentada no seu trabalho, Resnik considera uma probabilidade de utilização de um dado conceito tendo como base um *corpus*. Sendo assim, a similaridade entre dois conceitos, **x** e **y**, é definida, pelo autor, como sendo a probabilidade de utilização do seu **gcme** (fórmula 3.4).

$$sim(x,y) = -\log P(gmce_{x,y}) \quad \textbf{(3.4)}$$

Apesar de Resnik ter considerado as probabilidades das palavras presentes no *BrownCorpus* em bruto, ou seja, sem estarem etiquetadas quanto ao seu significado, optámos por utilizar o *SemCor*[9] como *corpus* de base devido à sua integração com o WordNet. Como pré-processamento, optámos por calcular para cada um dos conceitos presentes no WordNet a sua probabilidade de utilização com base no corpus *SemCor*, existindo assim uma função p: $C \rightarrow$ [0,1], tal que para qualquer conceito **c** $\in C$, *P(c)* é a probabilidade de encontrar uma instância do conceito **c** (cujo estimador é definido pela equação 4.2). Esta probabilidade é estimada como sendo a razão entre a frequência de ocorrência de um conceito **c** no *corpus*, **freq(c)**, sobre o total **N** de conceitos que o *corpus* abrange.

$$\hat{P}(c) = \frac{freq(c)}{N} \quad \textbf{(4.2)}$$

A probabilidade de ocorrência de um conceito **c**, *P(c)*, é calculada com base na frequência deste conceito no *corpus* e de todos os outros que **c** generaliza. O valor final de frequência expresso para cada conceito é a soma do número de ocorrências do conceito em si mais o número de ocorrências de todos os conceitos que são generalizados por **c**.

Ao analisarmos a fórmula proposta por Resnik (3.4), verificámos que podem existir duas situações especiais de similaridades: entre conceitos que não possuam qualquer conceito generalizador comum; e entre conceitos que não apareçam no corpus. Para tratar a primeira situação, foi criado um conceito artificial mais geral (***root***), com probabilidade de ocorrência de valor 1, para os casos em que dois conceitos não possuam outro conceito generalizador na taxonomia. Em relação ao facto de um conceito do WordNet não existir no *corpus*, ao invés de lhe atribuir uma frequência de valor 0 (e consequentemente, com similaridade impossível de ser calculada), preferimos fazer um alisamento (*smoothing)* de uma unidade, atribuindo o valor 1 de frequência (o que perfaz uma probabilidade de 6,06e-07). Do total de conceitos presentes no WordNet, 81% encontrava-se nesta situação, enquanto que os restantes 19% de conceitos que se encontravam no *corpus* também tiveram as suas frequências incrementadas em 1 unidade.

Uma vez que os valores de probabilidade de existência no *corpus* (fórmula 4.2) estão compreendidos no intervalo [0,1], os valores possíveis de similaridade obtidos através da fórmula 3.4 estão no intervalo [0, +∞[, ou seja, quanto maior a probabilidade de um conceito generalizador comum (gcme), menor é o seu conteúdo de informação. De forma a ter valores de similaridade passíveis de comparação com os obtidos por outras fórmulas, e como a menor

[9] Disponibilizado livremente por um dos seus autores em [Mihalcea, 2003].

probabilidade (para conceitos muito específicos) possível é de 6,06e-07, decidimos utilizar o seguinte factor de normalização para a fórmula 3.4:

$$sim(x,y) = \frac{-\log P(gmce_{x,y})}{-\log P(6{,}06e-07)} \cong \frac{-\log P(gmce_{x,y})}{6{,}22} \quad \textbf{(4.3)}$$

Para exemplificar as probabilidades associadas aos conceitos do WordNet calculadas com base no corpus *SemCor*, apresentamos na Tabela 4.6 as probabilidades dos vários significados das palavras contextualizadas no Sub Capítulo 4.2 e de alguns conceitos **gmce** que serão utilizados mais adiante.

Conceito	P(Conceito)	Significado no WordNet
artifact	6,95e-2	a man-made object taken as a whole
bank#1	5,36e-4	a financial institution
bank#2	6,3e-5	sloping land
bank#6	4,24e-6	a container for saving money
canoe	1,82e-6	small and light boat
crocodile	6,06e-7	large voracious aquatic reptile
entity	0,49	that which has its own physical existence (living or nonliving)
instrumentality#3	1,95e-2	an artifact (or system of artifacts) used in accomplishing some end
object#1	0,28	a tangible and visible entity
organism#1	6,35e-2	an independent living thing
sand#1	6,67e-6	a loose material; grains of rock or coral
vertebrate	3,04e-3	animals having a bony or cartilaginous skeleton

Tabela 4.6: Probabilidade de ocorrência no corpus *SemCor* de alguns conceitos do WordNet.

Ao compararmos estes conceitos de acordo com a fórmula de Resnik normalizada (4.3), obtemos os valores de similaridade da Tabela 4.7. Os conceitos generalizadores mais específicos comuns (**gcme**) a dois conceitos podem ser verificados na Figura 4.3.

Podemos observar, de acordo com a fórmula actual, que o conceito ***sand#1*** "areia" é tão similar ao conceito ***bank#2*** "banco de areia" como ao conceito ***bank#6*** "cofre", o que claramente não corresponde ao senso-comum. E além disso, a similaridade entre ***sand#1*** e ***bank#2*** é a mesma para qualquer combinação de conceitos que sejam a especialização de ***sand#1*** e/ou ***entity#1***, que demonstra também a insuficiência dos parâmetros considerados nesta medida. Analisaremos em maior escala no Capítulo 5 o desempenho desta medida de similaridade de acordo com uma bateria de testes efectuados.

conceitos		$gcme_{x,y}$	$similaridade_{Resnik}$
canoe	**bank#1**	*root*	-log(1)/6,22 = **0**
	bank#2	object#1	-log(0,28)/6,22 ≅ **0,09**
	object#1	object#1	-log(0,28)/6,22 ≅ **0,09**
	bank#6	instrumentality#3	-log(1,95e-02)/6,22 ≅ **0,27**
	instrumentality#3	instrumentality#3	-log(1,95e-02)/6,22 ≅ **0,27**
crocodile	**duck#1**	vertebrate#1	-log(3,04e-03)/6,22 ≅ **0,40**
	vertebrate	vertebrate	-log(3,04e-03)/6,22 ≅ **0,40**
	duck#4	object#1	-log(0,28)/6,22 ≅ **0,09**
	object#1	object#1	-log(0,28)/6,22 ≅ **0,09**
sand#1	**bank#1**	*root*	-log(1)/6,22 = **0**
	bank#2	entity	-log(0,49)/6,22 ≅ **0,05**
	entity	entity	-log(0,49)/6,22 ≅ **0,05**
	bank#6	entity	-log(0,49)/6,22 ≅ **0,05**

Tabela 4.7: Cálculo de similaridade (com base na medida de Resnik) entre alguns pares de significados (**x** e **y**) das palavras presentes no mapa conceptual da Figura 4.2.

Face a estas incoerências, verificámos que a medida de distância semântica proposta inicialmente por Jiang e Conrath [1997] (fórmula 3.5), que depois foi adaptada para uma medida de similaridade (fórmula 4.1), considera a probabilidade de cada conceito a comparar, além do conceito generalizador comum mais específico. Esta medida foi normalizada tendo em conta os conceitos (*w* e *z*, como veremos a seguir) que maximizassem a distância semântica a fim de produzir um medida de similaridade entre [0,1] (fórmula 4.4):

$$sim(x,y) = 1 - \frac{2 * \log P(gmce_{xy}) - (\log P(x) + \log P(y))}{2 * \log P(gmce_{wz}) - (\log P(w) + \log(P(z))} \qquad \textbf{(4.4)}$$

Onde *w* e *z* são os conceitos que possuem a maior distância semântica de acordo com as probabilidades do *corpus SemCor* 1.7.1. e hierarquia taxonómica do WordNet 1.7.1. Constatou-se que o valor de distância máxima atribuída pela fórmula 3.5 seria entre conceitos que não estivessem presentes no *corpus SemCor* e possuíssem como conceito generalizador comum o conceito artificial criado com probabilidade de ocorrência de valor 1. Ao compararmos a similaridade entre os mesmos conceitos anteriores utilizando, desta vez, a fórmula adaptada 4.4 obtemos os valores apresentados na Tabela 4.8.

Por esta medida, é possível verificar que o conceito ***crocodile*** está mais próximo semanticamente do significado referente ao animal "pato" da palavra ***duck*** do que ao significado correspondente a um tipo de "tecido". Ao contrário da medida anterior, vemos uma menor distância entre ***crocodile*** e ***vertebrate*** (já que um conceito generaliza o outro), quando

confrontado com qualquer especialização do segundo conceito (no nosso caso ***duck#1***). Isto porque, em conjunto com a probabilidade do conceito generalizador comum, são também consideradas as probabilidades individuais dos conceitos a analisar. Consequentemente, quanto mais genérico, maior é a probabilidade de ocorrência do conceito, resultando numa similaridade maior (de acordo com a fórmula 4.4).

Tal como acontecia com as outras medidas que apenas consideram as relações semânticas de generalização/especialização de substantivos presentes no WordNet, o significado ***sand#1*** não é considerado tão próximo ao conceito "banco de areia" (***bank#2***), em comparação com outros significados de ***bank*** como seria de esperar. Isto porque estas duas entidades estão classificadas em ramos completamente distintos na taxonomia: enquanto que um é considerado uma substância, o outro é um tipo de objecto. Outro equívoco aparece na similaridade relativamente maior entre os conceitos ***sand#1*** e ***bank#1***. Isto acontece porque na língua Inglesa a palavra ***bank*** é correntemente mais utilizada como "instituição financeira", tendo este conceito uma probabilidade 9 vezes superior ao significado "banco de areia". Deste modo, este significado é sempre mais valorizado em detrimento dos outros possíveis da palavra *bank*.

conceitos		**$gcme_{a,b}$**	**$similaridade_{Adaptada}$**
w	**z**	*root*	**0**
canoe	**bank#1**	*root*	**0,28**
	bank#2	object#1	**0,29**
	object#1	object#1	**0,58**
	bank#6	instrumentality#3	**0,38**
	instrumentality#3	instrumentality#3	**0,70**
crocodile	**duck#1**	vertebrate	**0,46**
	vertebrate	vertebrate	**0,70**
	duck#4	object#1	**0,09**
	object#1	object#1	**0,54**
sand#1	**bank#1**	*root*	**0,32**
	bank#2	entity	**0,30**
	entity	entity	**0,61**
	bank#6	entity	**0,20**

Tabela 4.8: Cálculo de similaridade (com base na medida de Jiang e Conrath) entre alguns pares de significados (**x** e **y**) das palavras presentes no mapa conceptual da Figura 4.2.

Para melhorar esta medida é importante verificar até que ponto não poderíamos considerar também a informação sobre a utilização das palavras em conjunto no cálculo da similaridade. Quando existem conceitos interligados, seja num mapa conceptual ou texto, que não partilhem de uma similaridade taxonómica, isto é, não pertençam a uma mesma classe na

árvore hierárquica de substantivos, mas que sejam frequentemente associados um a outro (*pinguim – Antárctica*, por exemplo) quaisquer das medida de similaridade (excepto em alguns casos a medida proposta por Hirst St. Onge) falham ao estabelecer uma relação cujo termo mais correcto não seja similaridade mas sim *semantic relatedness*[10]. Na Tabela 4.8, constatámos que os conceitos ***canoe*** e ***bank#6*** estão mais próximos na taxonomia do WordNet por serem ambos um tipo de instrumento com uma dada utilidade. No entanto, achamos mais natural a associação de ***canoe*** com o conceito ***bank#2***, por serem mais frequentemente utilizados em conjunto (por exemplo, na frase "*they pulled the canoe up on the bank*"). E, de facto, a utilização do primeiro par de conceitos (com base no *corpus SemCor*) é muito menos frequente em comparação com o segundo.

Apesar de não estarem disponíveis, até há bem pouco tempo, *corpora* em grande escala anotados (desambiguados) quanto ao seu significado semântico para a estimação de *n-grams*[11], tem sido feito um grande esforço neste sentido, nomeadamente na criação e manutenção do recurso *SemCor* e agora recentemente com o *ExtendedWordNet*, permitindo um primeiro passo para uma análise de *semantic relatedness* entre conceitos.

Como resultado do processo de contextualização, poderão surgir nós do mapa conceptual que se referem a um mesmo conceito do WordNet, caso estejam originalmente etiquetados com palavras sinónimas (por exemplo: *car* e *automobile*). Neste caso, optámos por fundir estes nós (agregando deste modo, todas as ligações dos nós anteriores) num único com maior peso no mapa. Deste modo, o grau de ligação deste nó será calculado com base na reunião das ligações originais.

Uma vez compreendida a similaridade semântica entre conceitos, iremos, na próxima Secção, generalizar a noção de similaridade entre conceitos para mapas conceptuais (que em primeiro lugar, são redes de conceitos que representam um dado domínio).

3.2. Similaridade Global entre Mapas Conceptuais

Nesta Secção, iremos averiguar se existirá alguma forma de comparar semanticamente mapas conceptuais com base nos seus elementos constituintes: nós (conceitos) e arcos (relações). Para tal, debruçámo-nos sobre abordagens existentes na teoria dos grafos e processamento semântico que lançassem algumas ideias para a resolução do problema.

[10] *Semantic relatedness* é um relacionamento que considera mais a utilização das palavras em conjunto do que o conceito de similaridade tradicional, que apenas considera características comuns (classificação) entre os objectos representados pelas palavras (esta contraposição é feita pela primeira vez no Sub Capítulo 3.1)

[11] Modelos estocásticos baseado em *corpus* que permitem prever a palavra que vriá a seguir numa dada sequência.

Uma vez que os mapas conceptuais são representados computacionalmente por grafos, existem métodos de comparação que se baseiam na estrutura dos grafos, nomeadamente isomorfismo e isomorfismo de subgrafos[12]. Mas, se por um lado estes métodos envolvem alguma complexidade computacional (sendo o isomorfismo de subgrafos reconhecido actualmente com sendo um problema NP-completo), por outro não queremos apenas uma comparação estrutural mas principalmente semântica entre o conteúdo representado pelos mapas.

No processamento semântico, existem diversas abordagens para a comparação de conceitos ente si, mas no que toca a conjuntos de conceitos presentes em textos ou representações mais estruturadas, geralmente recorre-se à teoria de conjuntos e classificações por palavras-chave para estabelecer uma medida de similaridade. No caso das estruturas conceptuais, começam a ser desenvolvidos métodos de pesquisa (*retrieval*), no entanto estes baseiam-se em métodos morosos de *matching* estrutural e o processamento semântico é limitado ao conhecimento introduzido manualmente sem recurso a outras bases de conhecimento.

A nossa ideia inicial foi procurar, de uma forma simples e eficiente, a comparação entre mapas conceptuais com base na sua estrutura e significado recorrendo a um recurso léxico-semântico suficientemente abrangente de forma a englobar os domínios representados por estes. Deste modo, foi formalizado o seguinte raciocínio: dados dois mapas conceptuais MC_1 e MC_2, a similaridade (ou inversamente, a distância semântica) entre estes reflecte a relação global dos seus elementos com base na organização taxonómica do WordNet.

De acordo com esta formulação, veremos a seguir algumas abordagens que serão adaptadas e outras originalmente propostas para a comparação semântico-estrutural de mapas conceptuais.

Com o objectivo de exemplificar cada uma destas futuras abordagens, iremos apresentar cada um dos algoritmos aplicando-os a um exemplo de comparação semântica entre 3 mapas conceptuais que designaremos MapaA, MapaB e MapaC (ver, respectivamente, Figuras 4.2, 4.6 e 4.7). Estes são assim designados para evitar uma pré-associação de ideias pelos títulos que lhe possam ser atribuídos, em vez da análise detalhada do seu conteúdo. Posteriormente, no Capítulo 5, serão apresentados uma variedade de testes de similaridade efectuados sobre um conjunto maior e mais representativos de mapas.

[12] Em anexo, encontram-se alguns tópicos sobre Teoria dos Grafos.

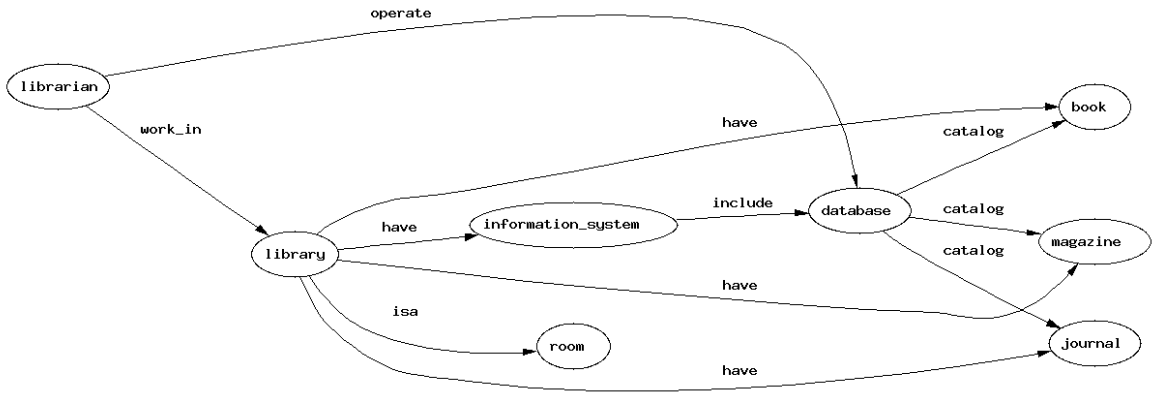


Figura 4.6: Exemplo simples de uma mapa conceptual.

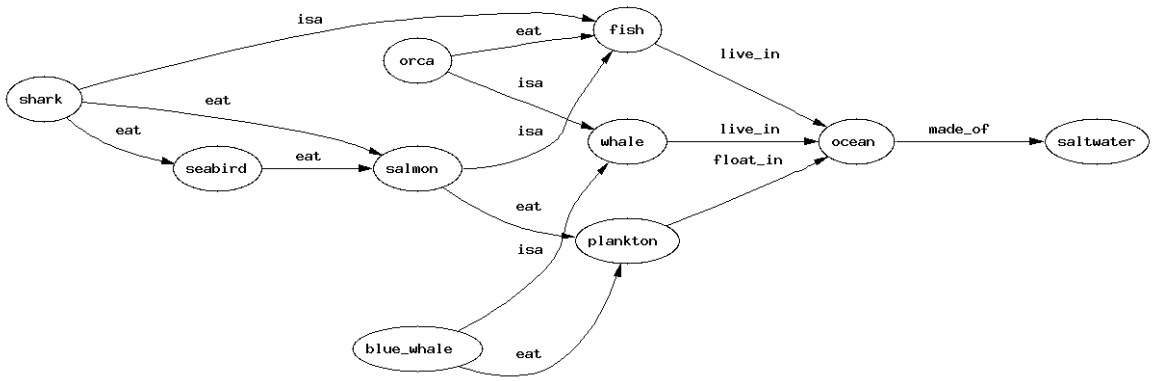


Figura 4.7: Exemplo simples de uma mapa conceptual.

Assim como o primeiro mapa conceptual (MapaA - Figura 4.2) foi contextualizado para ser possível a utilização da informação semântica presente no WordNet, o mesmo processo foi aplicado aos outros mapas conceptuais (MapaB - Figura 4.6 e MapaC - Figura 4.7) tendo como resultado do processo de desambiguação os valores das Tabelas 4.9 e 4.10 (respectivamente).

Na Tabela 4.9, podemos verificar que as palavras *book* e *magazine* possuem mais do que um significado correcto. Estes significados, apesar de muito semelhantes entre si, estão classificados independentemente no WordNet, e como tal, ambos ilustram correctamente a ideia transmitida pela palavra no contexto apresentado no MapaB. O facto de uma palavra poder possuir diferentes significados com um certo grau de semelhança é conhecido como polissemia[13] e, actualmente, não há nenhuma informação no WordNet sobre a semelhança entre os diversos significados de uma mesma palavra.

[13] Termo apresentado inicialmente na secção 2.2.2.

Conceito		Significados
book	**#1**	**a copy of a written work or composition that has been published**
	#2	physical objects consisting of a number of pages bound together
	#3	a record in which commercial accounts are recorded
	#4	a number of sheets (ticket or stamps etc.) bound together on one edge
	#5	a compilation of the known facts regarding something or someone
	#6	a major division of a long written composition
	#7	a collection of rules or prescribed standards on the basis of which decisions are made
	#8	a written version of a play or other dramatic composition; used in preparing for a performance
	#9	sacred writings of Islam revealed by God to the prophet Mohammed during his life at Mecca and Medina; divided into 114 chapters
	#10	the sacred writings of the Christian religion
database		**an organized body of related information**
information_system		**system consisting of the network of all communication channels used within an organization**
journal	**#1**	**a daily written record of (usually personal) experiences and observations**
	#2	a periodical dedicated to a particular subject
	#3	a ledger in which transactions have been recorded as they occurred
	#4	a record book as a physical object
	#5	the part of the axle contained by a bearing
librarian		**a professional person trained in library science and engaged in library services**
library#1	**#1**	**a room where books are kept**
	#2	a collection of literary documents or records kept for reference or borrowing
	#3	a depository built to contain books and other materials for reading and study
	#4	(computing) a collection of standard programs and subroutines that are stored and available for immediate use
	#5	a building that houses a collection of books and other materials
magazine	**#1**	**a periodic paperback publication**
	#2	product consisting of a paperback periodic publication as a physical object
	#3	a business firm that publishes magazines
	#4	a light-tight supply chamber holding the film and supplying it for exposure as required
	#5	a storehouse (as a compartment on a warship) where weapons and ammunition are stored
	#6	a metal frame or container holding cartridges; can be inserted into an automatic gun
room	**#1**	**an area within a building enclosed by walls and floor and ceiling**
	#2	space for movement
	#3	opportunity for
	#4	the people who are present in a room

Tabela 4.9: Resultado do processo de contextualização das palavras do mapa conceptual do MapaB. Dentro dos significados possíveis de cada conceitos, os que estão a sombreado são os escolhidos manualmente, enquanto que os em negrito são os resultantes do processo de contextualização automática. Apenas o conceito *journal* não é correctamente desambiguado, sendo o 2º significado desta palavra mais apropriado neste caso: *"a periodical dedicated to a paticular subject"*.

Conceito		Significado escolhido
blue_whale		**largest mammal ever known; bluish-gray migratory whalebone whale mostly of southern hemisphere**
fish	**#1**	**any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills**
	#2	the flesh of fish used as food
	#3	(astrology) a person who is born while the sun is in Pisces
	#4	the twelfth sign of the zodiac; the sun is in this sign from about February 19 to March 20
ocean	**#1**	**a large body of water constituting a principal part of the hydrosphere**
	#2	anything apparently limitless in quantity or volume
orca		**predatory black-and-white toothed whale with large dorsal fin; common in cold seas**
plankton		**the aggregate of small plant and animal organisms that float or drift in great numbers in fresh or salt water**
salmon	**#1**	**any of various large food and game fishes of northern waters; usually migrate from salt to fresh water to spawn**
	#2	a tributary of the Snake River
	#3	flesh of any of various marine or freshwater fish of the family Salmonidae
saltwater		**water containing salts**
seabird		**a bird that frequents coastal waters and the open ocean: gulls; pelicans; gannets; cormorants; albatrosses; petrels; etc.**
shark	**#1**	**any of numerous elongate mostly marine carnivorous fishes with heterocercal caudal fins and tough skin covered with small toothlike scales**
	#2	a person who is ruthless and greedy and dishonest
	#3	a person who is unusually skilled in certain ways
	#1	a very large person; impressive in size or qualities
	#2	any of the larger cetacean mammals having a streamlined body and breathing through a blowhole on the head

Tabela 4.10: Resultado do processo de contextualização das palavras do mapa conceptual do MapaC. Dentro dos significados possíveis de cada conceitos, os que estão a sombreado são os escolhidos manualmente, enquanto que os em negrito são os resultantes do processo de contextualização automática. Todos os conceitos foram desambiguados correctamente.

3.2.1. Contraste de Características entre Mapas Conceptuais

Tal como visto na Secção 3.2.1, este método consiste em comparar objectos descritos por características com base na teoria de conjuntos. Podemos adaptar este raciocínio ao nosso objecto de estudo e afirmar que um mapa conceptual é representado (e caracterizado) pelos conceitos que o integram. Deste modo, surge-nos a seguinte questão: como determinar se duas características são consideradas equivalentes? Iremos contabilizar somente coincidências exactas ou características similares? Nos mapas conceptuais, isto traduzir-se-ia em conceitos de idêntico significado ou com um certo grau de similaridade. Decidimos então, de forma a enriquecer o próprio método em si, considerar também um certo grau de abstracção na pesquisa de conceitos comuns. Ou seja, além dos conceitos presentes no mapa conceptual, o espaço de procura também é formado pelos conceitos que os generalizam até uma dada distância **n** (número de arcos *isa* na

árvore hierárquica do WordNet). Como, à medida que subimos na hierarquia, estamos a abranger cada vez mais conceitos, e consequentemente conceitos que representam um tipo de informação muito mais genérica, decidimos estipular um peso associado a cada conceito do espaço de procura que representa a sua distância no processo de indução. Deste modo, é estipulado um factor multiplicativo associado a cada generalização realizada para a obtenção do conceito.

A análise de similaridade pelos coeficientes estatísticos dos próximos pontos (*Dice*, *Jaccard,* Sobreposição e Cosseno) foi adaptada aos mapas conceptuais de acordo com os seguintes pressupostos:

i. |**X**| é o número de conceitos (já contextualizados e associados a significados-*synsets* do WordNet) presentes num mapa conceptual **X**;

ii. Um conceito $\mathbf{c_1}$ é considerado **comum** a um outro conceito $\mathbf{c_2}$ com um certo **factor de similaridade, fs**, cuja a distância total não ultrapasse um **limite L**. O factor de similaridade (4.4) entre dois conceitos assume que a semelhança entre dois conceitos diminui exponencialmente à medida que a distância na taxonomia hierárquica do WordNet aumenta.

$$fs(c_1, c_2) = \alpha^{len(c_1, c_2)}, \alpha \in]0,1[\quad \textbf{(4.4)}$$

Onde len(c_1,c_2) é o número de arcos *isa* entre $\mathbf{c_1}$ e **gmce** somado aos arcos entre $\mathbf{c_2}$ e **gmce** desde que este gmce se encontre a uma distância de até L arcos em relação a c_1 e c_2, ou 0 em caso contrário. O valor mínimo de α (>0) tem como objectivo evitar um factor de similaridade infinito (quando len() = 0). Já o seu valor máximo (<1), deve-se à intenção de preservar um valor maior para as correspondências originais entre conceitos.

iii. Na comparação de similaridade global entre dois mapas conceptuais, cada conceito presente num mapa poderá ter no máximo **uma** correspondência (conceito considerado comum) no outro mapa. Deste modo, caso dois conceitos $\mathbf{c_{x1}}$ e $\mathbf{c_{x2}}$ pertencentes ao mapa conceptual **X** sejam considerados comuns a um mesmo conceito $\mathbf{c_{y1}}$ do mapa **Y**, será contabilizado apenas o de **maior** factor de similaridade a este último.

iv. |**X∩Y**| é a **soma** dos factores de similaridade dos conceitos considerados comuns (correspondências) entre dois mapas conceptuais **X** e **Y**. O algoritmo de cálculo de intersecção de conceitos entre dois mapas conceptuais é apresentado no Quadro 4.2.

```
(1)Intersecção(mapa_x[], mapa_y[], limite)
     interseccao ← 0
     i ← limite
     j ← 0
(2)  enquanto j ≤ limite
     fazer
       para cada c_i ∈ mapa_x[]
       fazer
         para cada c_j ∈ mapa_y[]
         fazer
(3)        se c_i.gen[i] = c_j.gen[j]
           entao
             interseccao ← interseccao + fs(c_i,c_j)
(4)          retira c_i de mapa_x[]
             retira c_j de mapa_y[]
           fim se
         fim para
       j ← j + 1
     fim enquanto
 (5)  i ← 0
     j ← limite
     enquanto i < limite
     fazer
       para cada c_i ∈ mapa_x[]
       fazer
         para cada c_j ∈ mapa_y[]
         fazer
           se c_i.gen[i] = c_j.gen[j]
           entao
             interseccao ← interseccao + fs(c_i,c_j)
             retira c_i de mapa_x[]
             retira c_j de mapa_y[]
           fim se
         fim para
       fim para
       i ← i + 1
     fim enquanto
     devolve interseccao
```

Quadro 4.2: Intersecção de dois conjuntos de conceitos que representam mapas conceptuais. As linhas numeradas serão detalhadamente explicadas a seguir.

(1) Cada mapa conceptual é representado pelo conjunto de conceitos que constituem os seus nós (**mapa[]**). Cada conceito do mapa possui uma lista de outros conceitos ascendentes que o generaliza (**gen**), desde o mais específico, **gen[0]**, até o mais geral que se encontra a um dado **limite**, **gen[limite]**. Como consequência directa, o factor de similaridade, **fs**, entre o conceito original e estas generalização vai diminuindo à medida que subimos na hierarquia. A título de exemplo, utilizámos esta representação para os 3 mapas conceptuais (ver Tabela 4.11) anteriormente definidos: MapaA, MapaB e MapaC.

(2) O **limite** fornecido ao algoritmo é utilizado como valor máximo de profundidade na comparação entre conceitos. Assume-se aqui que a Intersecção entre dois mapas é feita

sucessivamente (ou seja, invocada por um outro algoritmo, como veremos adiante) com o valor do limite a crescer entre cada chamada.

(3) É feita uma comparação entre as generalizações do i-ésimo conceito, $\mathbf{c}_i$, do mapa X e do j-ésimo conceito, $\mathbf{c}_j$, do mapa Y a fim de saber se estes são idênticos.

(4) Se tal acontecer, ambos os conceitos originais são apagados dos mapas a que pertenciam até então. Isto garante que as correspondências sejam unívocas entre os conceitos.

(5) Como nem todas as correspondências poderão ter sido encontradas, é feita agora uma nova procura. Neste caso, a hierarquia de conceitos do primeiro mapa (X) é a que será percorrida no sentido ascendente.

Mesmo não tendo uma contextualização totalmente correcta para o MapaB devido à escolha de outro significado para a palavra *journal*, utilizaremos o resultado real da contextualização analisando o seu efeito sobre a comparação final sobre os mapas.

No Mapa C, o conceito *water#1* (*binary compound that occurs at room temperature as a clear odorless clearless tasteless liquid*), apesar de único, possui dois conceitos como hiperónimos directos (*binary_compound* e *liquid#1*) na hierarquia de substantivos do WordNet. Seguimos ambas generalizações até o limite estabelecido. Por uma coincidência na organização interna do WordNet, estes conceitos possuem um generalizador comum (*substance#1*). Uma vez que podemos chegar a este último por dois caminhos distintos, será escolhida para efeitos de correspondência apenas uma instância deste conceito.

Nível de Generalização	Factor de Similaridade	Mapas Conceptuais		
		A	**B**	**C**
0	1	bank#2 beach canoe crocodile duck#1 grass#1 river sand#1	book#1 database information_system journal#1 librarian library#1 magazine#1 room#1	blue_whale fish#1 ocean#1 orca plankton salmon#1 saltwater seabird shark#1 whale#2
1	α	slope#1 geological_formation small_boat crocodilian anseriform_bird gramineous_plant watercourse#2 soil#2	publication#1 information#1 system#2 written_material professional_person room#1 public_press area#4	baleen_whale aquatic_vertebrate body_of_water dolphin#2 organism#1 salmonid water#1 aquatic_bird selachian ceatacean
2	α^2	geological_formation natural_object boat#1 diapsid waterbird herb#1 body_of_water ground#3	piece_of_work subject_matter instrumentality#3 communication#1 adult#1 area#4 print_media structure#1	whale#2 vertebrate entity toothed_whale living_thing soft-finned_fish binary_compound liquid#1 bird cartilaginous fish aquatic_mammal
3	α^3	natural_object object#1 watercraft reptile aquatic_bird vascular_plant entity material#1	product#2 communication#2 artifact social_relation person#1 structure#1 medium#1 artifact	ceatacean chordate whale#2 object#1 teleost_fish chemical_compound fluid#1 vertebrate fish#1 pacental_mammal
4	α^4	object#1 entity craft vertebrate bird plant entity substance#1	object#1 social_relation object#1 abstraction#6 organism#1 artifact means#2 object#1	aquatic_mammal animal cetacean entity bony_fish substance#1 chordate aquatic_vertebrate mammal

Tabela 4.11: Representação interna dos mapas conceptuais A, B e C após o processo de contextualização, para a aplicação futura da abordagem de contraste de características. Podemos observar inclusive a evolução do factor de similaridade ao longo das generalizações.

Para exemplificar a operação de intersecção entre conjuntos de conceitos, iremos aplicá-lo sobre os 3 mapas conceptuais, em dois pares de cada vez (MapaA e MapaB) e (MapaA e

MapaC), como podemos verificar nas Tabelas 4.12 e 4.13. O resultado desta operação será utilizado futuramente nos próximos pontos.

Limite	Correspondências entre Mapas Conceptuais		A∩B
	A	B	
0			0
1			0
2			0
3			0
4	object#1$_A$[3] object#1$_A$[4]	object#1$_B$[4] object#1$_B$[4]	$\alpha^3\alpha^4$ + $\underline{\alpha^4}\underline{\alpha^4}$ $\alpha^8+\alpha^7$

Tabela 4.12: Intersecção entre os mapas conceptuais A e B por diferentes limites de pesquisa. Somente num nível considerado mais profundo (após 4 generalizações sucessivas) é que foram encontradas algumas correspondências entre conceitos. O valor da intersecção é fortemente dependente do grau atribuído às generalizações na taxonomia do WordNet.

Limite	Correspondências entre Mapas Conceptuais		A∩C
	A	C	
0			0
1			0
2	body_of_water$_A$[2]	body_of_water$_C$[1]	$\alpha^2\alpha = \alpha^3$
3	body_of_water$_A$[2] aquatic_bird$_A$[3] object#1$_A$[3]	body_of_water$_C$[1] aquatic_bird$_C$[1] object#1$_C$[3]	$\alpha^2\alpha$ + $\alpha^3\alpha$ + $\underline{\alpha^3}\underline{\alpha^3}$ $\alpha^6+\alpha^4+\alpha^3$
4	body_of_water$_A$[2] aquatic_bird$_A$[3] object#1$_A$[3] vertebrate$_A$[4] substance#1$_A$[4]	body_of_water$_C$[1] aquatic_bird$_C$[1] object#1$_C$[3] vertebrate$_C$[2] substance#1$_C$[4]	$\alpha^2\alpha$ + $\alpha^3\alpha$ + $\alpha^3\alpha^3$ + $\alpha^4\alpha^2$ + $\underline{\alpha^4}\underline{\alpha^4}$ $\alpha^8+2\alpha^6+\alpha^4+\alpha^3$

Tabela 4.13: Intersecção entre os mapas conceptuais A e C por diferentes limites de pesquisa. Com um limite 2 de pesquisa na taxonomia do WordNet já é possível estabelecer um valor de intersecção não-nulo entre os dois mapas.

À medida que são estabelecidas correspondências entre os mapas, os conceitos são retirados dos conjuntos a que pertencem, o que implica num menor número de generalizações efectuadas. Na Tabela 4.13, os conceitos *body_of_water* e *aquatic_bird* são eliminados em ambos os conjuntos de pesquisa após a verificação de correspondência entre os dois mapas. Se

apenas considerássemos o valor resultante da intersecção entre mapas conceptuais, poderíamos já concluir que o MapaA está semanticamente mais próximo do MapaC face ao MapaB.

- **Coeficiente *Dice***

O coeficiente de *Dice* (3.9) aplicado ao cálculo da similaridade entre mapas conceptuais segue o algoritmo apresentado no Quadro 4.3, onde é feita uma procura em largura bidireccional a partir de ambos os conceitos de cada mapa conceptual de modo a dividir a complexidade e tempo de processamento do algoritmo (o valor final $O(L(|X||Y|+|X|+|Y|))$, será o custo uniforme de cada vez que se avança um nível (L), não havendo o perigo de encontrar uma combinação menos vantajosa em primeiro lugar). O número de conceitos originais (ou seja, sem contar com eventuais conceitos mais genéricos que foram explorados durante o processamento do algoritmo) de um mapa X é representado por |**X**|.

$$SIM(x,y) = \frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid} \qquad \textbf{(3.9)}$$

```
(1)Similaridade_Por_Dice(X, Y, α, L)
(2)  mapa_x[] ← Conceitos(X)
     mapa_y[] ← Conceitos(Y)
     nivel ← 0
     valor_comum ← 0
(3)  para cada c_i ∈ mapa_x[]
       c_i.gen[nivel] ← c_i
     fim para

     para cada c_j ∈ mapa_y[]
       c_j.gen[nivel] ← c_j
     fim para
     enquanto nivel ≤ L
     fazer
(4)    valor_comum ← valor_comum + Intersecção(mapa_x[], mapa_y[], nivel)
       nivel ← nivel + 1
       para cada c_i ∈ mapa_x[]
       fazer
(5)      c_i.gen[nivel] ← Generalização(c_i.gen[nivel-1])
       fim para
       para cada c_j ∈ mapa_y[]
       fazer
         c_j.gen[nivel] ← Generalização(c_j.gen[nivel-1])
       fim para
     fim enquanto
     similaridade ← 2 * valor_comum/(|X| + |Y|)
     devolve similaridade
```

Quadro 4.3: Algoritmo para o cálculo de similaridade pelo coeficiente de Dice entre dois mapas conceptuais representados pelos conjuntos de conceitos X e Y (respectivamente). As linhas numeradas serão detalhadamente explicadas a seguir.

(1) O algoritmo recebe como dados de entrada: os mapas a serem comparados (**X** e **Y**), o grau do factor de similaridade à medida que os conceitos dos mapas vão sendo generalizados (**α**) e, finalmente, o limite máximo de pesquisa na taxonomia do WordNet.

(2) Serão apenas considerados os conceitos de cada mapa conceptual. A função **Conceitos** devolve a lista dos nós de um mapa conceptual.

(3) Os conjuntos de conceitos são inicializados com os conceitos originais dos respectivos mapas (nível 0 de generalização).

(4) É chamada a função de **Intersecção** definida no Quadro 4.2 consecutivamente em que apenas é variado o nível de generalização. Isto permite maximizar as correspondências, já que quanto menor é o nível de generalização dos conceitos maior é o seu peso. Os conjuntos de conceitos são actualizados à medida que as correspondência vão sendo detectadas.

(5) Subir mais um nível na taxonomia para os conceitos restantes através da função **Generalização.** Esta obtém basicamente a generalização (relação hiperónimo-*isa* do WordNet) imediata (um nível acima) de um dado conceito, caso este exista. Se não existir, por ser um conceito de topo na hierarquia, a função não retorna nada.

Para ilustrar a utilização deste método iremos aplicá-lo sobre os 3 mapas conceptuais de exemplo e verificar a similaridade relativa entre estes. Como a similaridade é calculada comparando-se 2 objectos, propomos esclarecer a questão: o MapaA é semanticamente mais similar ao MapaB ou ao MapaC? Intuitivamente, ao analisar o conteúdo dos mapas conceptuais, verificamos que o contexto apresentado pelo MapaC é mais próximo do MapaA, só que até que ponto, neste caso particular, as medidas de similaridade confirmam esta afirmação? Ao considerar a intersecção de conceitos entre os mapas conceptuais em questão (Tabelas 4.12 e 4.13), obtemos as seguintes similaridades globais ente os mapas conceptuais:

$$Similaridade\_por\_Dice(A,B,\alpha,4) = \frac{2\,|\,A \cap B\,|}{|\,A\,| + |\,B\,|} = \frac{2(\alpha^8 + \alpha^7)}{8+8} = \frac{2\alpha^7(\alpha+1)}{16} = \frac{\alpha^7(\alpha+1)}{8}$$

$$Similaridade\_por\_Dice(A,C,\alpha,4) = \frac{2\,|\,A \cap C\,|}{|\,A\,| + |\,C\,|} = \frac{2(\alpha^8 + 2\alpha^6 + \alpha^4 + \alpha^3)}{8+10}$$

$$= \frac{2\alpha^3(\alpha^5 + 2\alpha^3 + \alpha + 1)}{18} = \frac{\alpha^3(\alpha^5 + 2\alpha^3 + \alpha + 1)}{9}$$

$$\forall \alpha \in]0,1[, Similaridade\_por\_Dice(A,C,\alpha,4) > Similaridade\_por\_Dice(A,B,\alpha,4), qed$$

A escolha do valor 4 para o limite de generalizações possíveis deve-se ao facto de se querer considerar um valor não-nulo na intersecção entre os Mapas A e B. No Capítulo de Testes e

Resultados (5), será realizado um conjunto de teste com o objectivo de verificar o valor óptimo deste dado.

- **Coeficiente *Jaccard***

O cálculo da similaridade entre mapas conceptuais pelo coeficiente de *Jaccard* estabelece uma proporcionalidade entre os conceitos que são comuns a ambos os mapas face aos próprios de cada mapa. Além da operação de intersecção, este coeficiente requer informação sobre a união entre mapas conceptuais. Uma vez que, na Teoria dos Conjuntos, temos $|X \cup Y| = |X \cap Y| + |X - Y| + |X - Y|$, no Quadro 4.4 iremos alterar o algoritmo de intersecção entre mapas conceptuais até aqui considerado (Quadro 4.2) de forma a este devolver, além do factor de similaridade comum aos mapas, o número de elementos de comuns incluídos na intersecção.

$$SIM(x,y) = \frac{|X \cap Y|}{|X \cup Y|} \qquad \textbf{(3.10)}$$

```
Intersecção_Actualizada(mapa_i[], mapa_j[], limite, elementos_comuns)
  interseccao ← 0
  i ← limite
  j ← 0
  enquanto j ≤ limite
  fazer
    para cada c_i ∈ mapa_i[]
    fazer
      para cada c_j ∈ mapa_j[]
      fazer
        se c_i.gen[i] = c_j.gen[j]
        entao
          interseccao ← interseccao + fs(c_i, c_j)
          elementos_comuns ← elementos_comuns + 1
          retira c_i de mapa_i[]
          retira c_j de mapa_j[]
        fim se
      fim para
    fim para
    j ← j + 1
  fim enquanto
  i ← 0
  j ← limite
  enquanto i < limite
  fazer
    para cada c_i ∈ mapa_i[]
    fazer
      para cada c_j ∈ mapa_j[]
      fazer
        se c_i.gen[i] = c_j.gen[j]
        entao
          interseccao ← interseccao + fs(c_i, c_j)
          retira c_i de mapa_i[]
          retira c_j de mapa_j[]
        fim se
      fim para
    fim para
    i ← i + 1
  fim enquanto
 devolve interseccao
```

Quadro 4.4: Intersecção entre dois conjuntos de conceitos que representam mapas conceptuais actualizado para retornar também o número de elementos comuns considerados.

O algoritmo apresentado no Quadro 4.5 ilustra o processamento dos mapas para o cálculo de similaridade pelo coeficiente de *Jaccard*. É de realçar que o conjunto formado pelos conceitos de cada mapa e as suas respectivas generalizações ($mapa_i[]$) são actualizados (ou seja, os seus elementos são retirados) à medida que são estabelecidas correspondências. Deste modo, cada mapa resultante representa a diferença entre os dois conjuntos.

```
Similaridade_Por_Jaccard(X, Y, α, L)
  mapa_x[] ← Conceitos(X)
  mapa_y[] ← Conceitos(Y)
  nivel ← 0
  valor_comum ← 0
  elementos_comuns ← 0
  para cada c_i ∈ mapa_x[]
  fazer
    c_i.gen[nivel] ← c_i
  fim para
  para cada c_j ∈ mapa_y[]
  fazer
    c_j.gen[nivel] ← c_j
  fim para
  enquanto nivel ≤ L
  fazer
    valor_comum ← valor_comum + Intersecção_Actualizada(mapa_x[], mapa_y[],
                  nivel, elementos_comuns)
    nivel ← nivel + 1
    para cada c_i ∈ mapa_x[]
    fazer
      c_i.gen[nivel] ← Generalização(c_i.gen[nivel-1])
    fim para
    para cada c_j ∈ mapa_y[]
    fazer
      c_j.gen[nivel] ← Generalização(c_j.gen[nivel-1])
    fim para
  fim enquanto
  similaridade ← valor_comum/(elementos_comuns + Dimensão(mapa_x[]) +
                 Dimensão(mapa_y[]))

  devolve similaridade
```

Quadro 4.5: Algoritmo para o cálculo de similaridade pelo coeficiente de Jaccard entre dois mapas conceptuais representados, respectivamente, pelos conjuntos de conceitos X e Y.

Deste modo, o factor de similaridade (ou intersecção) obtida é a mesma que o algoritmo anterior para a operação de intersecção, só que neste caso recebemos uma informação adicional por parâmetro (número de elementos comuns considerados na intersecção) e utilizamos outros conjuntos para dimensionar (através da função **Dimensão** que retorna o número de elementos de um mapa). Sendo assim, os valores de intersecção apresentados nas Tabelas 4.12 e 4.13 continuam válidos.

Tal como o algoritmo do ponto anterior, o MapaA será comparado com o MapaB e MapaC, desta vez utilizando o coeficiente de *Jaccard*:

$$Similaridade\_por\_Jaccard(A,B,\alpha,4) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{count(A \cap B) + |A - B| + |A - B|}$$
$$= \frac{\alpha^8 + \alpha^7}{3+6+6} = \frac{\alpha^7(\alpha+1)}{15}$$

$$Similaridade\_por\_Jaccard(A,C,\alpha,4) = \frac{|A \cap C|}{|A \cup C|} = \frac{|A \cap C|}{count(A \cap C) + |A - C| + |A - C|}$$
$$= \frac{\alpha^8 + 2\alpha^6 + \alpha^4 + \alpha^3}{5+3+5} \quad \frac{= \alpha^3(\alpha^5 + 2\alpha^3 + \alpha + 1)}{13}$$

$\forall \alpha \in]0,1[, Similaridade\_por\_Jaccard(A,C,\alpha,4) > Similaridade\_por\_Jaccard(A,B,\alpha,4)$

qed

Onde a função **count** determina quantos conceitos foram considerados na intersecção entre dois conjuntos, e a diferença entre conjuntos (|A-B|, |B-A|, |A-C| e |C-A|) é obtida através da verificação de número de conceitos (sem contar com as suas respectivas generalizações) restantes após intersecção (penúltima linha, 3ª e 4ª colunas das Tabelas 4.11 e 4.12). Pelo coeficiente de *Jaccard*, além de comprovarmos, tal como no ponto anterior, que a similaridade entre o MapaA e MapaC é maior, também o fosso entre as duas medidas é muito maior, uma vez que este coeficiente atribui um valor muito menor de similaridade quando temos poucos elementos em comum:

$sim_{Dice}(A,C) - sim_{Dice}(A,B) > sim_{Jaccard}(A,C) - sim_{Jaccard}(A,B)$

- **Coeficiente de Sobreposição (*Overlap*)**

Assim como nos pontos anteriores, iremos aplicar o coeficiente de Sobreposição (ou Inclusão), que segue o algoritmo do Quadro 4.6, sobre os mapas A, B e C. É importante notar que, ao contrário do ponto anterior, não é necessário efectuar nenhuma alteração ao algoritmo de intersecção inicialmente proposto (Quadro 4.2).

$$SIM(x,y) = \frac{|X \cap Y|}{\min(|X|,|Y|)} \qquad \textbf{(3.11)}$$

```
Similaridade_Por_Sobreposição(X, Y, α, L)
  mapa_x[] ← Conceitos(X)
  mapa_y[] ← Conceitos(Y)
  nivel ← 0
  valor_comum ← 0
  para cada c_i ∈ mapa_x[]
  fazer
    c_i.gen[nivel] ← c_i
  fim para
  para cada c_j ∈ mapa_y[]
  fazer
    c_j.gen[nivel] ← c_j
  fim para
  enquanto nivel ≤ L
  fazer
    valor_comum ← valor_comum + Intersecção(mapa_x[], mapa_y[], nivel)
    nivel ← nivel + 1
    para cada c_i ∈ mapa_x[]
    fazer
      c_i.gen[nivel] ← Generalização(c_i.gen[nivel-1])
    fim para
    para cada c_j ∈ mapa_y[]
    fazer
      c_j.gen[nivel] ← Generalização(c_j.gen[nivel-1])
    fim para
  fim enquanto
  similaridade ← valor_comum/min(|X|,|Y|)
  devolve similaridade
```

Quadro 4.6: Algoritmo para o cálculo de similaridade pelo coeficiente de Sobreposição entre dois mapas conceptuais representados, respectivamente, pelos conjuntos de conceitos X e Y.

$$Similaridade\_por\_Sobreposição(A,B,\alpha,4) = \frac{|A \cap B|}{\min(|A|,|B|)} = \frac{\alpha^8 + \alpha^7}{8} = \frac{\alpha^7(\alpha^7+1)}{8}$$

$$Similaridade\_por\_Sobreposição(A,C,\alpha,4) = \frac{|A \cap C|}{\min(|A|,|C|)} = \frac{\alpha^8 + 2\alpha^6 + \alpha^4 + \alpha^3}{8}$$

$$= \frac{\alpha^3(\alpha^5 + 2\alpha^3 + \alpha + 1)}{8}$$

$$Similaridade\_por\_Sobreposição(A,C,\alpha,4) > Similaridade\_por\_Sobreposição(A,B,\alpha,4)$$

$$\forall \alpha \in]0,1[, qed$$

Mais uma vez, a similaridade entre os mapas A e C é relativamente maior que entre A e B. Este coeficiente valoriza muito mais a quantidade de características partilhadas, comparativamente à situação de uma menor semelhança, tal que, em valor absoluto, estes são os maiores valores de similaridade em relação aos outros coeficientes.

- **Coeficiente Cosseno**

O coeficiente Cosseno, Quadro 4.7, aplica a propriedade geométrica da distância euclidiana no cálculo de similaridade entre dois objectos. Este é o coeficiente mais popular para a similaridade

entre textos devido à sua flexibilidade na comparação entre conjuntos de tamanhos muito diferentes. Sendo assim, não é sensível à quantidade de informação, mas sim ao conteúdo da informação. A título de exemplo, este coeficiente também será aplicado aos mapas A, B e C.

$$SIM(x,y) = \frac{|X \cap Y|}{\sqrt{|X| x |Y|}} \quad \textbf{(3.12)}$$

```
Similaridade_Por_Cosseno(X, Y, α, L)
  mapa_x[] ← Conceitos(X)
  mapa_y[] ← Conceitos(Y)
  nivel ← 0
  valor_comum ← 0

  para cada c_i ∈ mapa_x[]
  fazer
    c_i.gen[nivel] ← c_i
  fim para

  para cada c_j ∈ mapa_y[]
  fazer
    c_j.gen[nivel] ← c_j
  fim para

  enquanto nivel ≤ L
  fazer
    valor_comum ← valor_comum + Intersecção(mapa_x[], mapa_y[], nivel)
    nivel ← nivel + 1
    para cada c_i ∈ mapa_x[]
    fazer
      c_i.gen[nivel] ← Generalização(c_i.gen[nivel-1])
    fim para
    para cada c_j ∈ mapa_y[]
    fazer
      c_j.gen[nivel] ← Generalização(c_j.gen[nivel-1])
    fim para
  fim enquanto

  similaridade ← valor_comum/√(|X|*|Y|)

  devolve similaridade
```

Quadro 4.7: Algoritmo para o cálculo de similaridade pelo coeficiente Cosseno entre dois mapas conceptuais representados, respectivamente, pelos conjuntos de conceitos X e Y.

$$Similaridade\_por\_Cosseno(A,B,\alpha,4) = \frac{|A \cap B|}{\sqrt{|A| \times |B|)}} = \frac{\alpha^8 + \alpha^7}{\sqrt{8 \times 8}} = \frac{\alpha^7(1+\alpha)}{8}$$

$$Similaridade\_por\_Cosseno(A,C,\alpha,4) = \frac{|A \cap C|}{\sqrt{|A| \times |C|)}} = \frac{\alpha^8 + 2\alpha^6 + \alpha^4 + \alpha^3}{\sqrt{8 \times 10}}$$

$$= \frac{\alpha^3(\alpha^5 + 2\alpha^3 + \alpha + 1)}{4\sqrt{5}}$$

$$\forall \alpha \in]0,1[, Similaridade\_por\_Cosseno(A,C,\alpha,4) > Similaridade\_por\_Cosseno(A,B,\alpha,4)$$
$$qed$$

Este coeficiente parece ser o mais adaptado à comparação entre mapas conceptuais devido a sua independência do tamanho dos conjuntos, ou seja, mapas de diferentes dimensões mas que apresentem informação semelhante serão beneficiados por esta medida.

A simplicidade da abordagem de contraste de características que considera os mapas conceptuais a comparar como conjuntos de conceitos esconde o facto de alguns destes conceitos poderem ter diferentes graus de importância no contexto apresentado pelo mapa e, consequentemente, terem um maior ou menor peso na comparação. Por seu turno, estes conceitos podem estar interligados, dando forma a uma certa estrutura ou hierarquia que não é valorizada por esta abordagem. Outra desvantagem desta abordagem, devido à sua simplicidade, é o factor de similaridade apresentado considerar equivalentes todas as relações hierárquicas no WordNet. Tal como vimos na Secção 4.3.1, onde foram discutidas diferentes métricas para a quantificação da similaridade entre conceitos, um dos factores a ser observado seria a existência de diferenças no grau de detalhe em que alguns domínios se encontram classificados hierarquicamente (por exemplo, o domínio da biologia encontra-se densamente representado no WordNet, com várias classes e uma grande profundidade taxonómica). Serão abordadas nas próximas subsecções outras abordagens para o cálculo da similaridade entre mapas conceptuais. Como veremos a seguir, estas abordagens irão considerar, além do conhecimento semântico, informação sobre a organização interna do mapa em questão.

3.2.2. Procura Simples pela Maior Similaridade entre Conceitos

Após estudar os métodos anteriores baseados na teoria dos conjuntos, decidimos adaptar um destes utilizando uma medida de similaridade entre conceitos com melhor desempenho[14] para determinar se um conceito é equivalente a outro na operação de intersecção. Deste modo, estudaremos a aplicabilidade e eficácia da medida de similaridade entre conceitos com melhores resultados (a medida adaptada de Jiang e Conrath) associada ao método adaptado de contraste de características que, no nosso entender, é o mais indicado para a comparação entre mapas conceptuais. O coeficiente Cosseno foi escolhido por ser menos sensível à variação de tamanho dos mapas conceptuais.

Neste caso, ainda não é considerada a importância de cada conceito (nó do grafo) no mapa conceptual e, consequentemente, a informação sobre a estrutura interna não influencia este

[14] Como poderemos comprovar no Capítulo 5 (Testes Efectuados).

primeiro algoritmo. Deste modo, todos os conceitos do mapa conceptual são tratados de forma igual.

A nível mais formal, propomos estabelecer um mapeamento entre conceitos de dois mapas conceptuais através de uma função injectiva de similaridade: para cada conceito de um mapa, é encontrado no máximo um equivalente no outro mapa. Na comparação global entre dois mapas, foram escolhidas as correspondências entre conceitos que maximizassem o valor de intersecção (similaridade) entre os mapas. Sendo assim, tal como podemos ver no Quadro 4.8, para cada combinação entre os conceitos dos dois mapas, foram seleccionadas a **maiores** medidas de similaridade para o mapeamento entre conceitos.

```
  Similaridade_Melhores_Pares(X, Y)
    índice  ← 0
    valor_comum ← 0
    para cada c_i ∈ X
    fazer
      para cada c_j ∈ Y
      fazer
(1)      matriz_similaridades[i][j] ← sim_jcn(c_i,c_j)
      fim para
    fim para
(2) num_correspondências ← min(|X|,|Y|)
    definir vector_correspondências[num_correspondências]

    enquanto num_correspondências > 0
    fazer
(3)   maior ← MaiorSimilaridade(matriz_similaridades[][], i, j)
      valor_comum ← valor_comum + maior
(4)   EliminaLinha(i, matriz_similaridades)
      EliminaColuna(j, matriz_similaridades)
      vector_correspondências[índice] ← (c_i,c_j)
      índice ← índice + 1
      num_correspondências ← num_correspondências - 1
    fim enquanto
(5) similaridade ← 1 + ( valor_comum - min(|X|,|Y))/√(|X|*|Y|)

    devolve similaridade
```

Quadro 4.8: Algoritmo para o cálculo de similaridade que considera o melhor mapeamento (maior similaridade entre conceitos) entre dois mapas conceptuais X e Y.). As linhas numeradas serão detalhadamente explicadas a seguir.

(1) Inicialmente, é criada uma matriz de similaridade que guarda as similaridades (baseada em Jiang e Conrath) entre os conceitos de cada mapa.

(2) Já que os mapas poderão ter tamanhos diferentes (sendo o menor de tamanho **N**), existem, no máximo, N correspondências entre os mapas.

(3) Através da função **MaiorSimilaridade**, é determinado o maior valor de similaridade actualmente presente na matriz de similaridades e em que posição este se encontra (linha **i** e coluna **j**).

(4) Uma vez que cada conceito só poderá corresponder a outro conceito, e que cada linha e coluna representam respectivamente os conceitos dos mapas X e Y, de cada vez que é encontrada uma correspondência entre conceitos, estes não serão mais explorados no futuro. A complexidade de processamento é aproximadamente *O(N|X||Y|))*.

(5) Para o cálculo da similaridade final entre dois mapas, foi utilizada uma adaptação do coeficiente Cosseno. A razão desta adaptação prende-se com o facto deste algoritmo ter sido inicialmente implementado utilizando a proposta original de *distância semântica* apresentada por Jiang e Conrath. Assumindo que a distância semântica entre conceitos produz resultados complementares em relação à similaridade entre conceitos, ou seja *sim(x,y) = 1 – dist(x,y)*, podemos comprovar que a similaridade global entre mapas conceptuais corresponde a:

$$SIM(X,Y) = 1 - dist(X,Y) = 1 - \frac{\sum dist(x_u, y_j)}{\sqrt{|X| \times |Y|}} = 1 - \frac{\sum 1 - sim(x_u, y_j)}{\sqrt{|X| \times |Y|}}$$

$$= 1 - \frac{\min(|X|,|Y|) - \sum sim(x_u, y_j)}{\sqrt{|X| \times |Y|}} = 1 - \frac{\min(|X|,|Y|) - |X \cap Y|}{\sqrt{|X| \times |Y|}} = 1 + \frac{|X \cap Y| - \min(|X|,|Y|)}{\sqrt{|X| \times |Y|}}$$

As Tabelas 4.14 e 4.15 representam, respectivamente, as matrizes de similaridades entre o mapaA e o mapaB, e entre o mapaA e o mapaC. Na Tabela 4.14, podemos verificar a evolução do algoritmo através do escurecimento das suas células. De cada vez que foi efectuada uma correspondência (células em *bold*) a linha e coluna desta são eliminadas da matriz. Como resultado do mapeamento, foram estabelecidas as seguintes correspondências entre os conceitos dos dois mapas: {(*bank#2*, *room#1*), (*grass#1*, *book#1*), (*canoe*, *magazine#1*), (*river*, *database*), (*beach*, *library#1*), (*duck#1*, *librarian*), (*sand#1*, *journal#1*), (*crocodile*, *information_system*)}; com a seguinte similaridade final entre os Mapas A e B:

$$Similaridade\_Melhores\_Pares(A,B) = 1 + \frac{|A \cap B| - \min(|A|,|B|)}{\sqrt{|A| \times |B|)}}$$

$$= 1 + \frac{0{,}497 + 0{,}481 + 0{,}410 + 0{,}371 + 0{,}231 + 0{,}168 + 0{,}146 + 0{,}088 - 8}{\sqrt{8 \times 8}} = 1 - \frac{5{,}839}{\sqrt{8 \times 8}} \cong 0{,}270$$

	book#1	database	information_system	journal#1	librarian	library#1	magazine#1	room#1
bank#2	0,495	0,362	0,250	0,225	0,274	0,298	0,357	**0,497**
beach	0,427	0,295	0,182	0,157	0,206	**0,231**	0,279	0,429
canoe	0,470	0,238	0,314	0,111	0,150	0,273	**0,410**	0,472
crocodile	0,333	0,200	**0,088**	0,063	0,112	0,136	0,185	0,335
duck#1	0,389	0,256	0,144	0,119	**0,168**	0,192	0,241	0,391
grass#1	**0,481**	0,348	0,236	0,210	0,260	0,284	0,332	0,483
river	0,465	**0,371**	0,220	0,233	0,245	0,269	0,317	0,467
sand#1	0,379	0,284	0,134	**0,146**	0,158	0,182	0,230	0,381

Tabela 4.14: Matriz de similaridades entre os conceitos dos mapas A e B.

	blue_whale	fish#1	ocean#1	orca	plankton	salmon#1	saltwater #1	seabird	shark#1	whale#2
bank#2	0,250	0,381	0,339	0,250	0,250	0,274	0,250	0,327	**0,274**	0,250
beach	0,182	0,314	0,271	0,182	**0,182**	0,206	0,183	0,259	0,206	0,182
canoe	0,126	0,257	0,215	0,126	0,126	0,150	0,127	0,203	0,150	**0,126**
crocodile	0,405	**0,536**	0,177	0,405	0,088	0,429	0,088	0,482	0,429	0,405
duck#1	0,461	0,592	0,233	0,461	0,145	0,485	0,144	**0,872**	0,485	0,461
grass#1	0,236	0,367	0,325	0,236	0,236	**0,260**	0,236	0,312	0,260	0,236
river	0,220	0,352	**0,816**	0,220	0,220	0,245	0,259	0,298	0,245	0,220
sand#1	0,134	0,265	0,261	0,134	0,134	0,158	**0,385**	0,210	0,158	0,134

Tabela 4.15: Matriz de similaridades entre os conceitos dos mapas A e C.

Já neste segundo exemplo de comparação entre mapas, foi encontrado o seguinte conjunto de correspondências entre conceitos: {(*duck#1*, *seabird*), (*river*, *ocean#1*), (*crocodile*, *fish#1*), (*sand#1*, *saltwater#1*), (*bank#2*, *shark#1*), (*grass#1*, *salmon#1*), (*beach*, *plankton*), (*crocodile*, *whale#2*)}. É importante salientar que, em alguns casos onde a similaridade entre pares de conceitos era a mesma (por exemplo os pares: (*bank#2, salmon#1*) e (*bank#2*, *shark#1*); (*beach, blue_whale*), (*beach, orca*), (*beach, plankton*) e (*beach, whale#2*); (*canoe, blue_whale*), (*canoe, orca*) e (*canoe, whale#2*); possuem o mesmo valor de similaridade), foi seleccionado o

par formado pelos conceitos mais gerais, ou seja, aqueles que se encontram a um menor nível de profundidade do topo (por exemplo, o conceito *shark#1* é mais genérico que *salmon#1* já que este último se encontra a mais de dois níveis de profundidade em relação ao conceito generalizador, **gcme**, entre os dois). A similaridade final calculada entre os Mapas A e C será:

$$Similaridade\_Melhores\_Pares(A,C) = 1 + \frac{|A \cap C| - \min(|A|,|C|)}{\sqrt{|A| \times |C|)}}$$

$$= 1 + \frac{0{,}872 + 0{,}816 + 0{,}536 + 0{,}385 + 0{,}274 + 0{,}260 + 0{,}182 + 0{,}126 - 8}{\sqrt{|A| \times |C|)}} = 1 - \frac{4{,}549}{\sqrt{8 \times 10}} \cong 0{,}491$$

Deste modo, podemos concluir que, tal como se esperava, por esta medida confirmamos o raciocínio intuitivo do MapaA ser semanticamente mais próximo do MapaC em relação ao MapaB. Na próxima Sub Secção, abordaremos uma variante desta medida de similaridade entre mapas conceptuais que considera também alguma informação sobre a organização interna dos mapas.

3.2.3. Procura Ponderada pela Maior Similaridade entre Conceitos

Como todos os conceitos são tratados com igual importância para o cálculo da similaridade final, será sensato pensar que deveríamos dar mais relevância aos conceitos mais centrais (e consequentemente com maior grau de ligação) do mapa e só depois verificar a similaridade entre os conceitos mais periféricos a comparar? Qual seria a influência desta informação estrutural sobre a comparação semântica entre mapas conceptuais?

Face a estas questões, decidimos introduzir uma ligeira alteração do algoritmo anterior (ver Quadro 4.9) para efectuar um mapeamento, em primeiro lugar, entre os conceitos mais importantes (por terem um maior peso, ou grau de ligação) de cada mapa conceptual. Para determinar a importância de um conceito num dado mapa, consideramos dois factores de relevância: o grau de ligação (grau de entrada somado ao grau de saída do nó) e probabilidade de ocorrência do conceito num *corpus* como factor de desempate entre dois conceitos de igual peso (ou seja, é considerado, em primeiro lugar, o conceito mais utilizado). Antes de calcularmos e criarmos a matriz de similaridade entre os dois mapas, cada conjunto de conceitos é ordenado através da função **OrdenaPorPeso()** segundo os dois factores já anteriormente descritos. O *SemCor* é o *corpus* utilizado para o cálculo da probabilidade, tal como podemos observar no exemplo de ordenação aplicada para um conjunto de conceitos (ver Tabela 4.16).

```
Similaridade_Melhores_Pares_Por_Peso(X, Y)
  índice  ← 0
  valor_comum_X ← 0
  valor_comum_Y ← 0
  X_ordenado ← OrdenaPorPeso(X)
  Y_ordenado ← OrdenaPorPeso(Y)
  para cada x_i ∈ X_ordenado
  fazer
       para cada y_j ∈ Y_ordenado
    fazer
      matriz_similaridades_X[i][j] ← sim_jcn(x_i,y_j)
          matriz_similaridades_Y[i][j] ← sim_jcn(x_i,y_j)
      fim para
  fim para
  num_correspondências ← min(|X|,|Y|)
  definir vector_correspondências_X[num_correspondências]
  definir vector_correspondências_Y[num_correspondências]
  para x_i ∈ X_ordenado
  fazer
    se num_correspondências = 0
    então
      sair ciclo
    fim se
    maior ← MaiorSimilaridadePorPeso(matriz_similaridades_X[i][], j)
    valor_comum_X ← valor_comum_X + maior
    EliminaLinha(i, matriz_similaridades_X)
    EliminaColuna(j, matriz_similaridades_X)
    vector_correspondências_X[índice] ← (x_i,y_j)
    índice ← índice + 1
    num_correspondências ← num_correspondências – 1
  fim para
  índice ← índice - 1
  para y_j ∈ Y_ordenado
  fazer
    se índice < 0
    então
      sair ciclo
    fim se
    maior ← MaiorSimilaridadePorPeso(matriz_similaridades_Y[][j], i)
    valor_comum_Y ← valor_comum_Y + maior
    EliminaLinha(i, matriz_similaridades_Y)
    EliminaColuna(j, matriz_similaridades_Y)
    vector_correspondências_Y[índice] ← (x_i,y_j)
    índice ← índice - 1
  fim para
  similaridade_X ← 1 + ( valor_comum_X - min(|X|,|Y))/√(|X|*|Y|)
  similaridade_Y ← 1 + ( valor_comum_Y - min(|X|,|Y))/√(|X|*|Y|)
  similaridade ← max(similaridade_X,similaridade_Y)
  devolve similaridade
```

Quadro 4.9: Algoritmo para o cálculo de similaridade que considera o melhor mapeamento (maior similaridade entre conceitos) observando a sua importância em cada mapa.

Cada mapa é analisado como sendo o objecto central de comparação. Assim, a similaridade total obtida será maximizada respeitando a ordem pela qual os conceitos foram ordenados. Ou seja, a procura é feita em paralelo tendo como ponto de partida o outro mapa, e o

objectivo é encontrar o melhor mapeamento que maximize a similaridade total calculada desde o conceito central até ao mais periférico. Deste modo, a função **MaiorSimilaridadePorPeso()** devolve o maior valor, desta vez, num vector de similaridades identificando o elemento seleccionado do vector. Finalmente, é considerada a maior similaridade, de entre as análises parciais dos dois mapas, como o valor final de comparação semântica entre os dois mapas conceptuais. Apesar de parecer mais complexo, o tempo de processamento deste algoritmo, tal como no algoritmo anterior, é de ordem *O(N|X ||*Y|*)*, onde N é a dimensão, em número de conceitos, do menor mapa na comparação entre X e Y.

MapaA			MapaB			MapaC		
Conceito	**Peso**	**P(Conceito)**	**Conceito**	**Peso**	**P(Conceito)**	**Conceito**	**Peso**	**P(Conceito)**
bank#2	7	6,3e-5	**library#1**	6	2,42e-6	**fish#1**	4	2,61e-5
river	3	8,00e-5	**database**	5	1,87e-4	**ocean#1**	4	2,30e-5
duck#1	3	3,03e-6	**book#1**	2	6,77e-4	**salmon#1**	4	1,21e-6
crocodile	3	6,06e-7	**magazine#1**	2	9,70e-6	**shark#1**	3	1,21e-6
grass#1	2	4,18e-5	**journal#1**	2	3,64e-6	**plankton**	3	6,06e-7
canoe	2	1,82e-6	**librarian**	2	1,21e-6	**whale#2**	3	6,06e-7
beach	1	9,09e-6	**information_system**	2	6,06e-7	**seabird**	2	5,45e-6
sand#1	1	6,67e-6	**room#1**	1	7,14e-4	**blue_whale**	2	6,06e-7
						orca	2	6,06e-7
						saltwater	1	1,82e-6

Tabela 4.16: Listas de conceitos dos mapas A, B e C, ordenadas pelos respectivos pesos e probabilidades.

Ao aplicarmos este algoritmo, verificamos que o mapeamento resultante entre os conceitos dos mapas A e B (ver Tabela 4.17) respeitando o ordenamento dos conceitos do mapaA corresponde a: {(*bank#2*, *room#1*), (*river*, *book#1*), (*duck#1*, *database*), (*crocodile*, *magazine#1*), (*grass#1*, *library#1*), (*canoe*, *information_system*), (*beach*, *librarian*), (*sand#1*, *journal#1*)}. A similaridade parcial em ordem ao mapaA é:

$$Similaridade\_Melhores\_por\_Peso_A(A,B) = 1 + \frac{|A \cap B|_A - \min(|A|,|B|)}{\sqrt{|A| \times |B|)}}$$

$$= 1 + \frac{0{,}497 + 0{,}465 + 0{,}256 + 0{,}185 + 0{,}284 + 0{,}314 + 0{,}206 + 0{,}146 - 8}{\sqrt{8 \times 8}} = 1 - \frac{5{,}647}{\sqrt{8 \times 8}} \cong 0{,}294$$

Se o mapaB for analisado em primeiro lugar na escolha do melhor mapeamento em relação ao mapaA (Tabela 4.18) surgem as seguintes correspondências entre conceitos: {(*bank#2*, *library#1*), (*river*, *database*), (*grass#1*, *book#1*), (*canoe*, *magazine#1*), (*beach*, *journal#1*), (*duck#1*, *librarian*), (*sand#1*, *information_system*), (*crocodile*, *room#1*)}.

A similaridade em ordem ao MapaB é a seguinte:

$$Similaridade\_Melhores\_por\_Peso_B(A,B) = 1 + \frac{|A \cap B|_B - \min(|A|,|B|)}{\sqrt{|A| \times |B|)}}$$

$$= 1 + \frac{0{,}298 + 0{,}371 + 0{,}481 + 0{,}410 + 0{,}157 + 0{,}168 + 0{,}134 + 0{,}335 - 8}{\sqrt{8 \times 8}} = 1 - \frac{5{,}646}{\sqrt{8 \times 8}} \cong 0{,}294$$

$$Similaridade\_Melhores\_por\_Peso(A,B) = \max(Sim_A(A,B), Sim_B(A,B)) = 0{,}294$$

	library#1	database	book#1	magazine#1	journal#1	librarian	information_system	room#1
bank#2	0,298	0,362	0,495	0,357	0,225	0,274	0,250	**0,497**
river	0,269	0,371	**0,465**	0,317	0,233	0,245	0,220	0,467
duck#1	0,192	**0,256**	0,389	0,241	0,119	0,168	0,144	0,391
crocodile	0,136	0,200	0,333	**0,185**	0,063	0,112	0,088	0,335
grass#1	**0,284**	0,348	0,481	0,332	0,210	0,260	0,236	0,483
canoe	0,273	0,238	0,470	0,410	0,111	0,150	**0,314**	0,472
beach	0,231	0,295	0,427	0,279	0,157	**0,206**	0,182	0,429
sand#1	0,182	0,284	0,379	0,230	**0,146**	0,158	0,134	0,381

Tabela 4.17: Matriz ordenada de similaridades entre os conceitos dos mapas A e B. As correspondências foram seleccionadas observando o ordenamento dos conceitos do mapaA.

	library#1	database	book#1	magazine#1	journal#1	librarian	information_system	room#1
bank#2	**0,298**	0,362	0,495	0,357	0,225	0,274	0,250	0,497
river	0,269	**0,371**	0,465	0,317	0,233	0,245	0,220	0,467
duck#1	0,192	0,256	0,389	0,241	0,119	**0,168**	0,144	0,391
crocodile	0,136	0,200	0,333	0,185	0,063	0,112	0,088	**0,335**
grass#1	0,284	0,348	**0,481**	0,332	0,210	0,260	0,236	0,483
canoe	0,273	0,238	0,470	**0,410**	0,111	0,150	0,314	0,472
beach	0,231	0,295	0,427	0,279	**0,157**	0,206	0,182	0,429
sand#1	0,182	0,284	0,379	0,230	0,146	0,158	**0,134**	0,381

Tabela 4.18: Matriz ordenada de similaridades entre os conceitos dos mapas A e B. As correspondências foram seleccionadas observando o ordenamento dos conceitos do mapaB.

Podemos constatar que, apesar de não haver o mesmo mapeamento entre as similaridades parciais a comparar os mapas A e B, os valores resultantes são muito próximos. Isto vem

confirmar o consenso em torno da similaridade atribuída por este algoritmo. O mesmo se passa quando comparamos os mapas A e C (Tabelas 4.19 e 4.20).

$$Similaridade\_Melhores\_por\_Peso_A(A,C) = 1 + \frac{|A \cap C|_A - \min(|A|,|C|)}{\sqrt{|A| \times |C|)}}$$

$$= 1 + \frac{0{,}381 + 0{,}816 + 0{,}485 + 0{,}482 + 0{,}260 + 0{,}127 + 0{,}182 + 0{,}134 - 8}{\sqrt{8 \times 10}} = 1 - \frac{5{,}133}{\sqrt{8 \times 10}} \cong 0{,}426$$

$$Similaridade\_Melhores\_por\_Peso_C(A,C) = 1 + \frac{|A \cap C|_C - \min(|A|,|C|)}{\sqrt{|A| \times |C|)}}$$

$$= 1 + \frac{0{,}592 + 0{,}816 + 0{,}429 + 0{,}274 + 0{,}236 + 0{,}182 + 0{,}203 + 0{,}134 - 8}{\sqrt{8 \times 10}} = 1 - \frac{5{,}134}{\sqrt{8 \times 10}} \cong 0{,}426$$

$$Similaridade\_Melhores\_por\_Peso(A,C) = \max(Sim_A(A,C), Sim_c(A,C)) = 0{,}426$$

Apesar de este método confirmar, tal como os algoritmos anteriores, a maior proximidade semântica entre os mapas A e C, em detrimento do mapaB, existe agora uma menor diferenciação entre os dois valores de similaridade global entre as duas comparações. Este comportamento pode estar ligado ao facto de, apesar de termos em conta a organização interna de cada mapa a comparar, talvez ainda falte diferenciar a contribuição de cada conceito no cálculo final de similaridade. Na próxima Sub Secção será apresentada uma última proposta para o cálculo de similaridade baseada também em alguma informação heurística.

	fish#1	**ocean#1**	**salmon#1**	**shark#1**	**plankton**	**whale#2**	**seabird**	**blue_whale**	**orca**	**saltwater #1**
bank#2	**0,381**	0,339	0,274	0,274	0,250	0,250	0,327	0, 250	0,250	0, 250
river	0,352	**0,816**	0,245	0, 245	0,220	0,220	0,298	0, 220	0,220	0,259
duck#1	0,592	0,233	**0,485**	0,485	0,145	0,461	0,872	0,461	0,461	0,144
crocodile	0,536	0,177	0,429	0,429	0,088	0,405	**0,482**	0,405	0,405	0,088
grass#1	0, 367	0,325	0, 260	**0,260**	0,236	0,236	0,312	0, 236	0,236	0,236
canoe	0,257	0,215	0,150	0,150	0,126	0,126	0,203	0,126	0,126	**0,127**
beach	0,314	0,271	0, 206	0, 206	**0,182**	0,182	0,259	0,182	0,182	0,183
sand#1	0,265	0,261	0,158	0,158	0,134	**0, 134**	0,210	0,134	0,134	0,385

Tabela 4.19: Matriz ordenada de similaridades entre os conceitos dos mapas A e C. As correspondências foram seleccionadas observando o ordenamento dos conceitos do mapaA.

	fish#1	ocean#1	salmon#1	shark#1	plankton	whale#2	seabird	blue_whale	orca	saltwater #1
bank#2	0,381	0,339	0,274	**0,274**	0,250	0,250	0,327	0, 250	0,250	0, 250
river	0,352	**0,816**	0,245	0, 245	0,220	0,220	0,298	0, 220	0,220	0,259
duck#1	**0,592**	0,233	0,485	0,485	0,145	0,461	0,872	0,461	0,461	0,144
crocodile	0,536	0,177	**0,429**	0,429	0,088	0,405	0,482	0,405	0,405	0,088
grass#1	0, 367	0,325	0, 260	0,260	**0,236**	0,236	0,312	0, 236	0,236	0,236
canoe	0,257	0,215	0,150	0,150	0,126	0,126	**0,203**	0,126	0,126	0,127
beach	0,314	0,271	0, 206	0, 206	0,182	**0,182**	0,259	0,182	0,182	0,183
sand#1	0,265	0,261	0,158	0,158	0,134	0, 134	0,210	**0,134**	0,134	0,385

Tabela 4.20: Matriz ordenada de similaridades entre os conceitos dos mapas A e C. As correspondências foram seleccionadas observando o ordenamento dos conceitos do mapaC.

3.2.4. Conceito Central e Média Ponderada das Similaridades entre Conceitos

Ao descrever um dado domínio, um mapa conceptual pode conter conceitos centrais com uma forte ligação aos outros conceitos presentes no mapa, e outros mais periféricos sem uma grande relevância para uma maior compreensão sobre o assunto retratado. Na comparação e mapeamento entre dois mapas conceptuais, a contribuição de cada par de conceitos correspondentes poderia ser proporcional ao peso que cada um representa no seu domínio. Esta última medida (fórmula 4.4) pretende, de uma forma simples e com pouco tempo de processamento, estabelecer uma comparação semântica entre mapas conceptuais que, além de se basear na importância de cada conceito, contabiliza o seu peso no cálculo final de similaridade.

O raciocínio empregue fundamenta-se na hipótese de existir um conceito suficientemente representativo que consiga por si só transmitir a quase totalidade da informação apresentada pelo mapa. Para deduzir tal conceito, consideramos, assim como no algoritmo anterior, o grau de ligação e probabilidade de ocorrência do conceito no *corpus SemCor*. Deste modo, o **conceito** mais **central** de uma mapa poderá ser o **conceito que possua o maior grau de ligação no mapa** e, caso exista mais do que um, é escolhido o que é mais utilizado no quotidiano.

$$Sim.C.Central(X,Y) = \max\left(\frac{\sum_{i=1}^{m} sim_{jcn}(x_i, centro_Y) \times peso_i}{m}, \frac{\sum_{j=1}^{n} sim_{jcn}(centro_X, y_j) \times peso_i}{n}\right) \quad \textbf{(4.4)}$$

Onde **n** e **m** são, respectivamente, o número total de conceitos dos mapas X e Y; **centro$_i$** é o nó de um mapa **i** que possua mais arcos incidentes ou dissidentes (conceito central do mapa) e **peso$_j$**

é calculado como a razão entre o grau de ligação do conceito **j** e o grau do conceito central do mapa.

Caso a hipótese de existência de um conceito central seja válida, será correcto analisar a similaridade não através de um mapeamento, mas sim como uma proximidade (ou o seu inverso, um desvio) a este conceito central. Desta forma, a similaridade global entre dois mapas conceptuais poderá ser comparada a uma análise sucinta aos conceitos mais importantes de cada mapa.

A título de exemplo, de acordo com os critérios descritos acima e com base nos dados apresentados pela Tabela 4.16, os conceitos centrais dos mapas A, B e C são, respectivamente, *bank#2*, *library#1* e *fish#1*. Quando aplicamos esta medida à comparação entre os mapas A e B, sob diferentes perspectivas temos o seguinte resultado:

$$Sim.C.Central(A,B) = \max\left(\frac{\sum_{i=1}^{m} sim(a_i, centro_B) \times Peso_i}{m}, \frac{\sum_{j=1}^{n} sim(centro_A, b_j) \times Peso_{ji}}{n}\right) =$$

$$= \max\left(\frac{\sum_{i=1}^{8} sim(a_i, library\#1) \times Peso_i}{8}, \frac{\sum_{j=1}^{8} sim(bank\#2, b_j) \times Peso_{ji}}{8}\right)$$

$$\sum_{i=1}^{8} sim(a_i, library\#1) \times Peso_i = sim(bank\#2, library\#1) + \left[sim(river, library\#1) + sim(duck\#1, library\#1) + sim(crocodile, library\#1)\right] \times \frac{3}{7} + \left[sim(grass\#1, library\#1) + sim(canoe, library\#1)\right] \times \frac{2}{7} + \left[sim(beach, library\#1) + sim(sand\#1, library\#1)\right] \times \frac{1}{7} \cong 0{,}152$$

$$\sum_{j=1}^{8} sim(bank\#2, b_j) \times Peso_{ji} = sim(bank\#2, library\#1) + sim(bank\#2, database) \times \frac{5}{6} + \left[sim(bank\#2, book\#1) + sim(bank\#2, magazine\#1) + sim(bank\#2, journal\#1) + sim(bank\#2, librarian) + sim(bank\#2, \inf ormation\_system)\right] \times \frac{2}{6} + sim(bank\#2, room\#1) \times \frac{1}{6} \cong 0{,}097$$

$$\therefore Sim.C.Central(A,B) = 0{,}152$$

Já em relação ao mapa conceptual C, existem três candidatos a serem considerados o conceito central por possuírem o mesmo grau de ligação: *fish#1*, *ocean#1* e *salmon#1*. Como segundo critério de selecção, constatou-se que o conceito *fish#1* é o mais utilizado entre os três, sendo este o escolhido como conceito mais representativo do mapa conceptual C. Deste modo, a similaridade entre os mapas A e C será:

$$Sim.C.Central(A,C) = \max\left(\frac{\sum_{i=1}^{m} sim(a_i, centro_C) \times Peso_i}{m}, \frac{\sum_{j=1}^{n} sim(centro_A, c_j) \times Peso_{ji}}{n}\right) =$$

$$= \max\left(\frac{\sum_{i=1}^{8} sim(a_i,\ fish\#1) \times Peso_i}{8}, \frac{\sum_{j=1}^{10} sim(bank\#2, c_j) \times Peso_{ji}}{10}\right)$$

$$\sum_{i=1}^{8} sim(a_i,\ fish\#1) \times Peso_i = sim(bank\#2,\ fish\#1) + \left[sim(river,\ fish\#1) + sim(duck\#1,\ fish\#1) + \right.$$

$$\left. sim(crocodile,\ fish\#1)\right] \times \frac{3}{7} + \left[sim(grass\#1,\ fish\#1) + sim(canoe,\ fish\#1)\right] \times \frac{2}{7} + \left[sim(beach,\ fish\#1) + \right.$$

$$\left. sim(sand\#1,\ fish\#1)\right] \times \frac{1}{7} \cong 0{,}205$$

$$\sum_{j=1}^{10} sim(bank\#2, c_j) \times Peso_{ji} = sim(bank\#2,\ fish\#1) + sim(bank\#2, ocean\#1) + sim(bank\#2, salmon\#1 + \right.$$

$$\left[sim(bank\#2, shark\#1) + sim(bank\#2,\ plankton) + sim(bank\#2, whale\#2)\right] \times \frac{3}{4} + \left[sim(bank\#2, seabird) + \right.$$

$$\left. sim(bank\#2, blue\_wahle) + sim(bank\#2, orca)\right] \times \frac{2}{4} + sim(bank\#2, saltwater\#1) \times \frac{1}{4} \cong 0{,}345$$

$$\therefore Sim..C.Central(A,C) = 0{,}345$$

Apesar de ser um algoritmo relativamente simples e económico ($O(|X|+|Y|)$, vem confirmar a proximidade entre os mapas conceptuais de exemplo A e C relativamente ao mapa B. No Capítulo 5, analisaremos um conjunto de testes empíricos que permitirá evidenciar o desempenho destas medidas sobre um alargado conjunto de mapas conceptuais.

5. Testes e Resultados

Os métodos de contextualização e similaridade semântica aplicados aos mapas conceptuais (discutidos no Capítulo 4) serão testados sobre um conjunto de 14 mapas extraídos da Internet [Leung, 2003] e 11 mapas recolhidos de diferentes colaboradores com formação em áreas diversas[1]. O primeiro conjunto de mapas refere-se à apresentação de domínios e assuntos científicos que vão desde a química (como um mapa conceptual que apresenta o ciclo do carbono) até à medicina (em que um outro mostra a anatomia do sistema nervoso central).

O objectivo desta bateria de testes é obter alguma forma de validação do resultado das abordagens aqui apresentadas face ao raciocínio de indivíduos que desempenharam a mesma tarefa.

1. Contextualização

Para o módulo de contextualização, cada mapa conceptual (dos 25 acima referidos) foi desambiguado manualmente por 3 juízes humanos, sendo o resultado final a intersecção entre as suas escolhas. Uma vez que as medidas de comparação entre mapas utilizam sobretudo o conhecimento semântico implícito nos nós dos mapas, esta etapa tem como objectivo descobrir o real significado de cada conceito presente no mapa, exceptuando as relações. Para cada mapa foi elaborado automaticamente um questionário com recurso ao WordNet 1.7.1 onde cada conceito é referido por diferentes significados. Na Figura 5.1 é apresentada a precisão de acerto do algoritmo de contextualização utilizando diferentes métricas de similaridades entre conceitos.

As 4 medidas de similaridades entre conceitos, discutidas na Secção 4.3.1 foram utilizadas no módulo de desambiguação e os conceitos escolhidos confrontados com os seleccionados manualmente por juízes humanos. Estas medidas foram implementadas com base numa *package* de *software* disponibilizada livremente pelo seu autor para fins científicos [Pedersen, 2003].

De forma a verificar a necessidade ou não de incluir uma medida de similaridade no processo de desambiguação, uma 5ª abordagem foi incluída na bateria de testes que consiste simplesmente em seleccionar o significado mais utilizado de uma palavra (*most-used-sense*) com base no corpus *SemCor*. Uma vez que uma palavra pode exprimir diferentes significados e até diferentes classes gramaticais dependendo do contexto onde está inserida, é seleccionado o seu significado mais utilizado dentre todas as classes gramaticais a que pode pertencer. Por exemplo,

[1] Estes mapas estão incluídos na documentação em anexo.

o conceito *variable* é utilizado mais frequentemente como um adjectivo (*liable to or capable of change*).

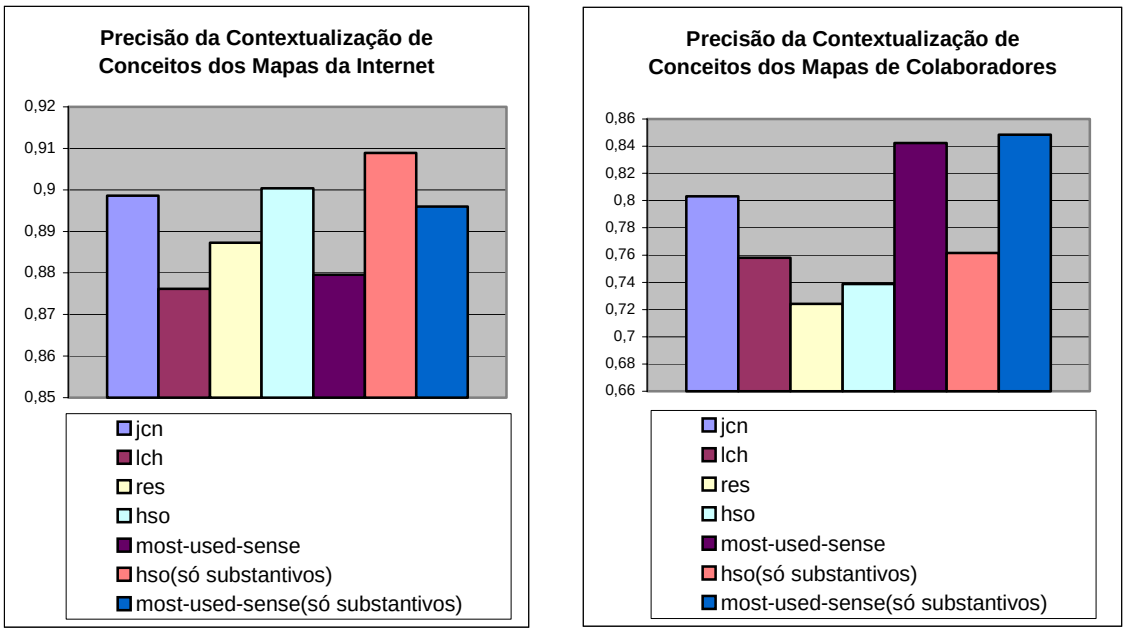


Figura 5.1: Desempenho das medidas de similaridade entre conceitos no processo de desambiguação sobre dois conjuntos distintos de mapas conceptuais (recolhidos da Internet e de colaboradores, respectivamente).

Como nem todas as medidas de similaridade consideram todas as classes gramaticais (como é o caso da medida proposta por Jiang e Conrath – *jcn*, Leacock e Chodorow – *lch*, Resnik – *res*), decidimos também verificar o desempenho da contextualização ao considerar apenas os substantivos no processo de desambiguação utilizando as outras duas abordagens mais abrangentes (medida de Hirst e St-Onge – *hso*, e significado mais utilizado – *most used sense*).

Através dos resultados apresentados na Figura 5.1, verificamos que o conteúdo dos mapas conceptuais influencia o desempenho das abordagens utilizadas, uma vez que os conceitos podem ser mais ou menos ambíguos. Ou seja, em mapas cujo domínio apresentado é mais preciso e detalhado, como é o caso dos mapas recolhidos da Internet, geralmente os métodos automáticos são mais fiáveis, visto que cada conceito apresenta poucas alternativas na escolha do significado correcto. Por seu lado, em mapas onde retratamos assuntos menos objectivos e utilizamos palavras com um grande grau de ambiguidade, o resultado obviamente é menos encorajador.

Outro facto a referir é o de que, quando consideramos apenas os substantivos nas medidas Hirst e St-Onge e na abordagem do significado mais utilizado, conseguimos uma maior percentagem de acerto. Isto deve-se à própria organização dos mapas, onde é maior a utilização de nomes para os conceitos expressos nos nós dos mapas conceptuais, em detrimento de outras classes gramaticais.

Dentre as medidas de similaridade entre conceitos utilizadas na desambiguação (considerando a escolha do significado mais utilizado apenas uma heurística e não uma medida de similaridade), a que tem melhor desempenho em ambos os conjuntos é a medida adaptada de Jiang e Conrath (precisão global de 0,857 ± 0,043 com um grau de confiança de 95%).

Globalmente, a abordagem com maior precisão na desambiguação de mapas conceptuais é a heurística de escolha do significado mais utilizado considerando apenas os substantivos como possíveis candidatos (precisão global de 0,875 ± 0,040 com um grau de confiança de 95%). Esta abordagem (assim como a segunda abordagem mais bem classificada – a de Jiang e Conrath) é fortemente dependente da abrangência do *corpus* utilizado como base de cálculo para a probabilidade de utilização, o que atribui a este recurso lexical uma grande responsabilidade no desempenho final do sistema.

2. Comparação Semântica

A fase decisiva de comparação de mapas conceptuais foi testada seguindo o mesmo raciocínio do módulo anterior. Uma vez que o objectivo deste trabalho não é estabelecer uma medida absoluta de similaridade já que esta pode variar consoante o número de elementos envolvidos na comparação, foi pedido a um conjunto de juízes humanos, desta vez, que ordenassem os mapas mais próximos semanticamente em relação a um mapa-alvo dentro de um dado conjunto de mapas conceptuais. Os mesmos dois conjuntos de mapas utilizados para teste de desambiguação também foram utilizados para testar a precisão da comparação semântica.

Na Secção 4.3.2, foram discutidos diferentes métodos de cálculo de similaridade entre mapas conceptuais. Para cada comparação de dois mapas foi obtido um valor que exprimia a similaridade semântica entre os conceitos de ambos. De forma a validar todo o processo, através destes valores de similaridades, foi elaborada uma lista dos mapas mais próximos quanto ao significado. Tendo os dois ordenamentos em mão (feito pelos juízes e os obtidos automaticamente) surgiu-nos a seguinte questão: como comprovar a eficácia ou não dos métodos propostos? Tentámos encontrar uma forma qualitativa de analisar os resultados, de modo a perceber o verdadeiro contributo ou eventuais falhas que o sistema oferecia. Deste modo, chegámos à conclusão de que era necessário avaliar principalmente dois aspectos: a similaridade semântica relativa e a similaridade semântica global entre mapas conceptuais. A primeira refere-se ao facto de conseguir determinar se um mapa é mais próximo a um outro em relação a um terceiro; a segunda correspondência tem a ver com a possibilidade de identificar o maior número possível de mapas semelhantes. Constatamos também que a partir do 5º mapa (num universo de 11 a 14 mapas) na lista de ordenamento, a similaridade já não era tão óbvia e começava a ser

difícil distinguir os que eram menos dos que eram totalmente dissimilares (uma vez que avaliar a dissimilaridade é muito mais subjectiva e metafórica que a própria semelhança, não havendo aqui um oposto directo do conceito). Deste modo, consideramos apenas os 5 mapas mais próximos para efeitos de validação do sistema face aos 5 mapas mais próximos escolhidos manualmente para cada mapa conceptual.

Nas próximas secções serão apresentados os diversos testes efectuados às abordagens propostas.

2.1. Contraste de Características

Após a apresentação do método para comparação semântica por contraste de características, em que foram discutidos 4 conhecidos coeficientes de similaridades formulados a partir da teoria dos conjuntos (Cosseno, Dice, Jaccard e de Sobreposição), foi proposta uma adaptação aos mapas conceptuais que se baseia no factor de similaridade entre dois conceitos pertencentes a diferentes mapas conceptuais. Este factor é a similaridade entre conceitos de acordo coma hierarquia de substantivos WordNet. A profundidade de pesquisa é também configurável de forma a permitir estudar qual a melhor relação entre esta e o desempenho final da medida. O próprio valor multiplicativo associado a cada generalização na árvore hierárquica à procura de correspondência pode ser ajustado no intervalo [0, 1[. Como foi comprovado anteriormente, este valor não influencia directamente no resultado relativo de comparação mas apenas no valor absoluto final de similaridade obtido (quanto maior é o valor multiplicativo, mais alto é o valor final de similaridade entre dois mapas).

Na Figura 5.2, é apresentada a precisão de acerto da similaridade relativa entre os mapas conceptuais ao longo de diferentes coeficientes de similaridade. Em cada conjunto de mapas, recolhidos da Internet e dos colaboradores, respectivamente, foram criados arranjos de 3 mapas a fim de determinar em cada tuplo qual o mapa mais próximo a um dado alvo.

De acordo com os resultados, nota-se um decréscimo global de performance quanto ao segundo conjunto de mapas (elaborado por colaboradores). Este valor parece ser uma consequência directa do processo de desambiguação, uma vez que, havendo alguma imprecisão no estabelecimento correcto dos significados, esta é transportada para comparação entre mapas. Os coeficientes de Jaccard e Cosseno são os que mais se sobressaem na definição de qual mapa é o mais próximo de um mapa-alvo em relação a um terceiro (precisão média de $0{,}871 \pm 0{,}018$ e $0{,}864 \pm 0{,}019$, respectivamente, com um grau de confiança de 95%).

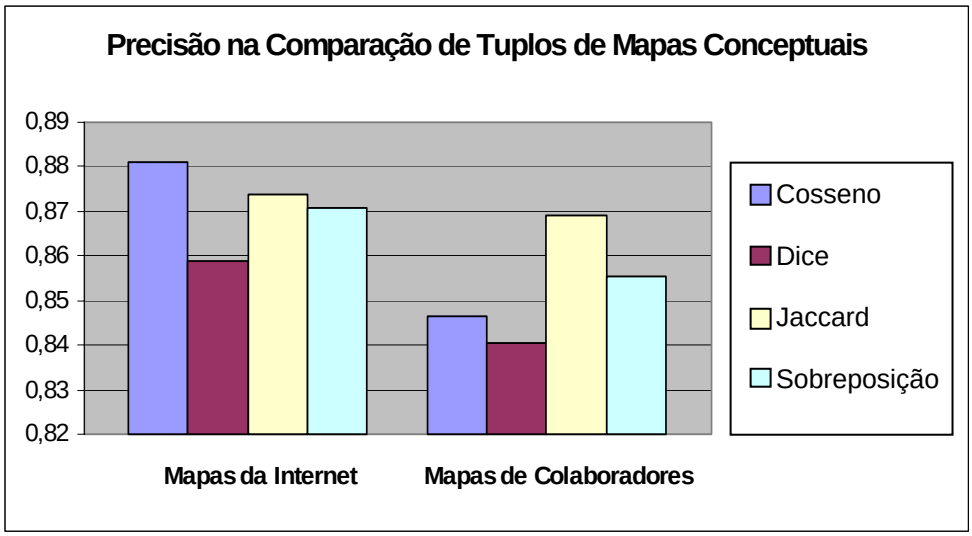


Figura 5.2: Desempenho dos diferentes coeficientes no contraste de característica sobre arranjos de 3 mapas conceptuais.

Na Figura 5.3, são analisados os mesmos coeficientes quanto à identificação dos mesmos mapas escolhidos pelos juízes como sendo os 5 mapas mais próximos a um dado mapa-alvo (similaridade global). Neste caso, o que interessa saber para cada mapa conceptual é o número de mapas considerados mais próximos que estão entre aqueles eleitos manualmente. Tal como na similaridade relativa, os coeficientes de Jaccard e Cosseno apresentam uma maior precisão de acerto face aos outros dois restantes coeficientes (precisão média de 0,669 ± 0,024 e 0,661 ± 0,018, respectivamente, com um grau de confiança de 95%).

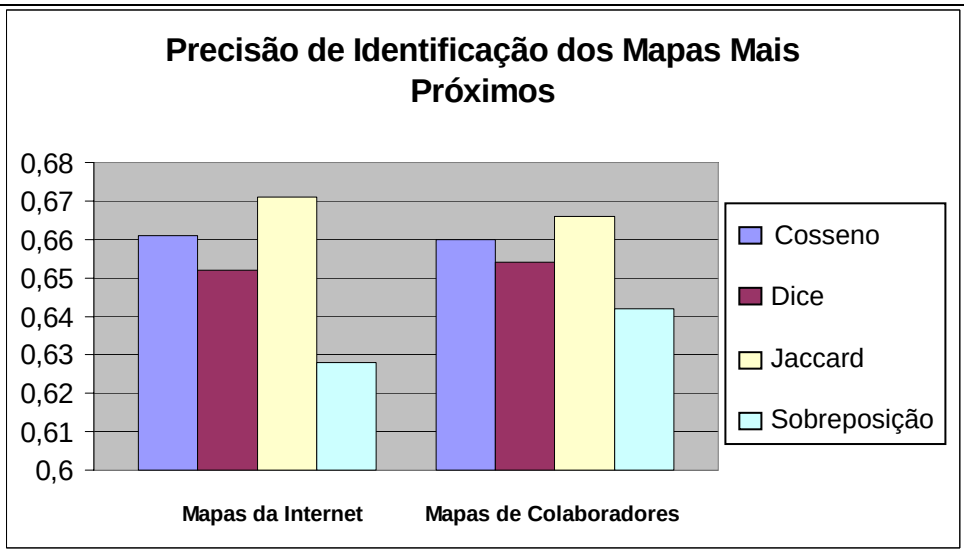


Figura 5.3: Desempenho dos diferentes coeficientes no contraste de característica em identificar os 5 mapas mais próximos para cada mapa-alvo.

Quanto à profundidade na pesquisa do factor de similaridade entre os conceitos dos mapas, foram testados diferentes níveis: desde a profundidade 0, onde não é feita qualquer pesquisa na árvore hierárquica do WordNet, sendo apenas considerados os conceitos que fossem

exactamente iguais entre dois mapas a comparar; até ao factor 4, onde é feita uma pesquisa até à 4ª generalização em busca de uma correspondência entre conceitos. É de notar que, com um nível de profundidade 0, existem alguns mapas que não possuem qualquer similaridade com outros mapas, não sendo possível determinar, por exemplo, dentre 2 mapas qual deles é o mais próximo a um mapa-alvo (numa situação extrema em que os dois mapas têm similaridade 0 em relação a este último). Deste modo, não foi considerada útil a análise da similaridade relativa para determinação do grau de profundidade que apresenta melhores resultados. Na Figura 5.4, os resultados dos diferentes níveis de profundidade são apresentados de acordo com uma comparação global de similaridade.

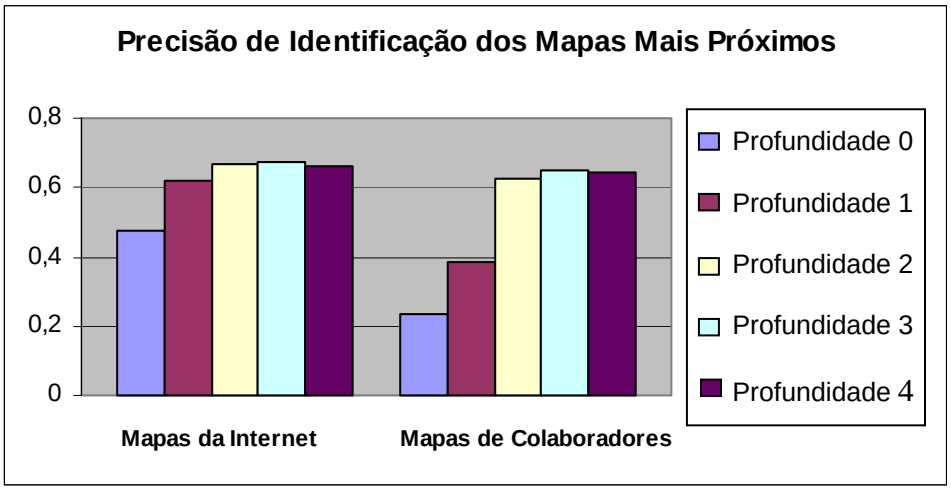


Figura 5.4: Desempenho dos diferentes níveis de profundidade no contraste de característica em identificar os 5 mapas mais próximos para cada mapa-alvo.

Quanto mais generalizarmos os substantivos para tentar encontrar correspondências entre os conceitos, maior é a precisão de acerto do algoritmo. De acordo com os resultados, o valor de profundidade 3 é o que identifica com maior eficácia os 5 mapas mais próximos (precisão média de 0,660 ± 0,034 com um grau de confiança de 95%). A partir do nível 4 em diante, devido à grande quantidade de informação (e consequentemente maior complexidade de pesquisa), há uma correspondência entre quase todos os conceitos dos dois mapas comparados, levando a que sejam, erradamente, considerados como semanticamente próximos.

2.2. Abordagens baseadas na similaridade entre conceitos

Na Secção 4.3.2, foram propostas, ainda, 3 abordagens baseadas na medida adaptada de Jiang e Conrath e no coeficiente Cosseno: Procura Simples pela Maior Similaridade entre conceitos, Procura Ponderada que ordenava os conceitos de acordo com a sua importância em cada mapa, e Procura do conceito central de um mapa para posterior comparação deste com os conceitos do outro mapa conceptual.

Apresentamos, agora, o desempenho destas propostas de similaridade entre mapas conceptuais face à abordagem de Contraste de Características. Para a implementação das 3 medidas descritas anteriormente, foi implementada, de raiz, a medida adaptada de Jiang e Conrath. Após termos testado o método de comparação por contraste de características, verificámos que o coeficiente de Jaccard com profundidade 3 de pesquisa era o que apresentava globalmente os melhores resultados. Deste modo, também é incluído o desempenho desta abordagem na Figura 5.5, que apresenta os resultados dos 4 algoritmos aqui indicados.

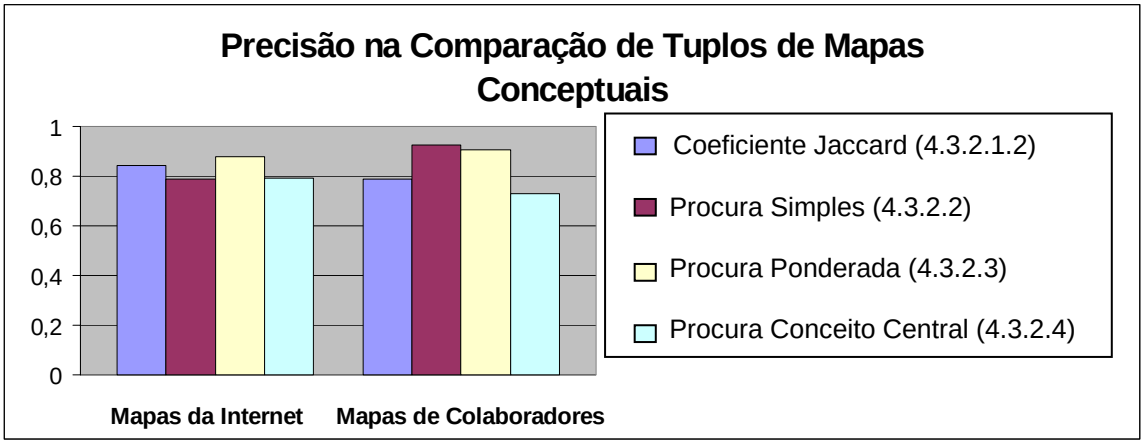


Figura 5.5: Desempenho das diferentes abordagens de comparação semântica (incluindo o Contraste de Características pelo Coeficiente de Jaccard) sobre arranjos de 3 mapas conceptuais

Globalmente, a Procura Ponderada e a Procura Simples apresentam os melhores resultados (precisão média de 0,892 ± 0,039 e 0,856 ± 0,062, respectivamente, com um grau de confiança de 95%). Apesar do método de Procura do Conceito Central ter o pior desempenho comparativamente às outras abordagens, este factor deve-se ao facto de ser muito limitado cingirmo-nos apenas a um único conceito central para todo um mapa conceptual. Deste modo, pelos valores obtidos, para obter uma melhor medida relativa entre dois mapas conceptuais em relação a um mapa-alvo, a abordagem por Procura Ponderada é a mais indicada.

Quanto à capacidade de identificação dos mapas mais próximos (Figura 5.6), entretanto, o método de Contraste de Características pelo coeficiente de Jaccard e profundidade de pesquisa 3 é o que claramente apresenta melhores resultados face às 3 restantes abordagens (precisão média de 0,627 ± 0,028 com um grau de confiança de 95%). No entanto, ao contrário deste método, as abordagens baseadas na similaridade semântica entre os conceitos possuem melhor desempenho nos mapas de colaboradores que são mais ambíguos.

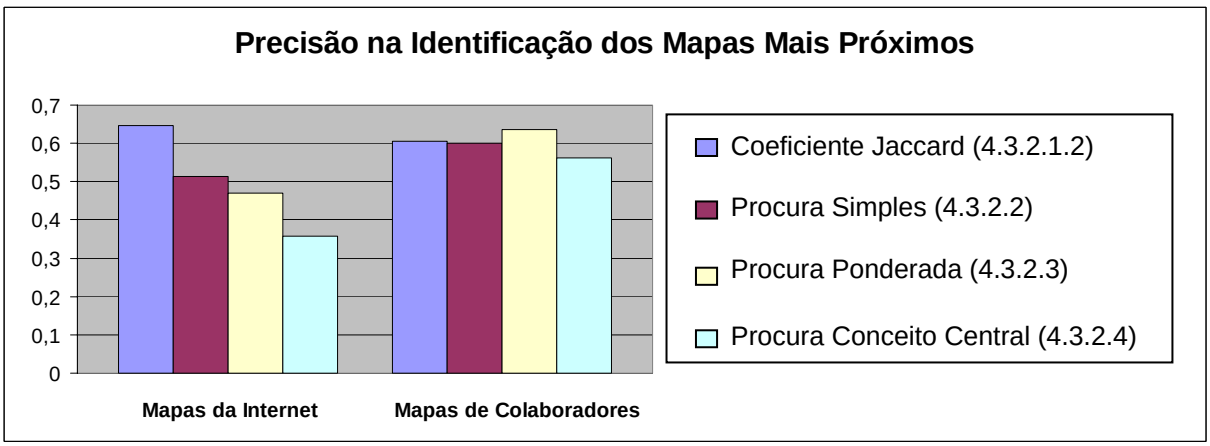


Figura 5.6: Desempenho das diferentes abordagens de comparação semântica (incluindo o Contraste de Características pelo Coeficiente de Jaccard) em identificar os 5 mapas mais próximos para cada mapa-alvo.

Para verificar o impacto do processo de desambiguação no desempenho final da comparação semântica entre mapas conceptuais (Figura 5.7), decidimos confrontar as mesmas abordagens utilizando mapas desambiguados por 3 métodos de contextualização distintos: baseado na medida adaptada de similaridade entre conceitos de Jiang e Conrath; escolha do significado mais utilizado; e desambiguação feita manualmente por juízes humanos. Por fim, de forma a comprovar ou não a necessidade de tal etapa intermédia, aplicámos os algoritmos de comparação sobre mapas originais sem qualquer pré-processamento semântico a fim de identificar a eventual contribuição da eliminação prévia de alguma ambiguidade dos mapas conceptuais alvos de comparação.

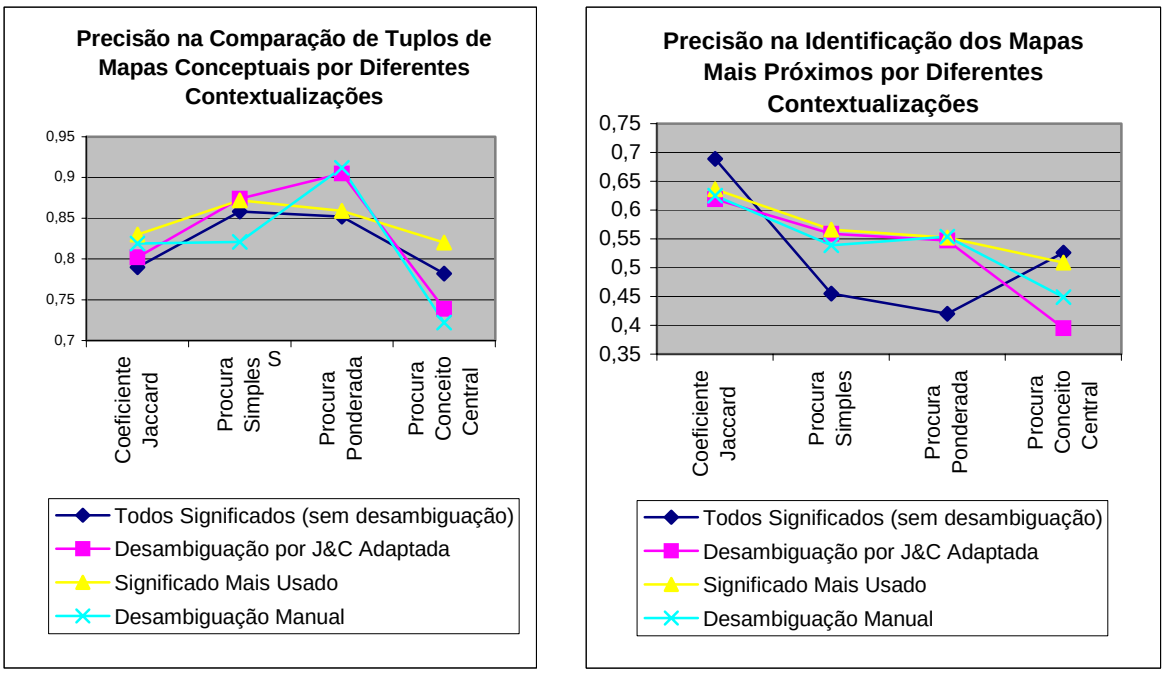


Figura 5.7: Desempenho das diferentes abordagens de comparação semântica com distintos métodos de desambiguação aplicados aos mapas.

Na determinação da similaridade relativa de um mapa em relação a outros dois, a utilização de um método de desambiguação favorece um maior desempenho das abordagens proposta. É importante notar que as medidas com melhor desempenho neste tipo de comparação (Procura Ponderada e Procura Simples) são mais precisas nos mapas que foram desambiguados utilizando a medida adaptada de similaridade entre conceitos proposta originalmente por Jiang e Conrath. Por outro lado, a abordagem que melhor identifica os mapas mais próximos (Contraste de Características pelo Coeficiente de Jaccard) não necessita de qualquer pré-processamento semântico a fim de obter uma maior precisão global. Este facto pode ser explicado pelo aumento de informação (uma vez que cada conceito terá todos os significados disponíveis para a intersecção entre os mapas) e, possivelmente, maior probabilidade de formação de um contexto significativo à volta das palavras polissémicas (em que possua significados muito semelhantes).

Curiosamente, o facto de termos um mapa correctamente desambiguado (correspondendo a todas as escolhas feitas pelo juízes) não implica necessariamente um melhor desempenho final dos algoritmos de comparação, o que sugere a existência de um ruído associado a todo o processo que deverá ser analisado futuramente. Finalmente, podemos concluir que a utilização do método de desambiguação escolhido de selecção do significado mais utilizado aumenta a performance do algoritmo de comparação semântica relativa.

Analisando globalmente este conjunto de experiências, constatámos que o desempenho das medidas propostas é influenciado por dois factores essenciais: 1) ambiguidade e a organização do conteúdo expresso nos mapas conceptuais; 2) tipo de comparação que esperamos conseguir modelar por um sistema computacional.

Se confrontarmos o estilo de representação dos dois conjuntos de mapas, verificamos que os mapas extraídos da Internet [Leung, 2003] exprimem informação de uma forma muito mais detalhada e específica sendo possível relacioná-los ao domínio que representam de uma forma directa, enquanto que os mapas construídos pelo colaboradores incluem geralmente informação mais subjectiva, na medida em que referem além de um assunto central (que muitas vezes não é directamente identificável) outros conceitos mais periféricos. A medida baseada no Contraste de Características utilizando o Coeficiente de Jaccard (com um nível de profundidade de pesquisa de até 3 níveis) demonstrou ser mais eficaz para o primeiro tipo de mapas, enquanto que as medidas que fazem um mapeamento entre os dois mapas à procura da maior similaridade entre os conceitos (Procura Simples e Procura Ponderada) apresentaram globalmente melhores resultados para o segundo tipo de mapas conceptuais.

Quanto à tarefa a que se destina a medida de comparação semântica, podemos concluir que quanto menor é o conjunto de mapas a escolher o mais próximo em relação a um mapa-alvo,

maior é a garantia de acerto do algoritmo. Todas as medidas apresentaram, de uma forma coerente, muito melhores resultados quando foram aplicadas à comparação relativa de mapas (dizer se uma mapa A é mais próximo de B ou C) do que à tarefa de âmbito alargado, e talvez um pouco ambiciosa, de criar um conjunto de mapas ordenados em relação a um alvo. Os próprios juízes humanos encontraram algumas dificuldades em criar listas ordenadas quanto à similaridade de mapas conceptuais. A principal constatação detectada foi o facto de, à primeira vista, parecer óbvio qual era o mapa mais próximo, enquanto que já não parecia tão fácil deduzir qual era o 2º, 3º e assim por diante. Neste contexto podemos verificar pelos resultados apresentados que as medidas de Procura Simples e Procura Ponderada conseguem se sobressair na detecção da similaridade relativa, ao passo que, o método de Contraste de Características tem um desempenho melhor na criação de listas ordenadas de mapas.

6. Conclusões e Trabalho Futuro

Apresentámos neste documento um sistema de comparação semântica entre mapas conceptuais que é resultante de uma investigação em Processamento Semântico e Raciocínio Cognitivo. Neste estudo procurámos, objectivamente, encontrar um padrão de comparação entre mapas conceptuais que se baseasse, fundamentalmente, nos seus significados. Este trabalho está ligado ao desenvolvimento anterior de uma aplicação interactiva de extracção e aprendizagem de mapas conceptuais, um sistema composto pelas aplicações *Clouds* e *TextStorm* [Pereira et. al, 2000; Alves et. al, 2001].

De forma a aproveitar o conhecimento científico já disponível nesta área, foram estudadas e testadas algumas medidas de similaridades entre conceitos já conhecidas [Hist e St-Onge, 1998; Jiang e Conrath, 1997; Leacock e Cheadorow, 1998; Resnik, 1995] a fim de verificar o real desempenho e contributo que este trabalho poderia oferecer. Foi feito um estudo sobre a viabilidade e influência da contextualização do significado inerente aos mapas conceptuais sobre o resultado final de comparação. De acordo com os resultados observados, a escolha do significado mais utilizado com base num *corpus* foi eleita como a técnica de desambiguação mais precisa com um desempenho de quase 88% de acerto.

Esta tarefa auxiliar de desambiguação revela-se fortemente dependente da abrangência do WordNet e do *corpus* de utilização dos conceitos. Alguns conceitos presentes no mapa conceptual não estão classificados e portanto não podem ser considerados na comparação semântica. Prevê-se um alargamento gradual tanto da base lexical como do *corpus* utilizado na medida em que o projecto WordNet tem sido continuamente aumentado com constantes actualizações e, muito brevemente, em diversas línguas de acesso livre.

Foi estudada e testada uma abordagem já conhecida, denominada Contraste de Características [Tversky, 1977], depois de adaptada a 4 coeficientes de similaridade utilizados actualmente para diversos fins (inclusive de similaridade textual com base em palavras-chave): Cosseno, *Dice*, *Jaccard* e Sobreposição, obtendo-se assim 4 variações de uma mesma medida. Propusemos 2 novas propostas de comparação semântica entre mapas conceptuais que consideravam, além do significado dos conceitos, também informação sobre a organização interna dos mapas: Procura Simples das maiores similaridades entre conceitos; e Procura Ponderada das maiores similaridades entre conceitos de acordo com a importância de cada conceito. O resultado desta comparação semântica foi um mapeamento entre os dois mapas conceptuais, sobre o qual era calculada a similaridade global com base numa medida já conhecida de similaridade adaptada entre conceitos, proposta inicialmente por Jiang e Conrath, e

no coeficiente Cosseno. Numa 3ª proposta, foi feita uma primeira aproximação da possibilidade de resumir um mapa conceptual, elegendo o conceito principal do mapa como o mais representativo deste. Apenas considerámos até agora o peso (grau de ligação) dos conceitos no mapa. Estamos a ponderar, de futuro, conjugar o Conteúdo de Informação de cada conceito de forma a ter uma ideia mais realista destes pesos. Esta última abordagem pretendia concentrar a comparação sobre os conceitos mais relevantes, produzindo um resumo dos mapas, e merece ser mais estudada de futuro a fim de ser explorado todo o seu potencial.

Constatámos que existem pelo menos duas formas de comparar semanticamente mapas conceptuais com algoritmos de reduzida complexidade computacional: relativa e globalmente. A primeira diz-nos, de dois mapas, qual o mais próximo em relação a um mapa-alvo; com ela conseguimos uma eficácia de quase 90% comparativamente ao raciocínio humano. Neste tipo de comparação, a contextualização de mapas conceptuais tornou-se uma ferramenta útil na medida em que a performance foi superior sobre mapas já desambiguados. Na segunda forma de comparação, dado um conjunto de mapas, é possível estabelecer um ordenamento quanto à similaridade de mapas conceptuais em torno de um mapa central, neste caso, foi possível chegar a uma precisão de 70%. Com esta medida será possível relacionar o conhecimento textual de uma forma mais aprofundada, ao invés dos conhecidos métodos de palavra-chave. A possível extensão desta medida para outras representações estruturadas, além dos mapas conceptuais, parece ser uma certeza, dado que as propostas de comparação semântica aqui apresentadas são suficientemente genéricas e pouco dispendiosas computacionalmente, não se restringindo a nenhuma particularidade específica dos mapas conceptuais.

Este trabalho tem como principal contributo oferecer um estudo de base às futuras pesquisas em similaridade semântica, na medida em que além de resumir as principais metodologias procura oferecer um conjunto de testes sobre um pequeno universo de mapas conceptuais que podem ser reproduzidos por outros investigadores. Este universo de testes só não é maior porque ainda é pouco atraente a criação de mapas de forma manual ou mesmo através do *TextStorm* e *Clouds* por não ser possível uma visualização imediata do mapa em construção. A tarefa de comparar semanticamente (para efeitos de validação) parece ser ainda menos aliciante aos colaboradores, uma vez que alegam o facto de ser-lhes pedido algo que é muito vago e que pode ser alvo da subjectividade de cada indivíduo.

Para resolver parte deste problema, consideramos útil, como proposta de trabalho futuro, uma ferramenta gráfica com ligação ao WordNet que permita construir estruturas conceptuais em geral (mapas conceptuais, redes semânticas, grafos conceptuais, etc.). Esta ferramenta poderá permitir a construção manual e automática (com auxílio do *TextStorm* e *Clouds*) de estruturas

conceptuais, facilitando a criação de uma maior base de conhecimento e uma maior uniformização das representações actualmente disponíveis. Deste modo, o processo de contextualização poderá ser feito em sincronização com a criação da representação conceptual, onde dada uma palavra o autor do mapa poderá escolher de entre os significados possíveis, e eventualmente incluir outros, o que mais correctamente define o conceito que pretende exprimir.

Outra dificuldade encontrada ao longo do nosso trabalho foi a definição de como seria o processo de validação dos resultados obtidos pelas medidas semânticas. Sendo óbvio que o objectivo principal não era oferecer valores absolutos de distâncias entre conceitos, este processo ainda não está completamente finalizado, na medida em que ainda queremos estudar analiticamente o desempenho das medidas perante um tipo de comparação (relativa) entre mapas conceptuais face a outros tipos menos proveitosos (comparação global).

Após inúmeras medidas de similaridade entre conceitos estudadas, verificámos que ainda não existe uma que seja realmente uma medida de *semantic relatedness*, oferecendo, além da similaridade de características, uma componente que contabilize a utilização das palavras em conjunto. Como base de todo o trabalho aqui desenvolvido, sendo empregue nas medidas propostas de Procura Simples e Procura Ponderada, pretendemos desenvolver uma medida de similaridade entre conceitos que tenha como principais componentes estes dois factores de comparação semântica entre conceitos: similaridade taxonómica e similaridade de co-ocorrência. Este próximo passo só será possível através da utilização de grande *corpus* de textos anotados com os conceitos da base lexical utilizada. Para tal poderá ser aproveitada informação já desambiguada, como é o caso do *Extended* WordNet [HLTRI, 2003] para aumentar a capacidade de previsão de utilização quotidiana dos conceitos. Sendo um elemento fundamental das propostas aqui apresentadas, esperamos que ao melhorar a eficácia da medida base de similaridade entre conceitos possamos aumentar o desempenho final das medidas de comparação de estruturas conceptuais.

7. Referências

Agirre E. e G. Rigau (1996) *Word Sense Disambiguation using Conceptual Density.* Proceedings of 15th International Conference on Computational Linguistics, COLING'96. Copenhagen, Denmark.

Alves, A. O, F. C. Pereira e A. Cardoso (2001) *Automatic Reading and Learning from Text*, in Proceedings of the International Symposium on Artificial Intelligence, pp. 302-310, ISBN: 81-7764-298-7, ISAI'2001, Fort Panhala (Kolhapur), India, http://ailab.dei.uc.pt/publications.php

Ausubel D. (1963) *The Psychology of Meaningful Verbal Learning*, Grune and Stratton, New York.

Axelrod, R. (1976) *Structure of Decision*, Princeton, New Jersey: Princeton University Press.

Borgatti, S. Disponível *on-line* em: http://www.analytictech.com/mb021/graphtheory.htm Última actualização em Janeiro de 2001.

Brooks, T. A. (1998) *The Semantic Distance Model of Relevance Assessment*.

Budanitsky, A. e G. Hirst (2001) *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*. Proc NAACL 2001 WordNet and Other Lexical Resources Workshop, 29-34, Pittsburgh.

Bunke, H. e K. Shearer (1998) *A graph distance metric based on the maximal common subgraph*, Pattern Recognition Letters, Vol 19, Nos 3 - 4, 255 – 259.

Catarci, T., L. Iocchi, D. Nardi, e G. Santucci (1997) *Conceptual Views over the Web*, in *Intelligent Access to Heterogeneous Information* - Proceedings of the 4th "Knowledge Representation Meets Databases" (KRDB) Workshop, Athens, Greece, http://SunSITE.Informatik.RWTH-Aachen.DE/Publications/CEUR-WS/Vol-8/

Collins, A. e M. Quillian (1969) *Retrieval time from semantic memory.* Journal of Verbal Learning and Verbal Behavior. p. 240.

Cormen, H., C. Leiserson e R. Rivest (1989) *Introduction to Algorithms* - Cambridge ; McGraw-Hill : The MIT Press : New York. 23: 463-493.

Corneil, D. e C. Gotlieb (1970) *An efficient algorithm for graph isomorphism*. Journal of the ACM, 17:51-64.

Crandell, T. L., N. A. Kleid, e C. Soderston (1996) *Empirical Evaluation of Concept Mapping: A Job Performance Aid for Writers*. Technical Communication (2nd quarter), pp. 157-163.

Cuena J., e M. Molina (1996) *Building Knowledge Models Using KSM*. Proceedings of Knowledge Acquisition of Knowledge Based Systems Workshop, KAW96. Banff, Canada.

Delugach, H. S. (1992) *An Exploration Into Semantic Distance*, Lecture Notes in Artificial Intelligence, no. 754, Chapter 9, Springer-Verlag, Berlin, 1993. (reprinted from Proc. Seventh Annual Workshop on Conceptual Graphs, pp. 29-37, New Mexico State University, Las Cruces, New Mexico, July 8-10, 1992)

Dress, A. e A. von Haeseler (1990) *Trees and hierarchical structures*. In S. Levin, Lecture Notes in Biomathematics, 84. Berlin: Springer-Verlag.

Duda, R. e P. Hart (1973) *Pattern Classification and Scene Analysis*. Wiley-Interscience, USA, 1st edition.

Ellman, J. (2000) , *Using Roget's Thesaurus to Determine the Similarity of Texts*. Phd's Thesis. University of Sunderland, UK, http://www.ellman.freeserve.co.uk/papers/ellman_phd.pdf

Fellbaum, C. (1998) *WordNet*: *An Electronic Lexical Database*. 2ª edição. MIT Press Cambridge, Massachussetts London.

Foo, N., B. J. Garner, N. Rao e E. Tsui (1989) *Semantic Distance in Conceptual Graphs*.

Garey, M., R., e D. S. Johnson (1979) *Computers and Intractability*: *A guide to the theory of NP-Completeness*. Freeman and Company, 1979.

Garner, B.J., D. LUKOSE, e E. TSUI (1987) *Parsing Natural Language through Pattern Correlation and Modification*, Proceedings of the 7th International Workshop on Expert Systems & Their Applications, Avignon, France, pp. 1285-1299.

Gentner, D., e I. France (1988) *The Verb Mutability Effect: Studies of the Combinatorial Semantics of Nouns and Verbs* in Small, S., Cottrell, G., and Tanenhaus, M. (eds.). *Lexical Ambiguity Resolution*. Los Altos, California: Morgan Kaufmann.

Goldsmith, T. e D. Davenport (1990) *Assessing structural similarity of graphs*. Em R. W. Schvaneveldt (Ed.), Pathfinder associative networks: Studies in knowledge organization. Norwood, NJ: Ablex, pp. 75-87.

Goldstone, R. L. (1999) *Similarity*, R.A. Wilson & F. C. Keil (eds.) MIT encylopedia of the cognitive sciences, pp. 763-765. Cambridge, MA: MIT Press, http://cognitrn.psych.indiana.edu/rgoldsto/pdfs/mitecs.pdf

Graesser, A. C. e L. F. Clark (1985) *Structures and Procedures of Implicit Knowledge,* New Jersey, Ablex.

Guarino, N., C. Masolo e G. Vetere (1999) *OntoSeek: Content-Based Access to The Web*, IEEE Intelligent Systems (14:3), 1999, pp. 70-80.

Guthrie, J., L. Guthrie, Y. Wilks e H. Aidinejad (1991) *Subject-Dependent Co-Occurrence and Word Sense Disambiguation*, ACL-91, pp. 146-152.

Hahn, U., e N. Chater (1998) *Understanding Similarity: a Joint Project for Psychology, Case-Based Reasoning and Law*. Artificial Intelligence Review, 12, pp. 393-427.

Harabagiu, S. e D. Moldovan (1998) *Knowledege Processing on Extended WordNet,* in *WordNet: An Electronic Lexidcal Database aned Some of ist Applications,* C. Fellbaum (ed), MIT Press, pp. 379-405.

Harabagiu, S., G. Miller e D. Moldovan (1999) *WordNet 2 – A Morphologically and Semantically Enhanced Resource*, em Proceedings of the SIGLEX Workshop.

Hibberd, R., A. Jones e E. Morris (2002) *The Use of Concept Mapping as a Means to Promote and Assess Knowledge Acquisition (CALRG Report No. 202).* Milton Keynes, UK: The Open University.

Hirst, G. e D. St-Onge (1998) *Lexical chains as representations of context for the detection and correction of malapropisms*, in *WordNet: An Electronic Lexidcal Database aned Some of ist Applications,* C. Fellbaum (ed), MIT Press, pp. 305–332.

Holley, C., D. Dansareau, B. McDonald, J. Garland e K. Collins (1979) *Evaluation of a hierarchical technique as an aid to text processing.* Contemporary Educational Psychology, 4, pp. 227-237.

Hoz, R., Y. Tomer e P. Tamir (1990). The relations between disciplinary and pedagogical knowledge and the length of teaching experience of biology and geography teachers. *Journal of Research in Science Teaching*, 27, 973-985.

James W. (1890) *The Principles of Psychology*, volume 1, capítulo 12: p. 459, http://psychclassics.yorku.ca/James/Principles/index.htm

Jiang, J. e D. Conrath (1997) *Semantic similarity based on corpus statistics and lexical* taxonomy, Proceedings of International Conference Research on Computational Linguistics, Taiwan.

Jonassen, D. (1996) *Computers in the classroom: Mindtools for critical thinking*. Eaglewoods, NJ: Merill/Prentice Hall.

Jonassen, D H, Reeves, T & Hong, N (1998) *Concept Mapping as Cognitive Learning and Assessment Tools.* Journal of Interactive Learning Research, 8,3/4, pp. 289-308.

Jonasson, D. e B. Grabowski (1993) *Handbook of Individual Differences, Learning, and Instruction*. Hillsdale, NJ: Lawence Erlbaum Associates.

Jurafsky, D. e J. Martin (2000) *Speech and Language Processing:An Introduction to Natural Language Processing,* Computational Linguistics and Speech Recognition. Prentice Hall, 1ª edição.

Kim, Y. W. e J. H. Kim (1990) *A model of knowledge based information retrieval with hierarchical concept graph. Journal of Documentation,* 2:113-37.

Kremer, R. e B. Gaines (1994) *Groupware Concept Mapping Techniques.*

Kremer, R. (1994) *Concept Mapping: Informal to Formal*. Proceedings of the Third International Conference on Conceptual Structures, Knowledge Represetnation Workshop, University of Maryland.

Kučera, H. e W. N. Francis (1967) *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lambiotte, J., D. Dansereau, D. Cross, e S. Reynolds (1989) *Multirelational Semantic Maps.* Educational Psychology Review, 1(4), pp. 331-367.

Landes, S., C. Leacock e I. Tengi (1998) *Building semantic concordances*, *WordNet: An Electronic Lexidcal Database and Some of its Applications,* C. Fellbaum (ed.), MIT Press, 1998, pp, 199-266.

Lanzing, J. (1997) *The Concept Mapping Homepage*, http://users.edte.utwente.nl/lanzing/cm_home.htm

Lawson, M. (1994) *Concept Mapping*. In T. Husén & T. N. Postlethwaite (Eds.), The international encyclopedia of education (2nd ed., Vol. 2, pp. 1026-1031). Oxford: Elsevier Science.

Leacock, C. e M. Chodorow (1998) *Combining local context and WordNet similarity for word sense identification*, in *WordNet: An Electronic Lexidcal Database aned Some of ist Applications,* C. Fellbaum (ed), MIT Press, pp. 265-283.

Lee, J., M. Kim e Y. Lee (1993) *Information Retrieval based on conceptual distance in IS-A hierarchies*, Journal of Documentation, 49(2), June, pp. 188-207.

Leung, J. (2003) *Concept Maps on Various Topics*, (actualizado em Março de 2004) http://www.fed.cuhk.edu.hk/~johnson/misconceptions/concept_map/concept_maps.html

Lesk, M. (1986) *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone*. Proceedings of the 5th annual international conference on Systems documentation, p.24-26, June, Toronto, Ontario, Canada

Li, X, S. Szpakowicz e S. Matwin (1995) *A WordNet-based algorithm for word sense disambiguation*. Em Proceedings of the 14th International Joint Conference on Artificial Intelligence,IJCAI-95, Montreal, Canada.

Lin, D. (1998) *An information-theoretic definition of similarity*. In Proceedings of the 15th International Conference on Machine Learning,Madison, WI.

Lomask, M., J. Baron, J. Greig, e C. Harrison (1992) *ConnMap:Connecticut's use of concept mapping to assess the structure of students knowledge of science*. Paper presented at the annual meeting of the National Association of Research in Teaching, Cambridge, MA.

Love, B. (2000) *A Computational Level Theory of Similarity*. Proceeding of the Cognitive Science Society, pp. 316-321.

Lukose, D., G. Mineau, M. L. Mugnier, J. U. Möller, P. Martin, R. Kremer e G. P. Zarri (1995) *Conceptual structure for Knowledge Engineering and Knowledge Modelling.*

Lytinen, S., L., N. Tomuro e T. Repede (2000) *The use of WordNet sense tagging in FAQFinder*, In *Proceedings of the AAAI-2000 workshop on AI and Web Search*, Austin, Texas.

Manning, C. e H. Schutze (1999) *Foundations of Statistical Natural Language Processing*. MIT Press. 1ª edição.

Marrafa, P. (2001) *WordNet do Português: uma base de dados de conhecimento linguístico*, Lisboa, Instituto Camões.

Markman, A. B. e D. Gentner (1993) *Structural Alignment during Similarity Comparisons*. Cognitive Psychology 25, pp. 431-467.

Martin, P. (1995) *Knowledge Acquisition using Documents, Conceptual Graphs and a Semantically Structured Dictionary*, em Proceedings of the 9th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff Conference Centre, Bannf, Alberta, Canada, artigo nº 9.

McAleese, R. (1994) *A theoretical view on concept mapping*. ALT-J, 2, 1, 38–48.

McClure, J. e P. Bell (1990) *Effects of an Environmental Education-Related STS Approach to Instruction on Cognitive Structures on Preservice Science Teachers.*University Park, PA.: Pennsylvania State University.

Messmer, B. (1995) *Efficient Graph Matching Algorithms for Preprocessed Model Graphs*, PhD thesis, Institut fur Informatik und Angewandte Matheatik, Universitat Bern, Switzerland.

Mihalcea, R. (2003) Software e Recursos Linguísticos disponíveis *on-line*: http://www.cs.unt.edu/~rada/downloads.html

Mihalcea, R. e D. Moldovan (2000) *Semantic Indexing Using WordNet Sense*.. Proceedings of ACL Workshop on IR & NLP, Hong Kong, October.

Miller, G. (1990) *An On-Line Lexical Database*. International Journal of Lexicography, 13(4), pp.235-312.

Miller, G. (1991). *The science of words.* New York: Scientific American Library.

Miller, G. e W. Charles (1991). *Contextual correlates of semantic similarity*. Language and Cognitive Processes, 6:1–28.

Miller, G., R. Beckwith, C. Felbaum, D. Gross e K. Miller (1993) *Introduction to WordNet: An On-line Lexical Database;* http://www.cogsci.princeton.edu/~wn/w3w n.html (revisto em Agosto de 1993)

Mitchell, T, R. Keller e S. Kedar-Cabelli (1986) *Explanation-based learning: A unifying view*, Machine Learning 1(1):47-80.

Moldovan, D. (2003) *The Extended WordNet Project*,(actualizado em 2003), http://xwn.hlt.utdalas.edu

Möller, J-U. e M. Willems (1995) *CG-Desire: Formal Specification using Conceptual Graphs,* em Proceedings of the 9th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff Conference Centre, Bannf, Alberta, Canada, artigo nº 11.

Montes-y-Gómes, M., A. Guelbukh e A. López-López (2000) *Comparison of Conceptual Graphs*. MICAI 2000, Avances en Inteligencia Artificial. Acapulco, Mexico.

Murray, P. *New language for new leverage: the terminology of knowledge management,* http://www.ktic.com/topic6/13_TERM5.HTM

Myaeng, S. e A. López-López (1992) *Conceptual Graph Matching: a flexible algorithm and experiments.* Journal of Experimental and Theoretical Artificial Intelligence. Vol. 4, 1992, pp. 107-126.

Nersessian, N. (1989) *Conceptual change in science and in science education*, *Synthese*, vol. 80, no. 1 pp. 163-184.

Nosek, J. T. e I. Roth (1990) *A comparison of Formal Knowledge Representation Schemes as Communication Tools: Predicate Logic vs. Semantic Networks.* International Jornal of Man-Machine Studies, 33, pp. 227-329.

Novak, J. (1965) *A model for the interpretation and analysis of concept formation*. Journal of Research in Science Teaching, 3, pp. 72-83.

Novak, J. e D.B. Gowin (1984) *Learning How To Learn*, New York: Cambridge University Press.

Novak, J. (2002) *The Theory Underlying Concept Maps and How To Construct Them*, Cornell University,http://cmap.coginst.uwf.edu/info/

Oxford English Dictionary *on-line* (2003), http://dictionary.oed.com

Pedersen, T. (2003) *Semantic Distance in WordNet Package*.
Disponível em : http://www.d.umn.edu/~tpederse/tools.html

Pereira, F. C., A. Oliveira e A. Cardoso (2000) *Extracting Concept Maps with Clouds*, in Proceedings of the Argentine Symposium on Artificial Intelligence, ASAI'00, Buenos Aires, Argentina, http://ailab.dei.uc.pt/publications.php

Quillian, M. (1968) *Semantic memory*, M. Minsky (Ed.), *Semantic information processing*. Cambridge, MA: MIT Press.

Rada, R. e E. Bicknell (1989) *Ranking Documents with a Thesaurus*. Journal of the American Society for Information Science 40(5); pp. 304-310.

Rada, R, H. Mili, E. Bicknell e M. Blettner (1989). *Development and application of a metric on semantic nets*.IEEE Transaction on Systems, Man, and Cybernetics, 19(1), pp. 17-30.

Rao, A. e N. Foo, N. (1987) *CONGRES - Conceptual Graph Reasoning System: Third Conference on Artificial Intelligence and Applications*, (IEEE), Florida.

Resnik, P. (1995) *Using information content to evaluate semantic similarity in a taxonomy*, em Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montréall; 1:448-53.

Resnik, P. (1999) *Disambiguating Noun Groupings with Respect to WordNet Senses*. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky (eds.), Natural Language Processing using Very Large Corpora, pp. 77-98. Kluwer Academic Press.

Richardson, R. e A. Smeaton (1995) *Using WordNet in a Knowledge-Based Approach to Information Retrieval.* Dublin City University School of Computer Applications Working Paper.

Rigau, G., J. Atserias e E. Agirre (1997). *Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation*. Proceedings of joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics ACL/EACL'97. Madrid, Spain.

Rips, L. J., E. J. Shoben. e E. E. Smith (1973) *Semantic Distance and the verification of semantic relations. Journal of Verbal Learning and Verbal Behavior,* 12, 1-20.

Rivest, R. (1987) *Learning decision lists*. Machine Learning, 2(3):229-246.

Ross, S. (1976) *A First Course in Probability*. Macmillan, NewYork. ISBN: 0024038806

Ruiz-Primo, M., S. Shavelson e S. Schultz (1997) *On The Validity Of Concept Map-Base Assessment Interpretations: An Experiment Testing The Assumption Of Hierarchical Concept Maps In Science*. CSE Technical Report 455, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE) Graduate School of Education & Information Studies University of California, Los Angeles, California.

Schank, R.C. e R. Abelson (1977) Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Earlbaum Assoc.

Schutze, H. (1998) *Automatic Word Sense Discrimination*. Computational Linguistics,Vol. 24:1,97-124.

Schvaneveldt, R. W., F. T. Durso, e B. R. Mukherji (1982) *Semantic distance effects in categorization tasks. Journal of Experimental Psychology: Learning, Memory and Cognition,* 8, 1-15.

Sharples, M., D. Hogg, C. Hutchison, S. Torrance e D. Young (1996) *What are semantic networks? A little light history*. *Computers and Thought: A pratical introduction to Artificial Intelligence,* http://www.cogs.susx.ac.uk/local/books/computers-and-thought

Shaw, M. L. G. e B. R. Gaines (1995) *Knowledge and Requirements Engineering,* em Proceedings of the 9th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff Conference Centre, Bannf, Alberta, Canada, artigo nº 44.

Shearer, K., H. Bunke, S. Venkatesh e D. Kieronska (1997) *Efficient graph matching for video indexing*. In: Jolion, J.-M.,. Kropatsch, W. Eds. , Preproceeding GbR'97: IAPR Workshop on Graph based Representations, Lyon.

Shoben, E. J. (1976) *The verification of semantic relations in a same-different paradigm: An Asymmetry in semantic memory.* Journal of Verbal Learning and Verbal Behavior, 15, 365-379.

Smeaton, A. e I. Quigley (1996) *Experiments on using semantic distances between words in image caption retrieval.* In H. Frei, Donna Harman, P. Schäuble, and R. Wilkinson, editors, Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 174-180. ACM Press.

Southey, F, e J. Linders (1999) *NOTIO-A Java API for Developing CG Tools*, in *Conceptual Structures: Standard and Pratices.* W. Tepfenhart & W. Cyre (eds.). Lecture Notes in Artificial Intelligence. Berlim, pp. 262-271, Springer.

Sowa, J. F. (1984) *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley, Reading, MA.

Sowa, J. F. e E. C. Way (1986) *Implementing a semantic interpreter using conceptual graphs.* Ibm Journal Res. Develop., 30(1): 57-69.

Stetina, J. et al. (1998) *General word sense disambiguation method based on a full setential context*. Proceedings COLLING-ACL Workshop, Usage of WordNet in Natural Language Processing Systems, pp. 33-38.

Sussna, M. (1993) Word *sense disambiguation for free-test indexing using a massive semantic network.* Proceedings of the 2nd International Conference on Information and Knowledge Management. Arlington, Virginia, USA.

Thadgard, P. (1992) *Conceptual Revolutions*, Princeton, New Jersey: Princeton University Press.

Tsui, E. (1988) *Canonical Graph Models*, Ph.D. thesis, Department of Computing and Mathematics, Deakin University, Australia.

Tversky, A. (1977) Features of Similarity.Psychological Review 84(4), pp. 327-352.

Ullman, J. (1976) *An algorithm for subgraph isomorphism*. *Journal of the ACM*, 23(1):31-42, Janeiro.

Voorhees, E. (1993) *Using WordNet to Disambiguate Word Senses for Text Retrieval*. Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval. Pittsburgh, PA. pp.171-180

Voorhees, E. (1998) *Using WordNet for Text Retrieval*. in: in *WordNet: An Electronic Lexidcal Database aned Some of ist Applications,* C. Fellbaum (ed), MIT Press, pp. 285-303.

Vossen, P. (1997) *EuroWordNet: a multilingual database for information retrieval*, Proceedings of the DELOS workshop on Cross-language Information Retrieval. Zurique, http://www.illc.uva.nl/EuroWordNet/docs/P011.ai

Whyte, L., A. Wilson e D. Willson (1969) *Hierarchical Structures*. American Elsevier Publishing Compagny, New York. ISBN 444-00069-0. p. 317.

Woods, W. A. (1975) *What's in a link: foundations for semantic networks*. Em D.G. Bobrow e A.M. Collins, (Eds.), Representation and Understanding: Studies in Cognitive Science,. New York: Academic Press, pp. 35-82.

Yarowsky, D. (1992) *Word-sense disambiguation using statistical models of Roget's categories trained on large corpora*. In Proceedings of COLING-92, pages 454-460, Nantes, France.

Zarri, G.P. (1997) *NKRL, a Knowledge Representation Tool for Encoding the 'Meaning' of Complex Narrative Texts*, Natural Language Engineering - Special Issue on Knowledge Representation for Natural Language Processing in Implemented Systems 3, 231-253.

Zimmaro, D. e J. Cawley (1998). *Concept map module*. Schreyer Institute for Innovation in Learning, The Pennsylvania State University, http://www.ttuhsc.edu/SOM/success/DHPS/Concept%20Map%20Module.htm

Anexos

A.1. Introdução à Teoria dos Grafos

Além do significado que os conceitos possuem, é importante observar a forma como a informação está interligada num mapa conceptual. Tal como qualquer linguagem de representação do conhecimento, além da sua semântica é necessário encontrar a sintaxe por detrás dos arcos e nós que representam uma visão do domínio representado. Neste anexo, iremos abordar a estrutura de dados representada por um mapa conceptual: os grafos; e, que contributos esta análise poderá adicionar ao processo de comparação de mapas conceptuais.

A.1.1 Estrutura de dados para representação computacional

Por representar uma estrutura complexa onde estão inseridos conceitos ligados por vários tipos de relações, um mapa conceptual pode ser também mais formalmente definido como um grafo dirigido acíclico n-dimensional composto por um conjunto de vértices representado por **m** conceitos ou vértices **V**={v_1,...,v_m} e um conjunto não-vazio de **n** nomes de relações ou arcos **E=**{e_1,...,e_n} que interliga um determinado subconjunto de vértices [McAleese, 1994] cujo o valor de cada arco é dado por uma função $\omega : E \to R$, onde **R** é o conjunto de etiquetas atribuídas aos arcos (ver exemplo na Figura A.1).

O recurso léxico-semântico escolhido, o WordNet, também pode ser computacionalmente definido como um grafo, uma vez que existem diversos tipos de relações semânticas (*is-a, part-of, atribute-of,...*) que unem os conceitos (*synsets*), formando um espaço multi-dimensional.

Computacionalmente, existem duas formas básicas de representar grafos, através de matrizes de adjacência ou listas de adjacência. Veremos de seguida em detalhe cada uma destas representações:

Matriz de Adjacências

Dado um grafo **G=(V,E)**, sendo **V** o seu conjunto de vértices, e **E** o seu conjunto de arcos, a matriz de adjacências **A=(a_{ij})** é uma matriz **m x m** tal que:

m é o número de vértices, |**V**|;

n é o número de arcos, |**E**|;

$a_{ij} = \omega_{ij}$ se **(v_i,v_j)** pertence a **E**, ou seja, se forem adjacentes e houver um arco que divirja de v_i e convirja em v_j;

$a_{ij} = 0$ caso contrário;

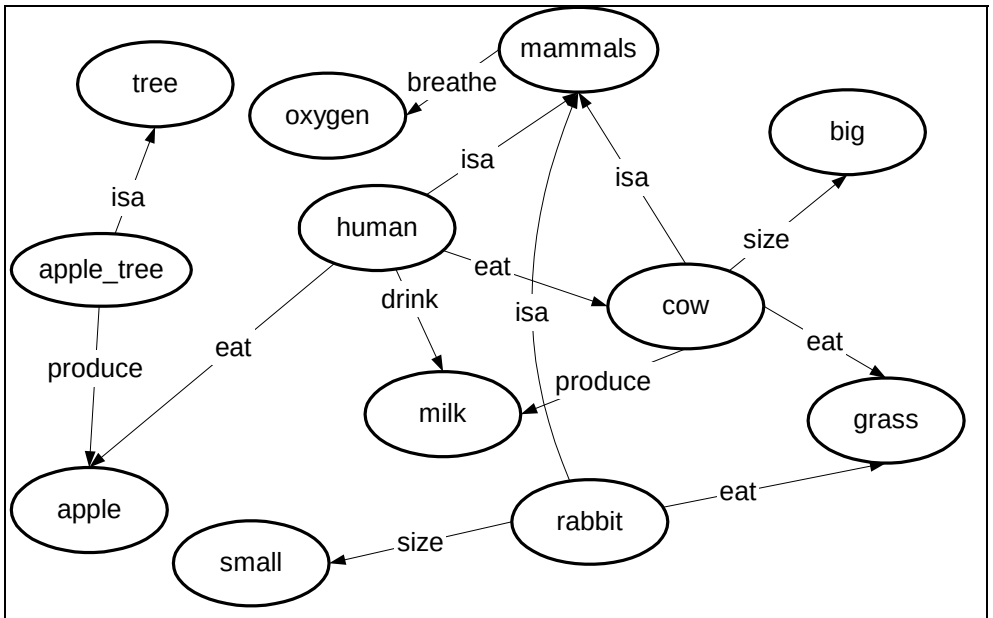


Figura A.1: Exemplo de um mapa conceptual e respectivos conjuntos para representação por grafo.

A matriz de adjacências (Tabela A.1) representa unicamente um grafo sem ambiguidade, porém a um mesmo grafo **G** podem corresponder várias matrizes diferentes, onde em cada uma destas o número de células diferentes de 0 é exactamente igual a **n**. Com a vantagem de permitir o acesso directo a qualquer arco, esta representação necessita de um espaço mínimo necessário em memória na ordem de $\Theta(m^2)$ independentemente do número de arcos existentes, sendo assim pouco apropriada para grafos esparsos (quando |**E**| é muito menor que |**V**|^2).

Lista de Adjacências

Seja **G(V,E)** um grafo, define-se a lista de adjacências **L=(l$_{ij}$)** um vector de **m** listas, uma para cada vértice do grafo. Para cada $v \in V$, a lista de adjacência **L[*v*]** aponta para todos os vértices *u*, tal que o arco (*v*, *u*) $\in E$ e cada um dos elementos da lista armazena o peso ou etiqueta atribuído ao arco pela função $\omega : E \to R$. Se G é um grafo dirigido, a soma do comprimento de todas as listas é |**E**|, requerendo um espaço de armazenamento máximo de *O(V+E)*. Na Figura A.2 é apresentada uma representação possível do mapa conceptual utilizada como modelo.

	apple	apple_tree	big	cow	grass	human	mammals	milk	oxygen	rabbit	small	tree
apple												
apple_tree	produce											isa
big												
cow			size		eat		isa	produce				
grass												
human	eat			eat			isa	drink				
mammals									breathe			
milk												
oxygen												
rabbit					eat		isa				size	
small												
tree												

Tabela A.1: Representação através de uma matriz de adjacências do mapa conceptual da Figura A.1.

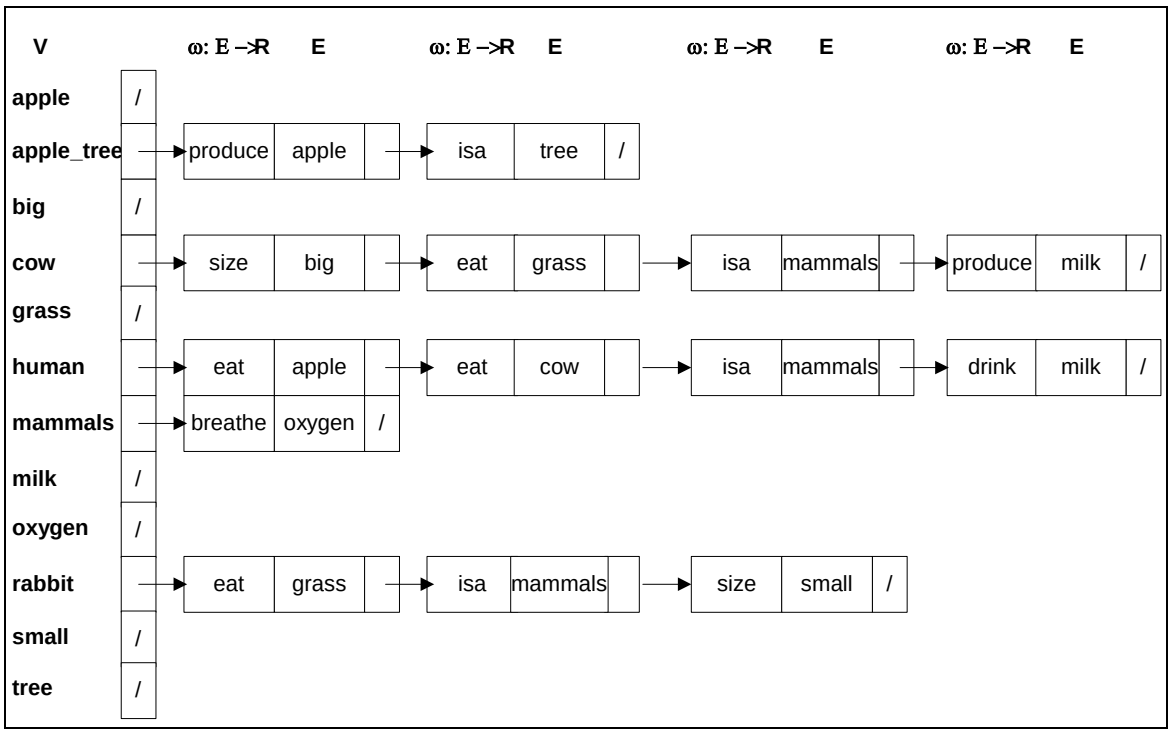


Figura A.2: Representação através de uma lista de adjacências do mapa conceptual da Figura A.1.

A.1.2 Relações entre Grafos

Na Teoria dos Grafos [Borgatti, WWW], já existem um conjunto de relações definidas de similaridade estrutural entre grafos que se baseiam não só na quantidade de vértices comuns como também na semelhança de organização entre estes. Estamos interessados em determinar se

dois grafos são ou não iguais, e caso não sejam, se existe alguma região partilhada por ambos os grafos. Com este objectivo, observaremos de seguida as relações estruturais entre grafos: Isomorfismo, Isomorfismo de Subgrafos e Máximo Subgrafo Comum.

Isomorfismo entre grafos

Sejam dois grafos $\mathbf{G_1(V_1,E_1)}$ e $\mathbf{G_2(V_2,E_2)}$, com $\mathbf{|V_1|=|V_2|=n}$. Se existe uma função unívoca $\mathbf{f:V_1->V_2}$, tal que **(v,w)** pertence a $\mathbf{E_1}$ sse **(f(v),f(w))** pertence a $\mathbf{E_2}$, para todo **v**, **w** pertencente a $\mathbf{V_1}$, então $\mathbf{G_1}$ e $\mathbf{G_2}$ são ditos isomorfos entre si. No caso de grafos etiquetados, as designações de cada arco do primeiro grafo devem ser mantidas no segundo. Por esta definição, dois grafos são ditos completamente equivalentes um ao outro e não iguais, uma vez que permite relacionar grafos que mantenham um mapeamento de um para um entre os vértices através de uma função de mapeamento (que num caso especial, pode ser a de igualdade).

Qualquer algoritmo para comparação entre grafos requer um determinado nível de complexidade computacional [Messmer, 1995]. Ainda não é consensual a classificação da complexidade do problema do isomorfismo entre grafos como sendo polinomial ou não (classe P ou NP) [Garey e Johnson, 1979], sendo que a maioria dos algoritmos para a detecção desta relação necessitam, no pior cenário, de um tempo exponencial de computação.

Dados dois grafos, todos os arcos do primeiro grafo devem estar presentes e unindo os vértices equivalentes no segundo grafo, sem excepção, caso contrário não estamos perante a uma relação de isomorfismo.

Isomorfismo de subgrafos (subgraph isomorphism)

Um subgrafo $\mathbf{G_2(V_2,E_2)}$ de um grafo $\mathbf{G_1(V_1,E_1)}$ é um grafo tal que $\mathbf{V_2}$ está contido em $\mathbf{V_1}$ e $\mathbf{E_2}$ está contido em $\mathbf{E_1}$. Se o subgrafo $\mathbf{G_2}$ de $\mathbf{G_1}$ satisfaz: para quaisquer **v**, **w** pertencente a $\mathbf{V_2}$, se **(v,w)** pertence a $\mathbf{E_1}$, então **(v,w)** pertence a $\mathbf{E_2}$. Dessa forma, $\mathbf{G_2}$ é dito subgrafo induzido pelo conjunto de vértices **V2**.

Na procura de isomorfismo entre os grafos $\mathbf{G_1}$ e $\mathbf{G_2}$, se um dos grafos envolvidos é maior do que outro, e.g. $\mathbf{G_1}$ contém mais vértices que $\mathbf{G_2}$, então estamos a tentar encontrar um isomorfismo entre um subgrafo **S** de $\mathbf{G_1}$ e $\mathbf{G_2}$. Nesta definição de isomorfismo de subgrafos é aplicada a ideia anterior de isomorfismo, só que agora, entre um grafo e uma parte de outro, ou seja, é possível saber se um grafo está contido ou não em outro.

Há muito é sabido que o problema do isomorfismo de subgrafos é NP-completo [Garey e Johnson, 1979] ao contrário do que acontece com o isomorfismo de grafos. Consequentemente, nenhum algoritmo poderia encontrar isomorfismos de subgrafos num tempo polinomial. No

entanto, alguns trabalhos têm mostrado a existência de métodos que se comportam razoavelmente bem na maioria dos casos e só se tornam computacionalmente intratável numa pequena parte.

Um dos métodos mais conhecidos para **detectar isomorfismos entre grafos e subgrafos** é baseado na procura em profundidade com *backtracking*, descrita pela primeira vez em [Corneil e Gotelieb, 1970]:

Dados dois grafos $\mathbf{G_1(V_1,E_1)}$ e $\mathbf{G_2(V_2,E_2)}$, os vértices em $\mathbf{V_1}$ são mapeados um por um nos vértices em $\mathbf{V_2}$ e depois de cada mapeamento é verificado se a estrutura dos arcos em $\mathbf{E_1}$ é preservada em $\mathbf{E_2}$ pelo mapeamento. Existe um isomorfismo entre os grafos $\mathbf{G_1}$ e $\mathbf{G_2}$, se estes forem de igual tamanho e todos os vértices de $\mathbf{G_1}$ forem mapeados com sucesso nos vértices de $\mathbf{G_2}$. Se $\mathbf{G_1}$ é de menor dimensão que $\mathbf{G_2}$ então temos um isomorfismo do subgrafo $\mathbf{G_1}$ com $\mathbf{G_2}$.

Apesar deste método ser eficiente para grafos de dimensões reduzidas (aproximadamente até 10 vértices), o número de operações aumenta exponencialmente à medida que o grafo é maior. Posteriormente, foi proposto por [Ullman, 1976] combinar *backtracking* com um método de verificação antecipada que reduz em grande parte a necessidade de operações desnecessárias. Ambos os algoritmos são óptimos uma vez que encontram todos os isomorfismos de grafos e subgrafos de $\mathbf{G_1}$ para $\mathbf{G_2}$, tendo como principal desvantagem o tempo não ser polinomial para grafos maiores. Face a este problema, podem ser adoptados algoritmos de optimização ou aproximação, que necessitam apenas de um tempo polinomial de execução [Messemer, 1995], contudo não garantem encontrar a solução óptima.

Máximo Subgrafo Comum (Maximum Common Subgraph)

Até agora, tanto o isomorfismo entre grafos como a possibilidade de saber se um grafo está contido noutro (subgrafo), não são capazes de determinar se existe algum subconjunto de arcos e vértices comuns partilhado por dois grafos para uma possível quantificação do grau de similaridade que possa haver entre eles. É neste contexto que surge a proposta de determinação do máximo subgrafo comum entre dois grafos [Bunke e Shearer, 1998] para a determinação de uma medida de distância entre grafos. Antes de analisarmos o método proposto neste trabalho, é importante definir o conceito de máximo subgrafo comum: Tendo $\mathbf{G}$, $\mathbf{G_1}$ e $\mathbf{G_2}$, $\mathbf{G}$ é máximo subgrafo comum de $\mathbf{G_1}$ e $\mathbf{G_2}$ denominado $\mathbf{msc(G_1,G_2)}$ se existe isomorfismo de subgrafos de $\mathbf{G}$ para $\mathbf{G_1}$ e de $\mathbf{G}$ para $\mathbf{G_2}$. Um subgrafo comum $\mathbf{G}$ é **máximo** de $\mathbf{G_1}$ e $\mathbf{G_2}$ se não existe nenhum outro subgrafo comum **G'** que possua mais vértices que $\mathbf{G}$.

A medida de distância entre grafos apresentada por [Bunke e Shearer, 1998] é definida pela fórmula (A.1), e possui as seguintes propriedades de uma métrica:

i. $0 \leq dist(G_1, G_2) \leq 1$

ii. $dist(G_1, G_2) = 0 \Leftrightarrow \mathbf{G_1}$ e $\mathbf{G_2}$ são isomorfos entre si.

iii. $dist(G_1, G_2) = dist(G_2, G_1)$

iv. $dist(G_1, G_3) \leq dist(G_1, G_2) + dist(G_2, G_3)$

$$dist(G_1, G_2) = 1 - \frac{|\, mcs(G_1, G_2)\,|}{\max(|\, G_1\,|, |\, G_2\,|)} \qquad \textbf{(A.1)}$$

Onde **|G|** é uma notação abreviada de **|V|** para o número de vértices de um grafo **G(V,E)**.

Apesar de não demonstrarem resultados práticos, a aplicação desta medida pode ser aplicada em diversas áreas (Pesquisa de dados multidimensionais em *Information Retrieval*, reconhecimento de padrões na Visão Computacional, entre outras) desde que seja possível em tempo útil determinar o máximo subgrafo comum de dois grafos. Os autores reconhecem que um grande esforço da comunidade científica tem sido dispensado nesta área para a concretização deste objectivo, uma vez que todos os algoritmos actualmente disponíveis têm um tempo de execução exponencial, afirmando ainda, que num outro trabalho publicado por eles [Shearer et. al, 1997] apresentam um novo algoritmo com o tempo de execução máximo de $O(2^n)$.

A.1.3 Propriedades dos Grafos

Internamente a um grafo existem um conjunto de propriedades estabelecidas que podem auxiliar qualquer pesquisa e possibilidade de focá-la nos elementos principais da estrutura. A seguir, serão apresentadas algumas definições que futuramente serão utilizadas, nomeadamente, a noção de caminho, semi-caminho, distância entre dois vértices, excentricidade, centro de um grafo e classificação do grau de ligação entre os vértices de um grafos.

Num grafo dirigido, o número de arcos que divergem ("saem") de um vértice *v* é denominado **grau de saída** de *v*, enquanto que **grau de entrada** de *v* é o número de arcos que convergem ("entram") a este vértice. Um vértice é classificado **sumidouro** se possuir grau de saída nulo, ou **fonte** se for nulo o seu grau de entrada. A soma dos graus de entrada de todos os vértices do grafo é obrigatoriamente igual a soma dos graus de saída.

Um **caminho** é uma sequência de vértices $v_1, v_2, ..., v_n$, sem repetições, tal que existam os arcos $v_1 \rightarrow v_2, v_2 \rightarrow v_3, \ldots, v_{n-1} \rightarrow v_n$. Este caminho tem origem em v_1, atravessa os vértices $v_2, v_3, \ldots, v_{n-1}$ e termina em v_n. A **distância** de um caminho é a soma do número de arcos presentes neste. Como caso especial, um vértice *v* possui um caminho até a si próprio de distância 0. Um **semi-caminho** é um caminho onde não é respeitada a direcção dos arcos, ou seja, é também uma sequência de vértices $v_1, v_2, ..., v_n$, sem repetições, tal que existam os arcos

$|v_1 \rightarrow v2 \vee v_2 \rightarrow v_1|, |v_2 \rightarrow v_3 \vee v_3 \rightarrow v_2|, \ ..., |v_{n-1} \rightarrow v_n \vee v_n \rightarrow v_{n-1}|$. Num grafo dirigido, o seu **grafo subjacente** é obtido pela retirada das direcções dos arcos.

A pesquisa de caminhos entre vértices pode ser feita basicamente por uma procura em **largura** ou em **profundidade** [Cormen et. al, 1989]. A procura em largura é a maneira mais simples de encontrar um caminho e é a base de muitos algoritmos para grafos mais complexos, nomeadamente o algoritmo de Dijkstra para descobrir o caminho mais curto a partir de um vértice de origem determinado. Dado um grafo **G(V,E)** representado por uma lista de adjacências ***Adj*** e um vértice de origem ***v***, a procura em largura analisa os arcos de **G** para descobrir todos os vértices que são acessíveis desde ***v***, calculando a distância para cada um. Este algoritmo (apresentado no Quadro A.1) explora gradualmente todos os vértices encontrados a uma distância *k* para posteriormente investigar todos os que estejam a uma distância *k+1*. Mantém um registo dos arcos visitados com a seguinte classificação para cada um: não-visitado, visitado e já investigado (onde todos os adjacentes já foram visitados). De seguida é apresentada uma adaptação do algoritmo de procura em largura de todos os caminhos em **G** a partir de um vértice de origem ***v***, onde ***d[w]*** é a distância do caminho desde ***v*** até ***w*** e ***π[w]*** armazena o antecedente do vértice ***w***, ou seja, o vértice adjacente a ***w*** pelo qual, ao longo do caminho com origem em ***v*** se chegou ***w***. É assumido que exista uma fila ***Q*** que obedeça a disciplina FIFO na organização e disponibilização de dados, onde são armazenados novos dados através da função **põe_na_fila(*Q*, *v*)** e **retira_da_fila(*Q*)**.

```
Procura_Em_Largura(G, v)
 para cada vértice u ∈ V – {v}
      fazer classificação[u]← não-visitado
            d[u] ← ∞
            π[u] ← NIL
 classificação[v] ← visitado
 d[v] ← 0
 π[v] ← NIL
 põe_na_fila(Q,v)
 enquanto Q ≠ 0
      fazer u ← retira_da_fila(Q)
            para cada w ∈ Adj[u]
                  fazer se classificação[w] = não-visitado
                        então classificação[w] ← visitado
                              d[w] ← d[u] + 1
                              π[w] ← u
                              põe_na_fila(Q,v)
            classificação[u] ← investigado
```

Quadro A.1: Algoritmo de procura em largura para descoberta de caminhos em grafos. Adaptado de [Cormen et. al, 1989]

Este algoritmo tem um tempo de execução proporcional à representação do grafo por listas de adjacências $O(|V|+|E|)$.

Numa procura em profundidade (ver Quadro A.2) é explorado cada um dos vértices adjacentes do vértice ***w***, tal que (***v,w***) antes de serem analisados os outros arcos que saem de ***v***. Neste algoritmo, tal como na procura em largura, a representação para o grafo utilizada é uma lista de adjacências e ao invés da distância do caminho percorrido, são registados instantes temporais referentes à primeira passagem por um vértice em **início[*v*]** (quando é visitado pela primeira vez) e ao término da pesquisa neste vértice em **fim[*v*]** (ou seja, todos os seus adjacentes já foram explorados). No entanto, ao contrário do tipo de procura anterior, onde o antecedente de um vértice ***π[w]*** era o adjacente anterior no caminho percorrido desde a origem, ou o seu "irmão", neste caso pelo modo recursivo como o grafo é percorrido, o antecedente armazenado em ***π[w]*** é o vértice imediatamente "superior" ou "pai" que seja adjacente a ***w,*** transmitindo uma determinada hierarquia em **árvore** na profundidade da pesquisa.

```
Procura_Em_Profundidade(G)
 para cada vértice u ∈ V
          fazer classificação[u]← não-visitado
                π[u] ← NIL
 tempo ← 0
 para cada vértice u ∈ V
          fazer se classificação[u] = não-visitado
                então Visitar_Em_Profundidade(u)


Visitar_Em_Profundidade(u)
 classificação[u]← visitado
 início[u] ← tempo ← tempo + 1
 para cada v ∈ Adj[u]
          fazer se classificação[v] = não-visitado
                então π[v] ← u
                      Visitar_Em_Profundidade(v)
 classificação[u]← investigado
 fim[u] ← tempo ← tempo + 1
```

Quadro A.2: Algoritmo de procura em profundidade para descoberta de caminhos em grafos. Adaptado de [Cormen et. al, 1989]

Uma vez que a função **Visitar_Em_Profundidade** é chamada apenas uma vez para cada vértice, $|V|$ vezes, e depois de chamada percorre todos os arcos adjacentes ao vértice em questão, $|E|$, este algoritmo tem um tempo de execução mínimo de $\Theta(|V|+|E|)$. Enquanto que a procura em largura é mais adequada para encontrar o caminho mais curto entre dois vértices, a procura em profundidade permite estudar em detalhe a estrutura do grafo de modo que é possível fazer uma analogia a uma árvore ao modo de procura. À medida que o algoritmo progride, uma ou mais árvores são dedutíveis a partir da informação temporal registada. Informação esta, que permite classificar cada um dos vértices de acordo com uma hierarquia em árvore com base no intervalo entre o instante em que o vértice foi encontrado até ao momento em que deixou de ser explorado

(porque já foi completamente investigado). Seja um grafo **G(V,E)** para quaisquer dois vértices ***u*** e ***v*** somente uma das três condições poderá existir:

i. os intervalos **[início[*u*],fim[*u*]]** e **[início[*v*],fim[*v*]]** são completamente disjuntos;

ii. o intervalo **[início[*u*],fim[*u*]]** está inteiramente contido no intervalo **[início[*v*],fim[*v*]]**, sendo o vértice ***u*** descendente de ***v***;

iii. a situação oposta a anterior, o intervalo **[início[*v*],fim[*v*]]** está inteiramente contido no intervalo **[início[*u*],fim[*u*]]**, sendo o vértice ***v*** descendente de ***u***.

Um vértice é ainda considerado de topo (**raiz**) de uma árvore obtida pela pesquisa em profundidade se o seu intervalo de procura não estiver contido em nenhum intervalo de outro vértice.

Além de percorrer caminhos num grafo dirigido, existem outras propriedades que podem ser úteis para a determinação da(s) principal(is) região(ões) existente(s) no grafo: Denomina-se **excentricidade de um vértice v** pertencente a **V** ao valor máximo de distância do menor caminho entre **v** e **w**, para todo **w** pertencente a **V** (fórmula A.2). Caso não seja possível estabelecer um caminho entre dois vértices, é considerada uma distância infinita entre eles (∞). Através desta definição, é possível determinar **centro de um grafo G** como sendo o subconjunto de vértices com excentricidade mínima (fórmula A.3).

$$excent(v) = {}_{\forall w \in V} \max\{\min length(v, w)\} \qquad \textbf{(A.2)}$$

$$center(V) = \{\forall v \in V \wedge \min(excent(v))\} \qquad \textbf{(A.3)}$$

Nem sempre é possível estabelecer um caminho entre todos os vértices de um grafo, como é o caso do exemplo da Figura A.1, onde só é possível encontrar semi-caminhos para cada vértice em relação a todos os outros, consequentemente possuindo todos excentricidade infinita. Neste caso, dizemos que um grafo dirigido é **fracamente conexo** pois apesar de não existir partes do grafo isoladas, sem ligações com os outros vértices do grafo, a direcção dos arcos não permite estabelecer caminhos entre todos os vértices (só é possível ligar alguns vértices através de semi-caminhos). Quando todos os vértices possuem caminhos entre si, ou seja, a partir de qualquer vértice é possível chegar a todos os outros seguindo a direcção dos arcos, temos um grafo dirigido **fortemente conexo**.

Uma **componente forte** de um grafo dirigido é o seu maior subgrafo fortemente conexo, ou seja, não existe nenhum vértice fora deste subgrafo que seja fortemente conexo a todos os vértices aí presentes (pois neste caso, teria que fazer parte da componente forte do grafo). De forma análoga, uma **componente fraca** é o maior subgrafo fracamente conexo que exista dentro do grafo.

É proposto em [Cormen et. al, 1989] um algoritmo linear (ver quadro A.3) para encontrar os componentes fortes contidos num grafo dirigido **G(V,E)** que utiliza o conceito de **grafo transposto** de **G** como sendo o grafo $\mathbf{G^T(V, E^T)}$ onde $\mathbf{E^T=\{(v,w):(w,v) \in E\}}$, ou seja, $\mathbf{E^T}$ é conjunto dos arcos de **G** com as suas direcções invertidas. Utilizando uma lista de adjacências para representar um grafo dirigido, o tempo necessário para obter $\mathbf{G^T}$ é $O(|V|+|E|)$, enquanto no caso da escolha de matrizes de adjacência, o tempo é consideravelmente maior, $O(|V|^2)$. É interessante notar que os componentes fortes de **G** são preservados em $\mathbf{G^T}$: existe um caminho de ***v*** para ***w*** e vice-versa em **G** se e somente se estes estão ligados por um caminho em ambas as direcções em $\mathbf{G^T}$.

```
Procura_De_Componentes_Fortes(G)
 Procura_Em_Profundidade(G)   //para determinar fim[u] para cada //vértice u

 calcular G^T
 Procura_Em_Profundidade(G^T)  //considerando no ciclo principal os //vértices
                               em ordem decrescente de fim[u]
 Cada componente forte é formada pela árvores obtidas nesta última pesquisa
```

Quadro A.3: Algoritmo de procura de componentes fortes com base na procura em profundidade em grafos. Adaptado de [Cormen et. al, 1989].

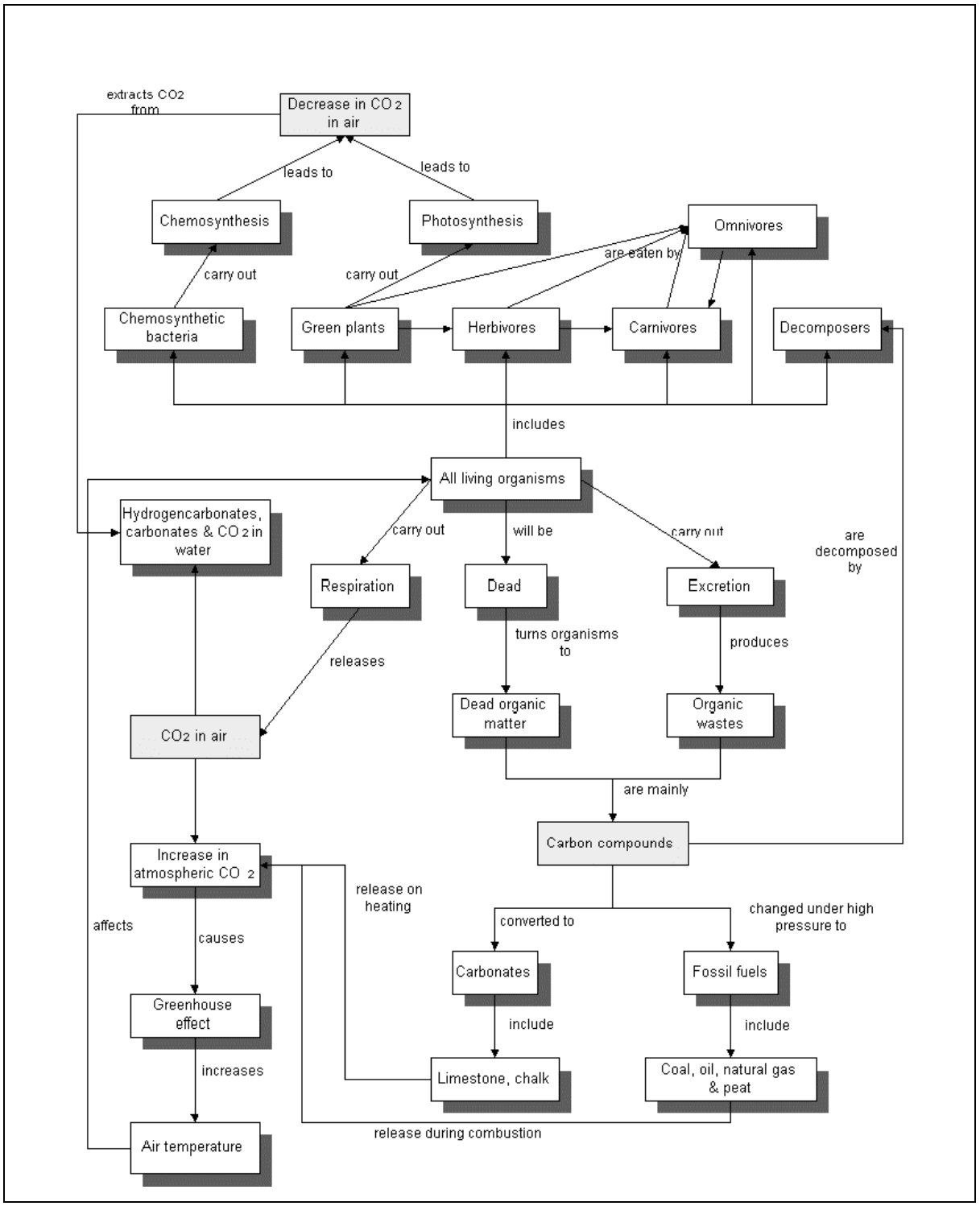
A.2. Mapas Conceptuais Utilizados

De seguida, é apresentado o conjunto de mapas conceptuais utilizados para a fase de testes dos algoritmos de contextualização e comparação semântica.

A.2.1. Mapas Conceptuais da Internet

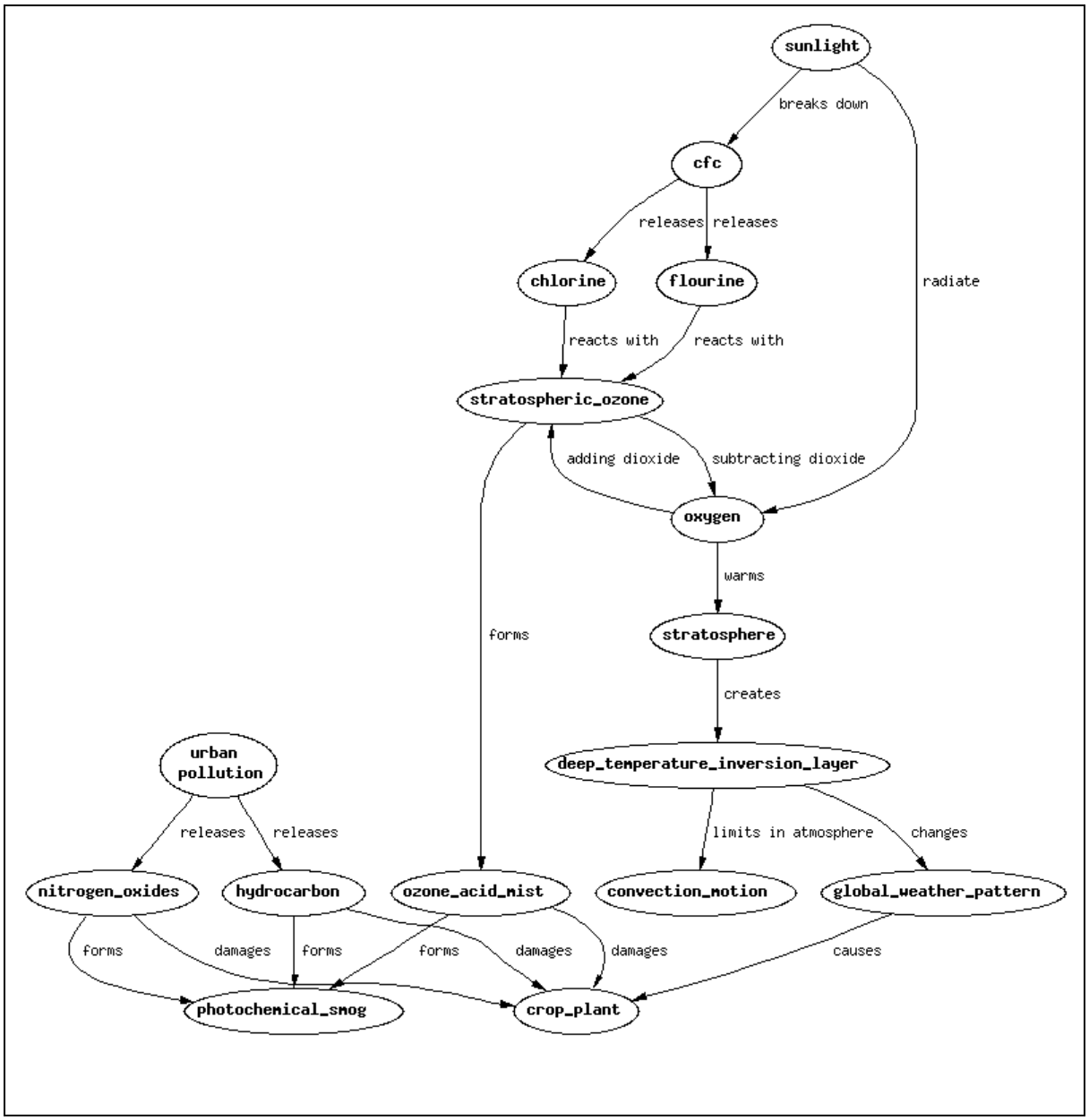
Seguem-se os mapas conceptuais recolhidos a partir da *WordWideWeb* [Leung,2003].

Ciclo do Dióxido de Carbono

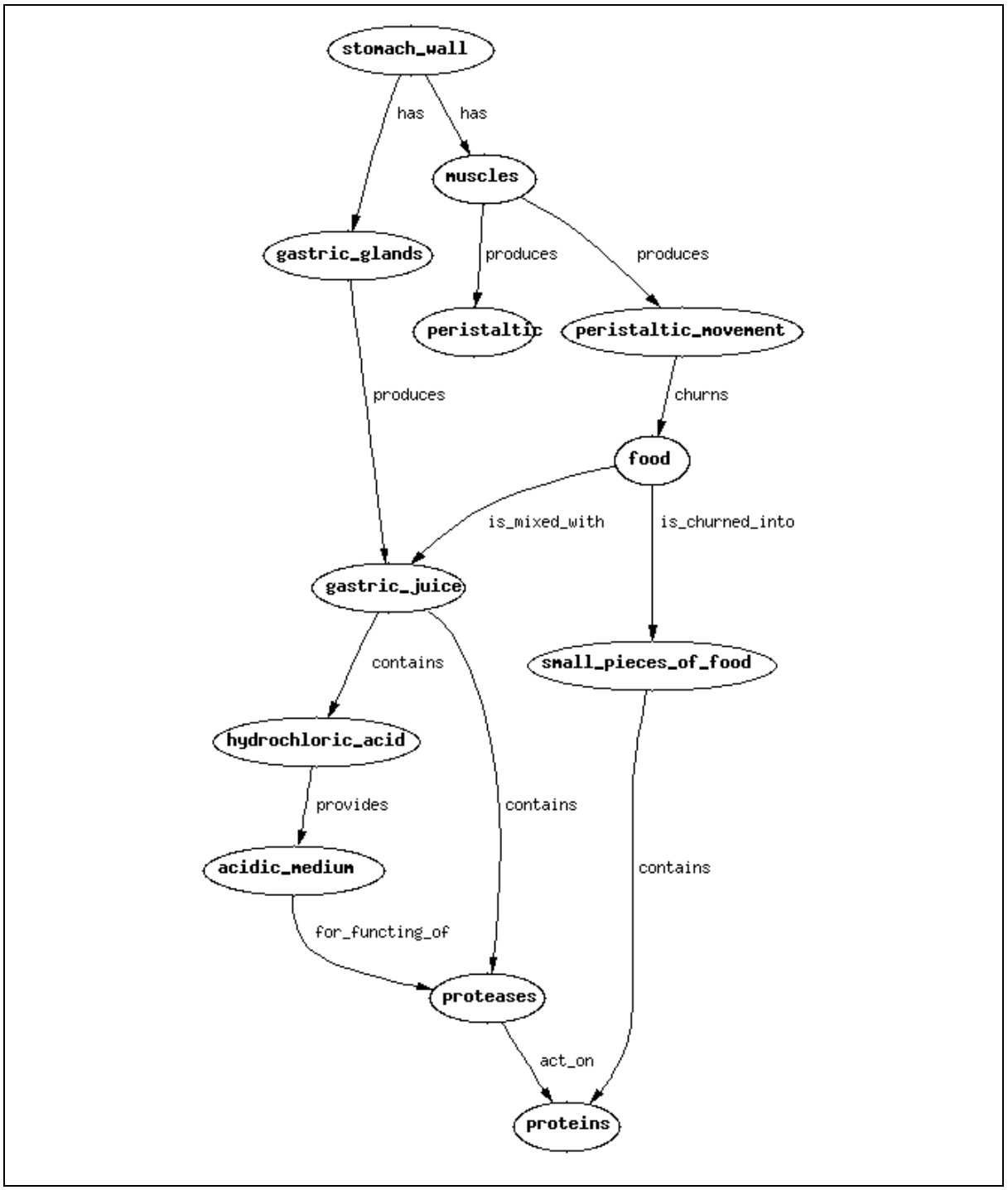


Anexo 2. Mapas Conceptuais Utilizados

Ciclo dos Clorofluorbonos (CFC)

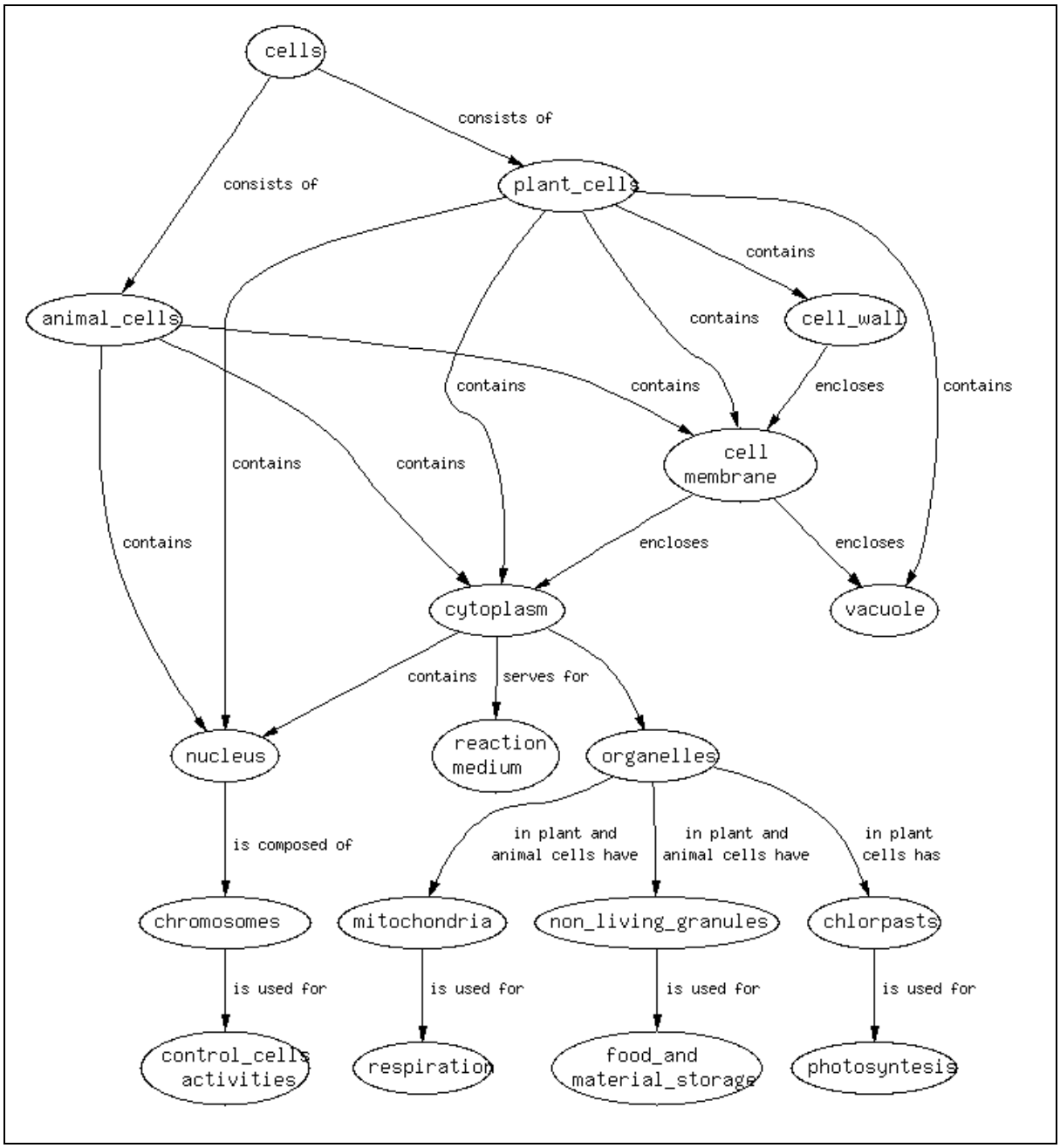


Digestão Humana

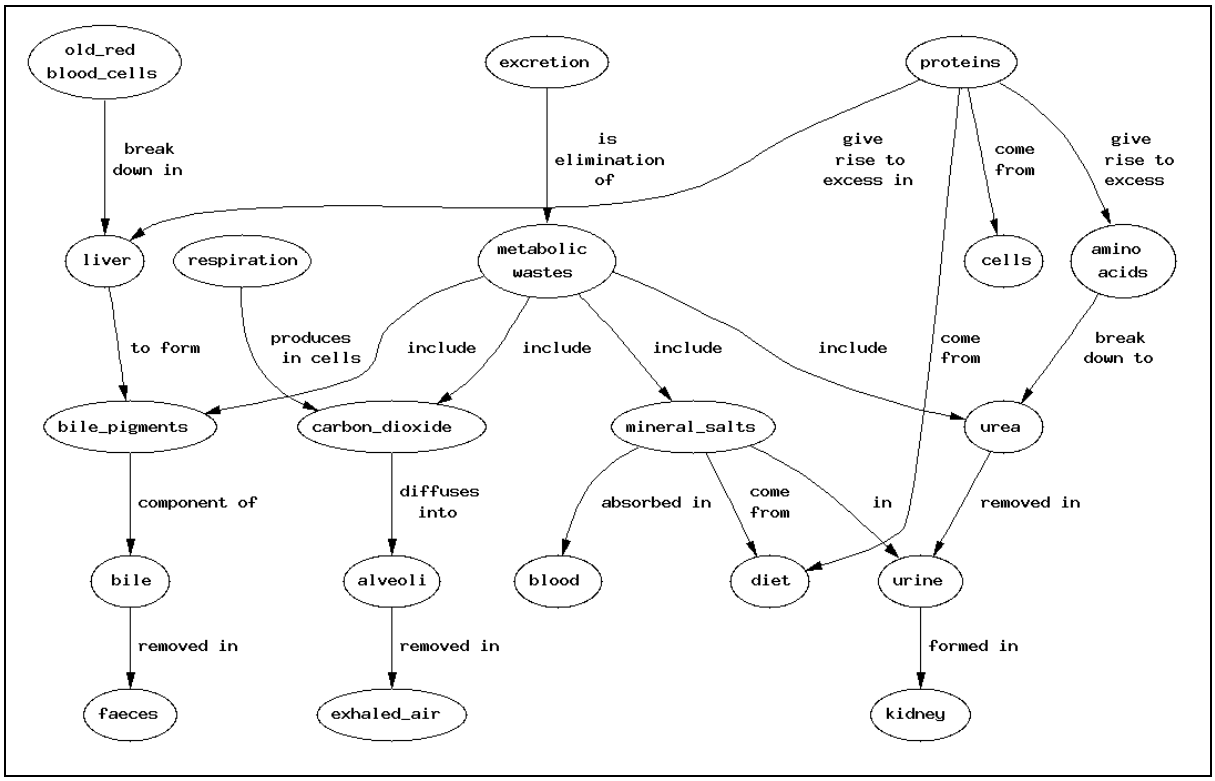


Anexo 2. Mapas Conceptuais Utilizados

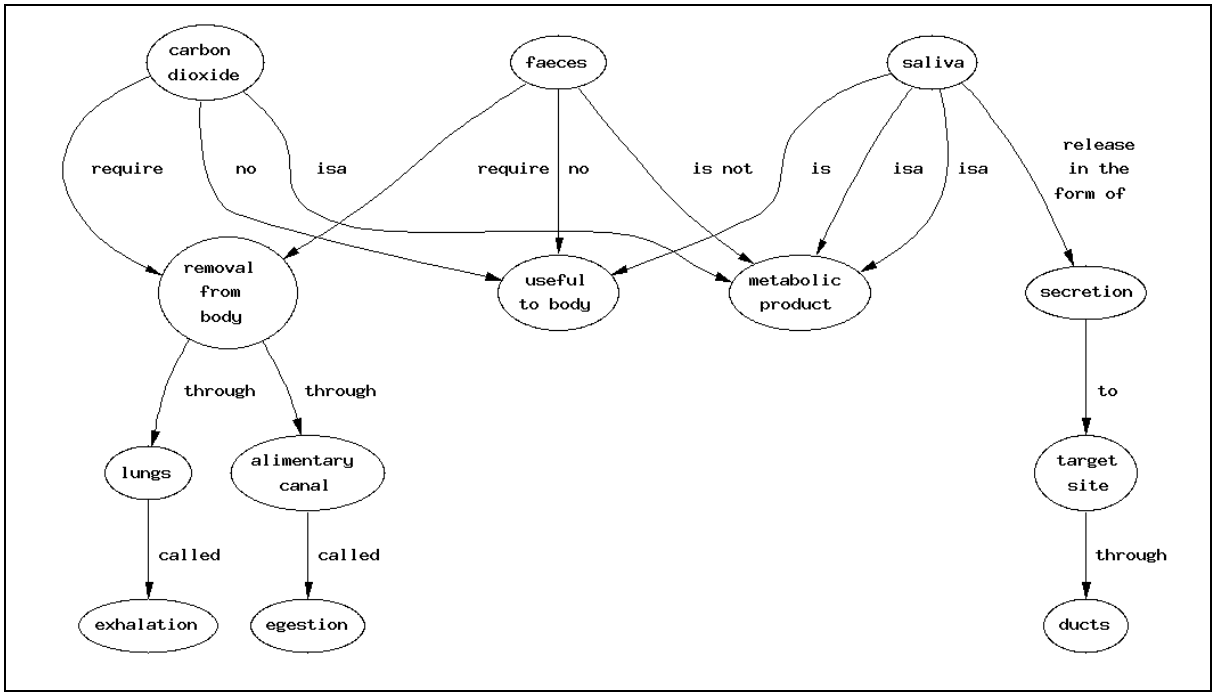
Estrutura Interna das Células



Excrecção Humana (1)

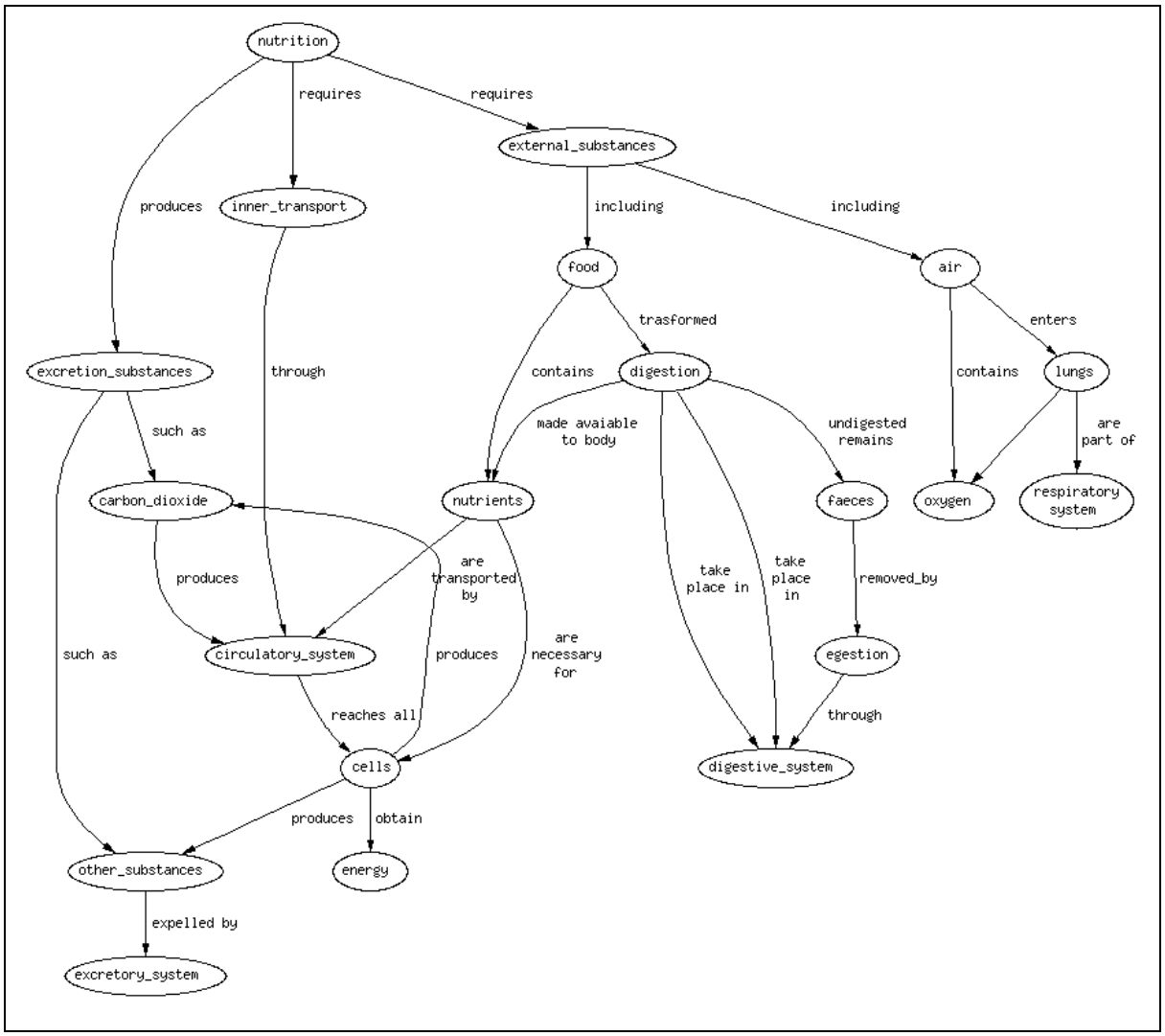


Excrecção Humana (2)

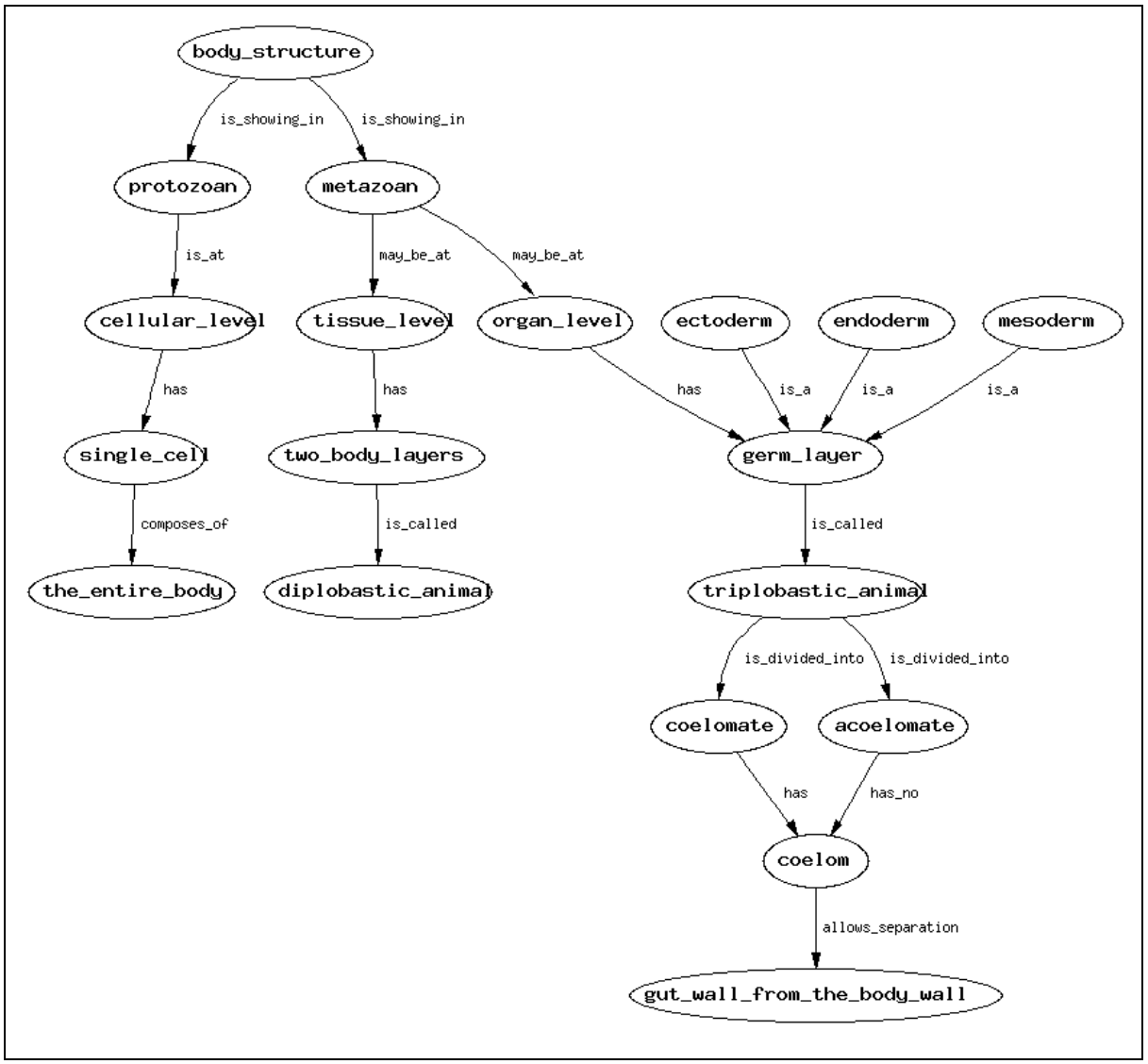


Anexo 2. Mapas Conceptuais Utilizados

Nutrição Humana

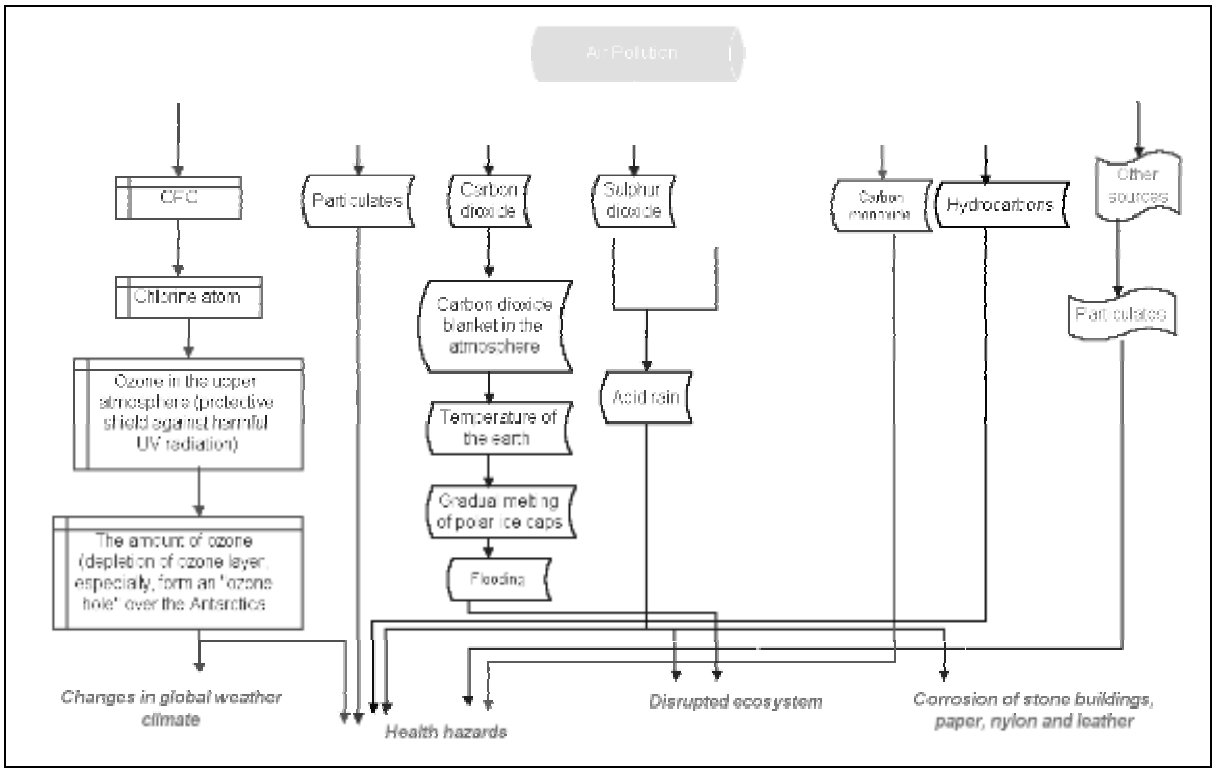


Organização Celular

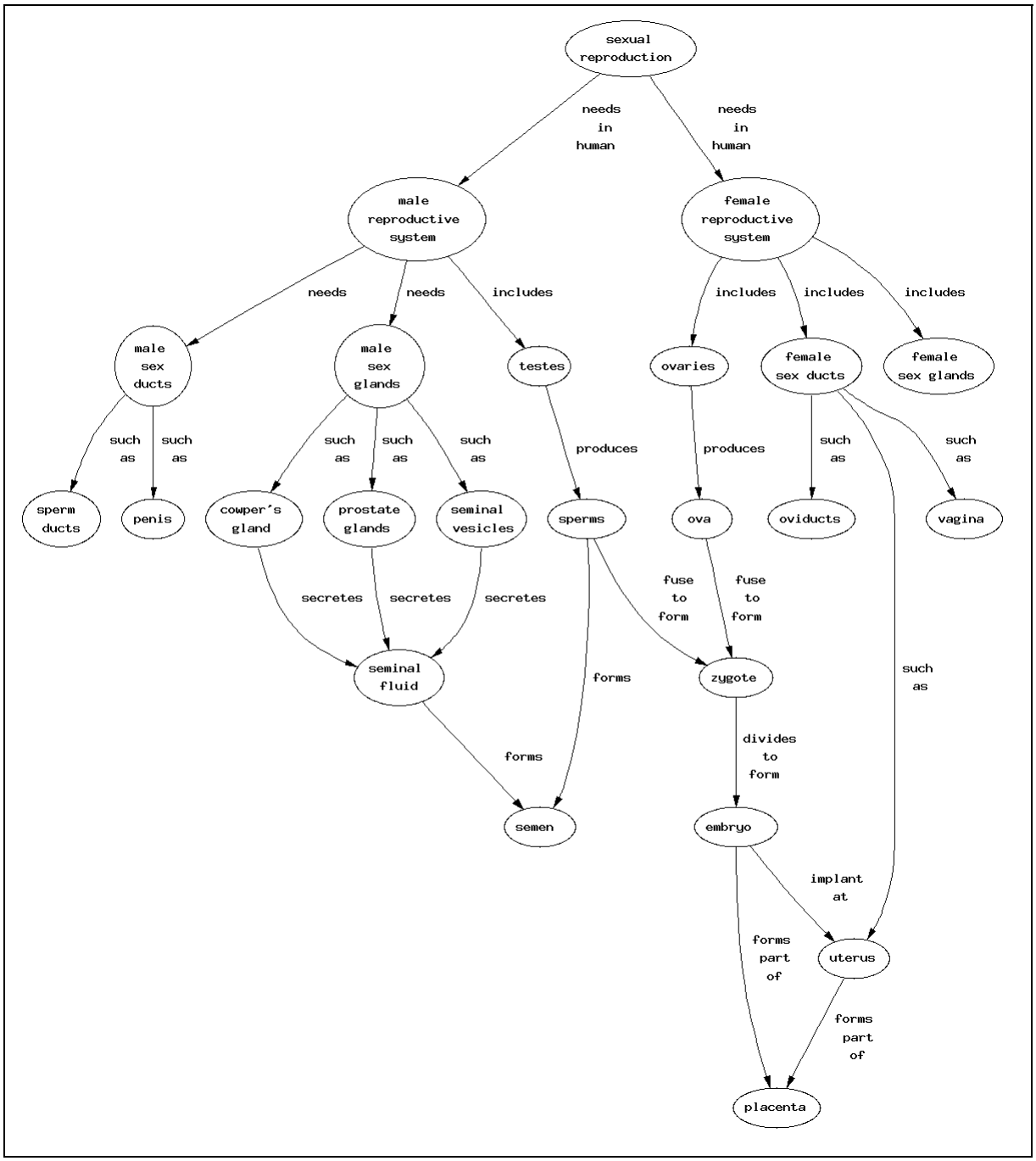


Anexo 2. Mapas Conceptuais Utilizados

Poluição do Ar

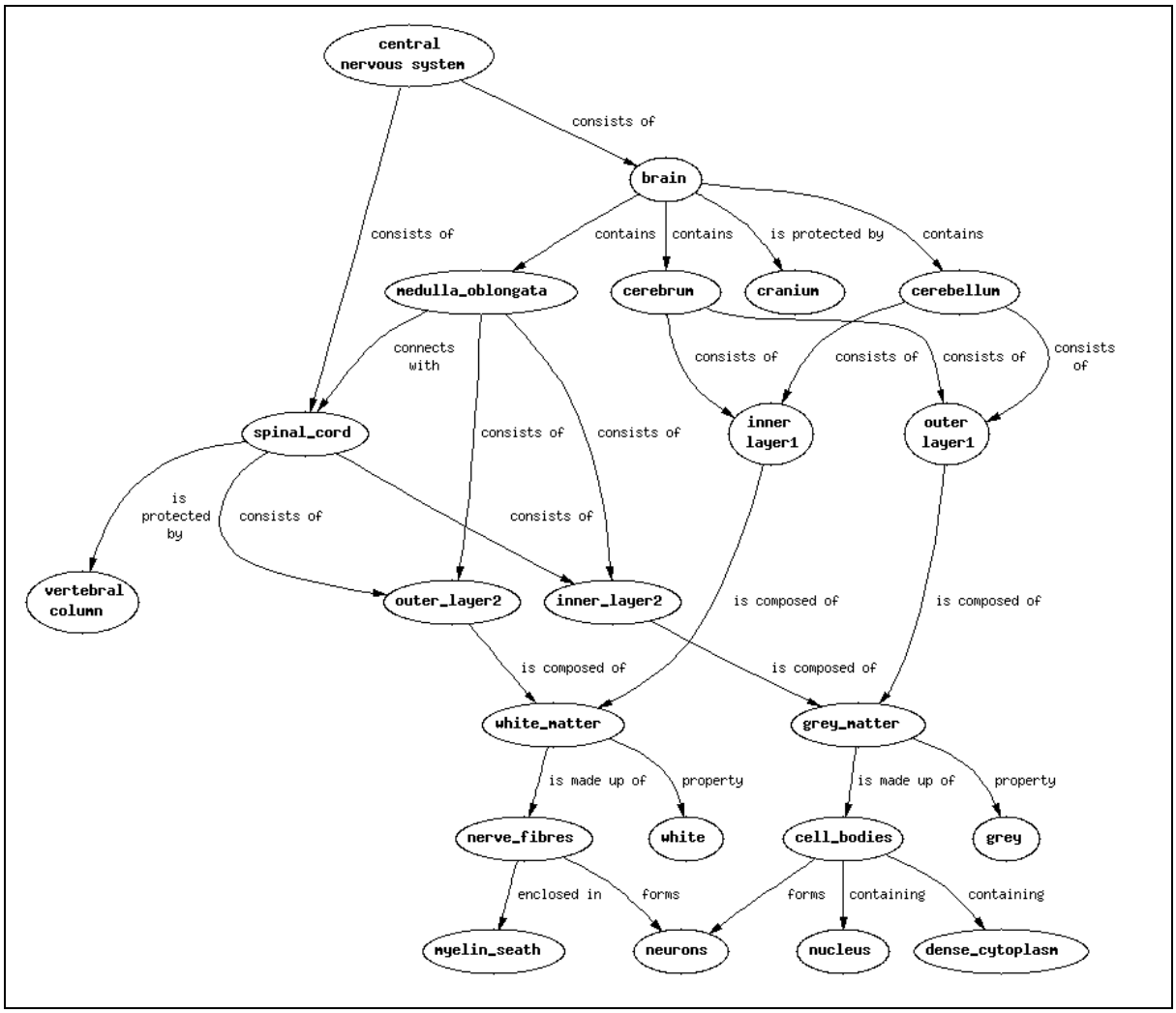


Reprodução Humana

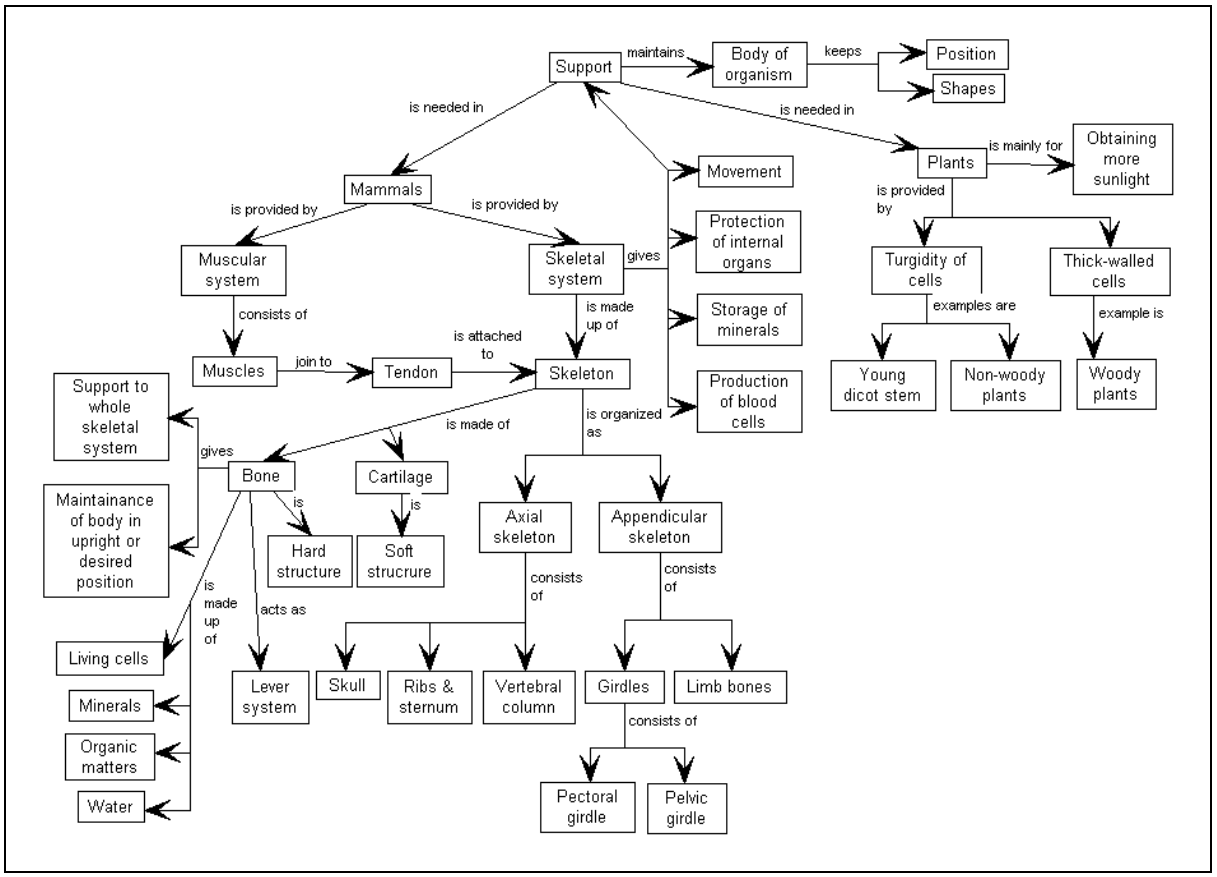


Anexo 2. Mapas Conceptuais Utilizados

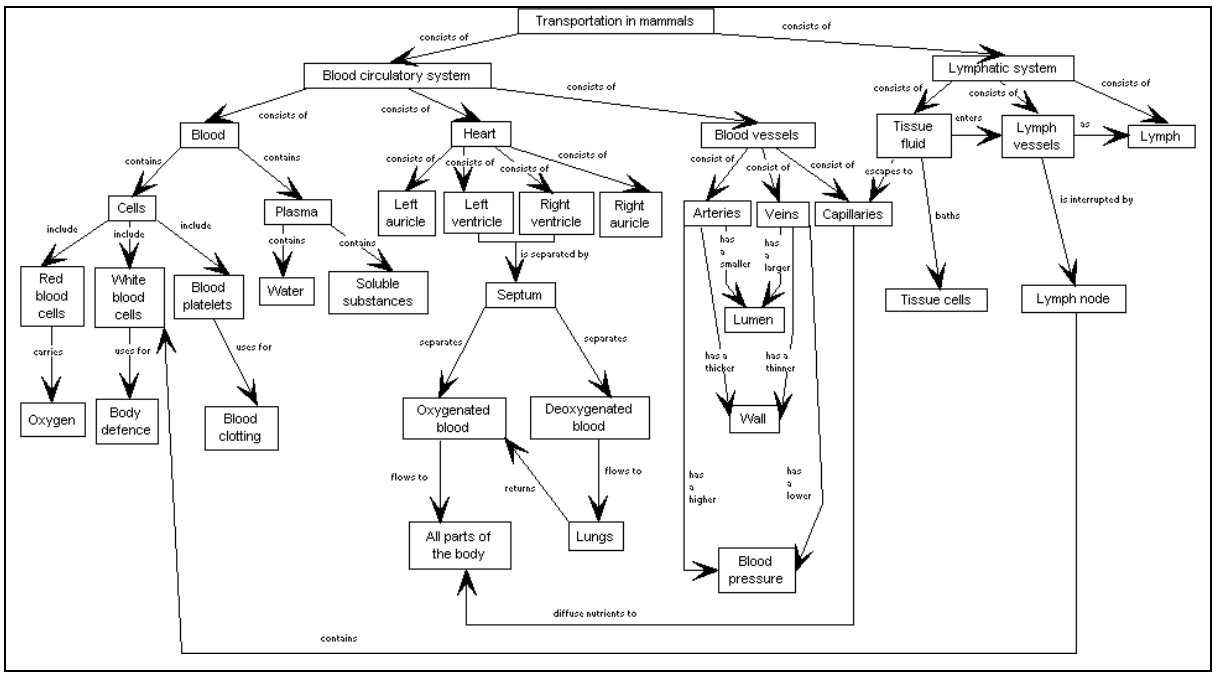
Sistema Nervoso Central



Suporte Corporal

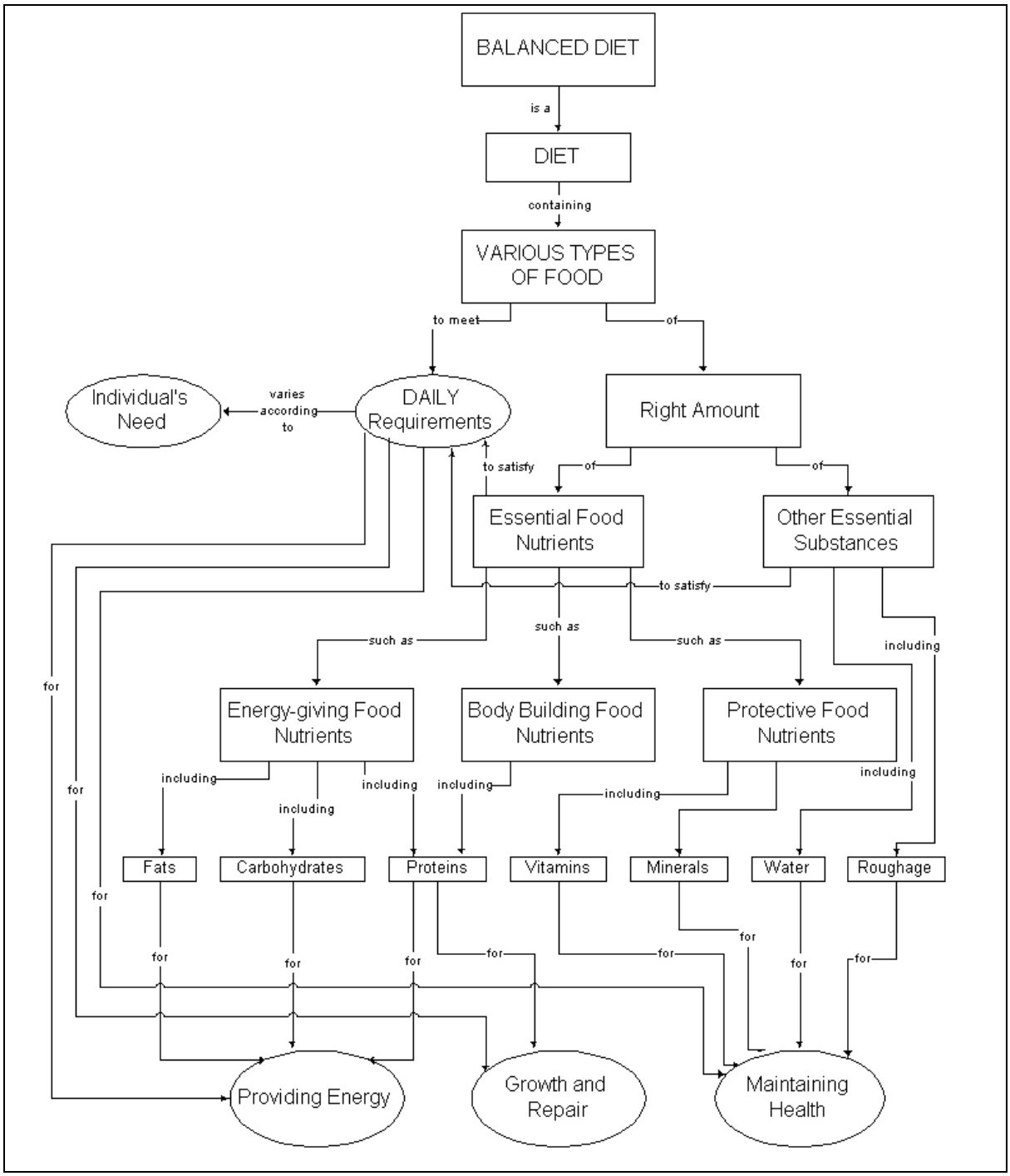


Transporte de Fluídos nos Mamíferos



Anexo 2. Mapas Conceptuais Utilizados

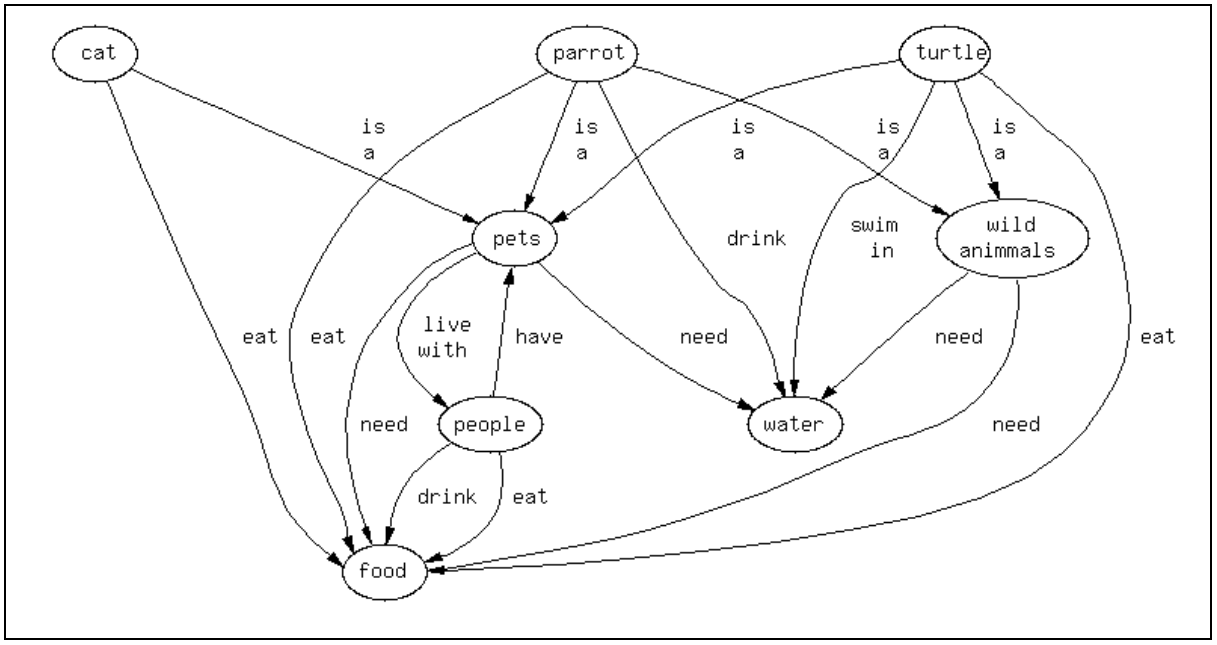
Uma dieta balanceada



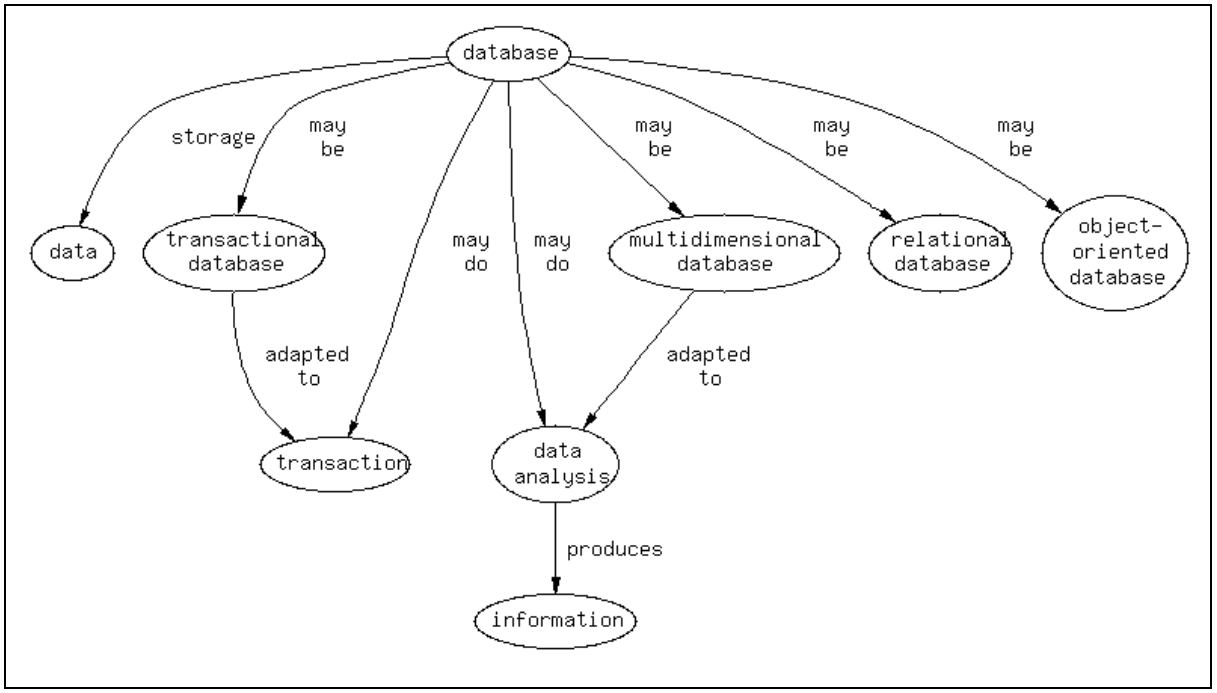
A.2.2. Mapas Conceptuais de Colaboradores

Os mapas conceptuais apresentados a seguir foram construídos por diversos colaboradores deste trabalho.

Animais

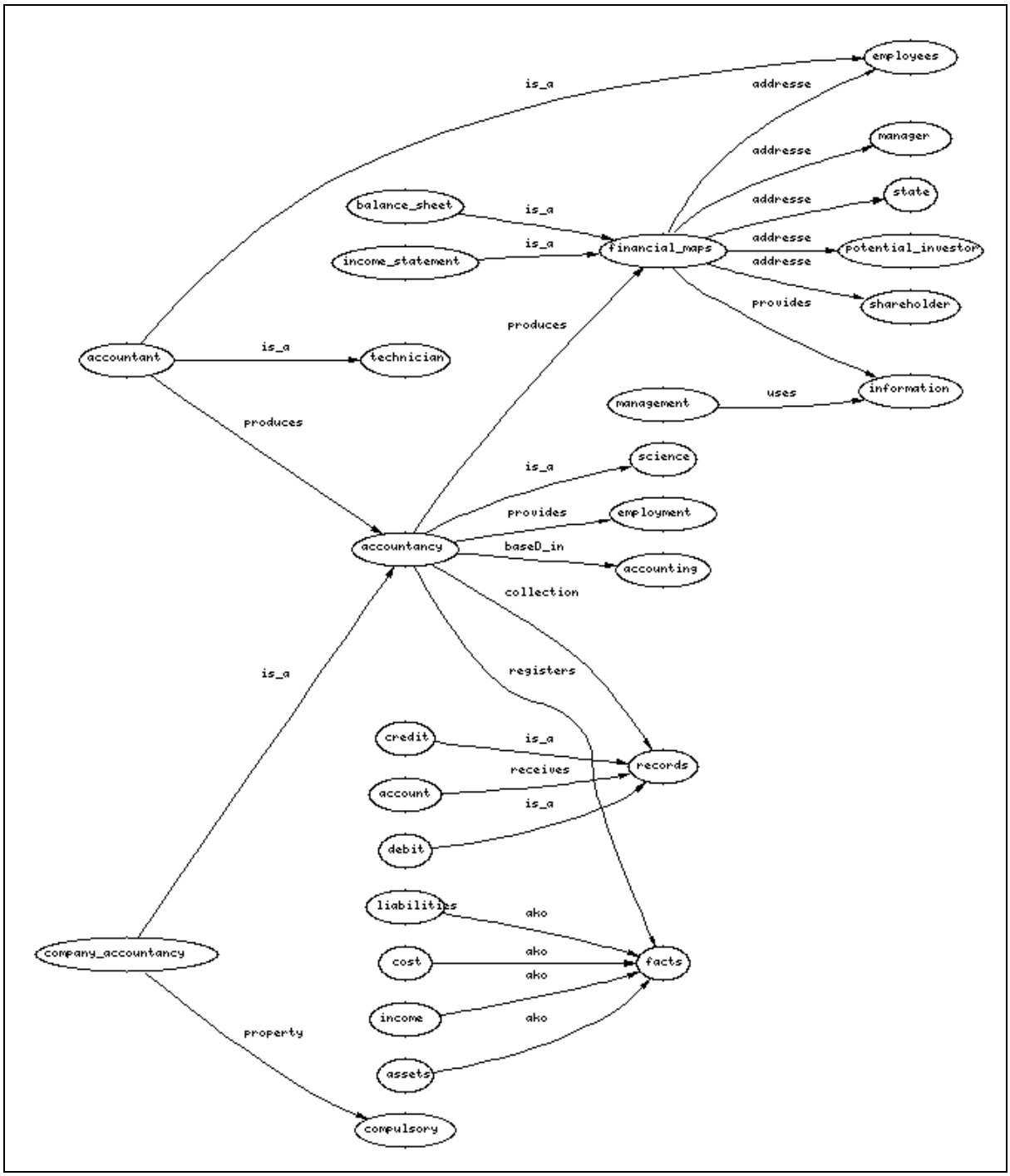


Bases de Dados

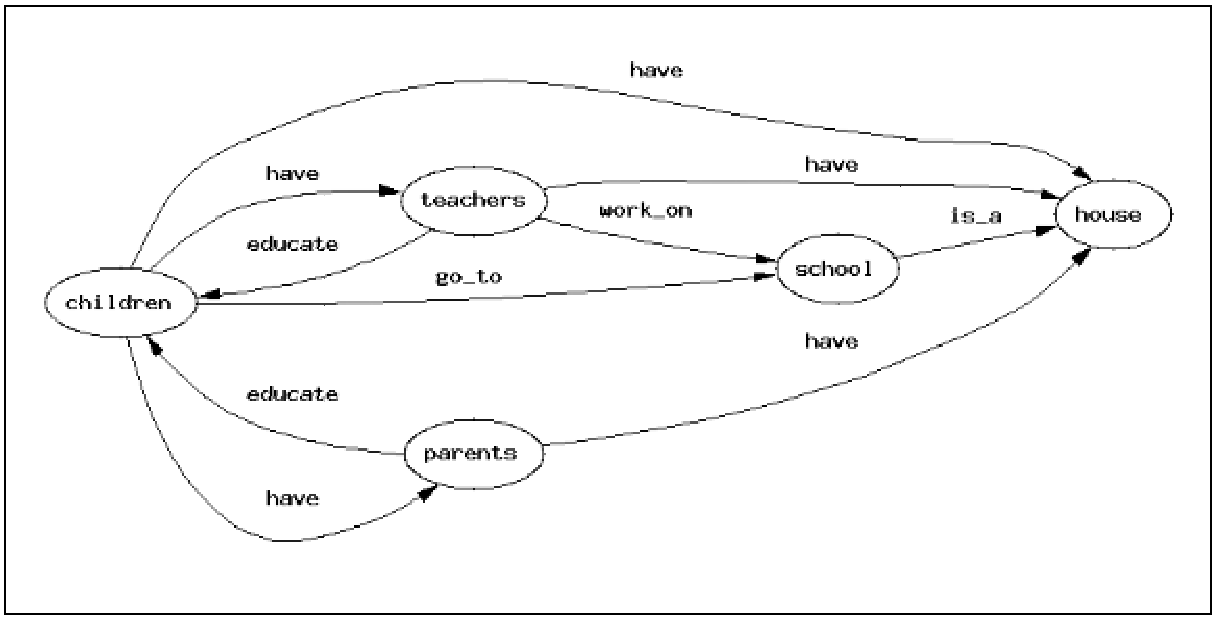


Anexo 2. Mapas Conceptuais Utilizados

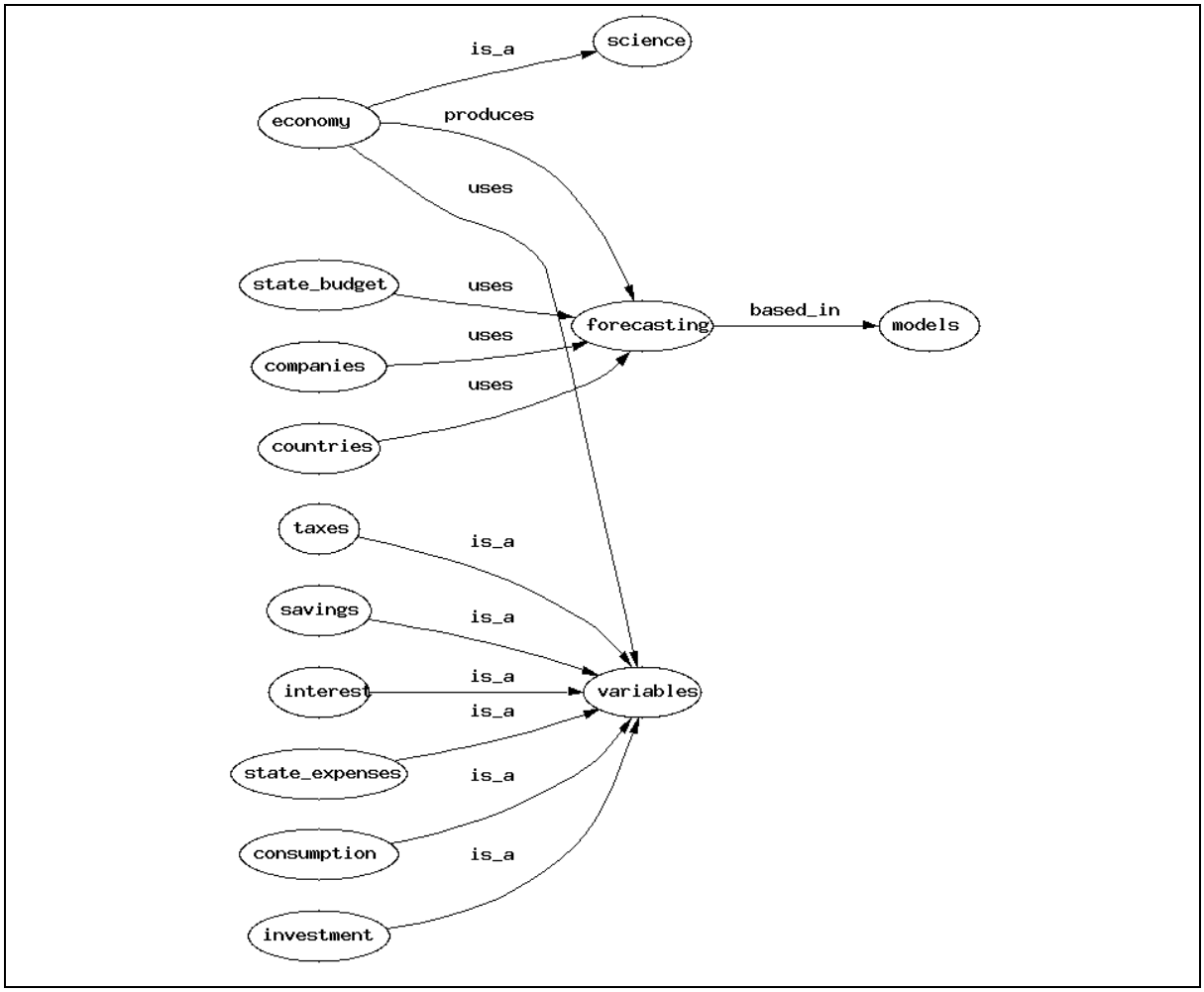
Contabilidade



Educação

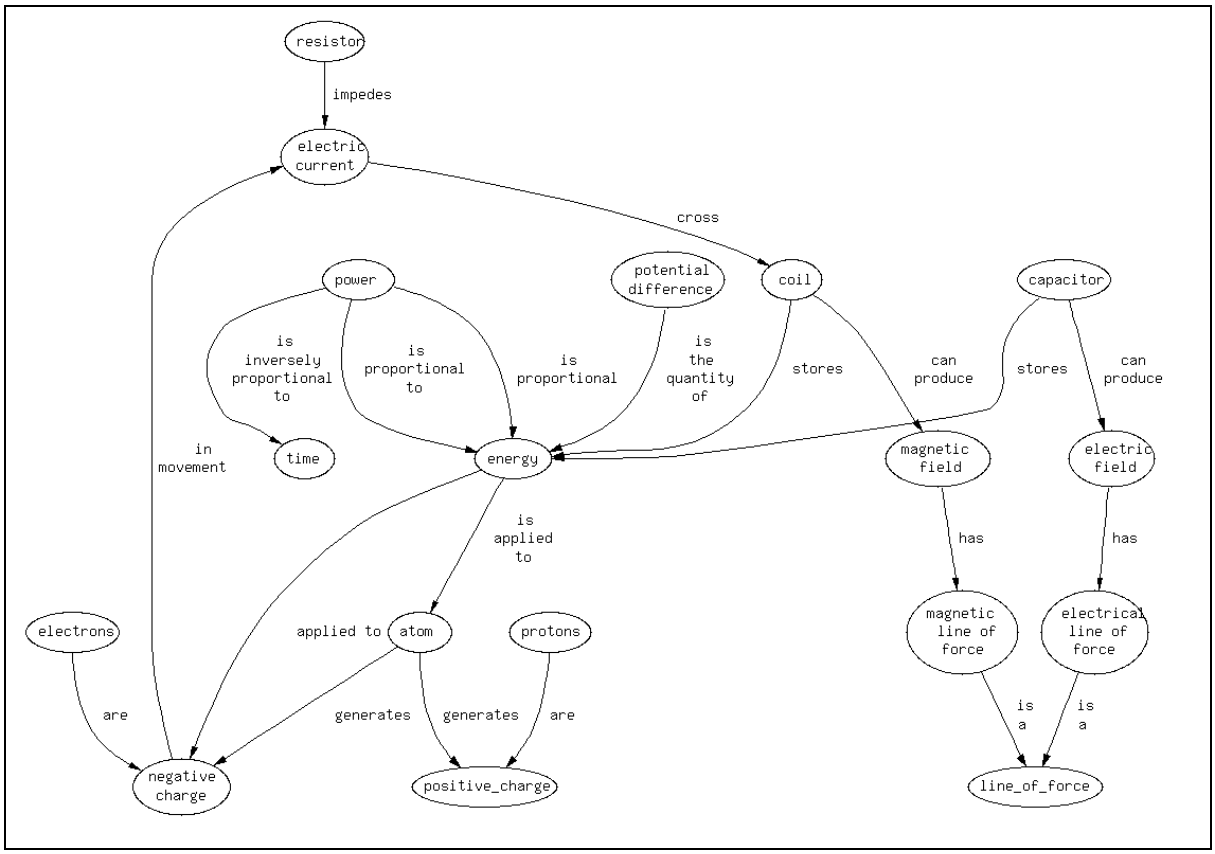


Economia

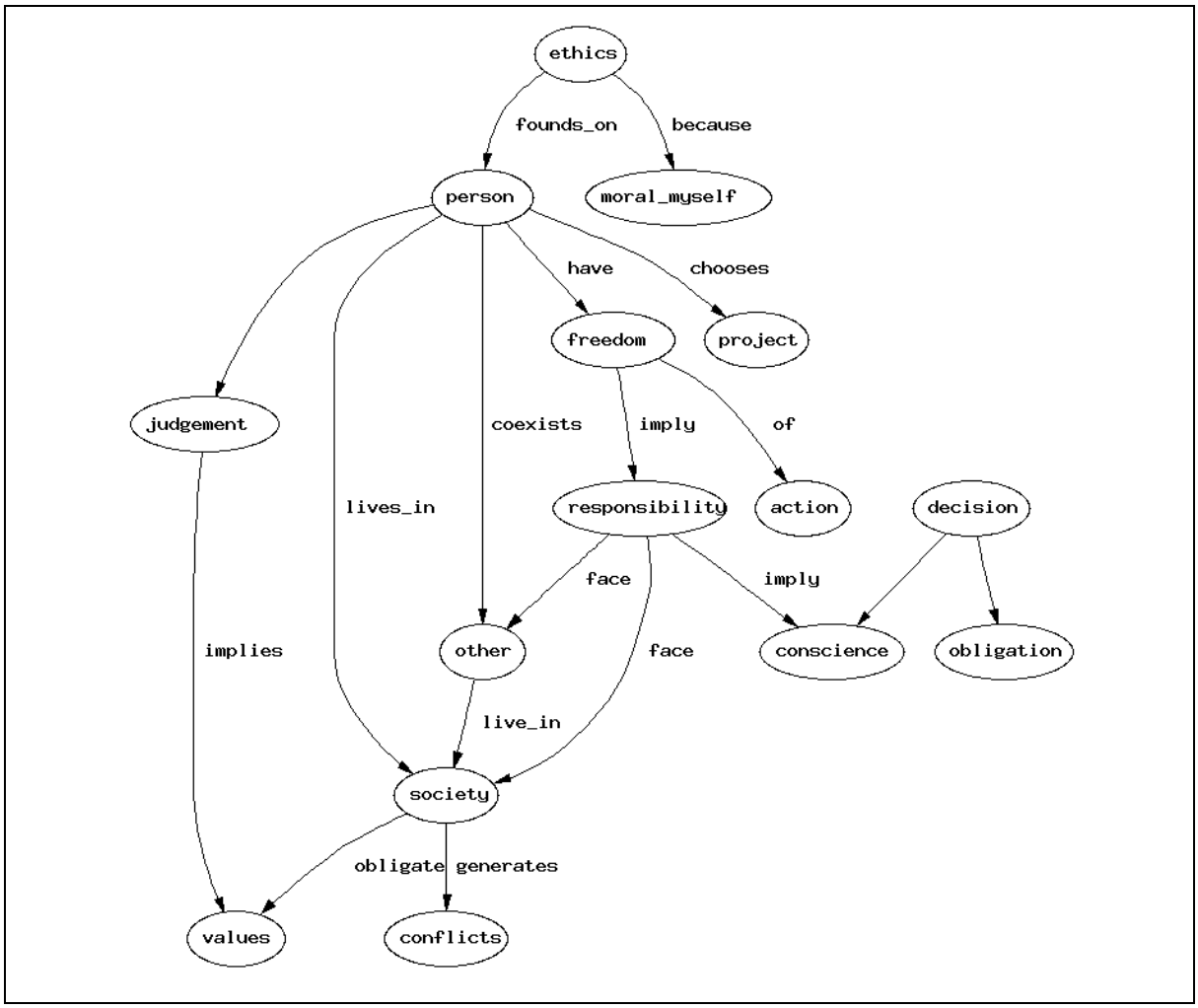


Anexo 2. Mapas Conceptuais Utilizados

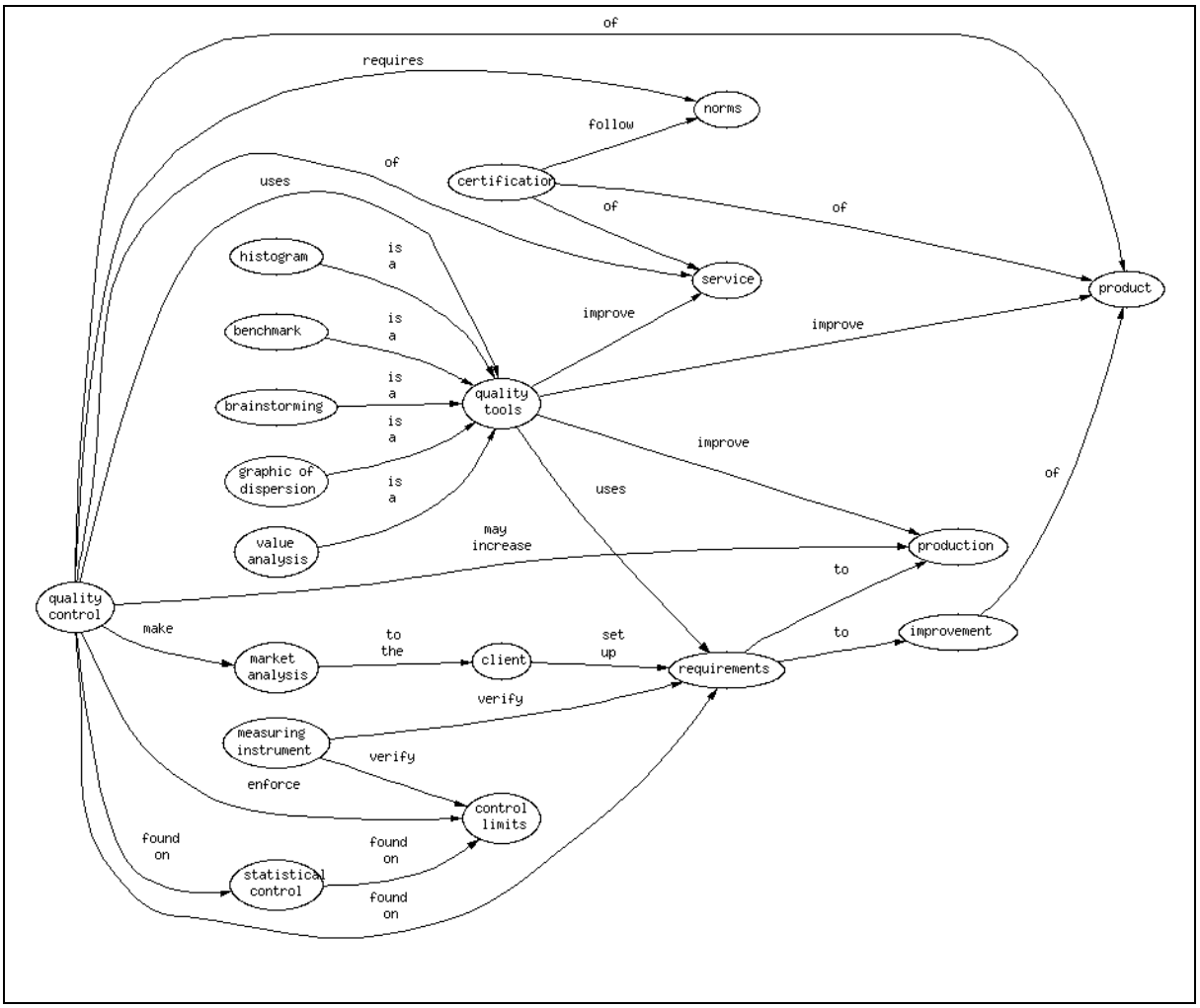
Electricidade



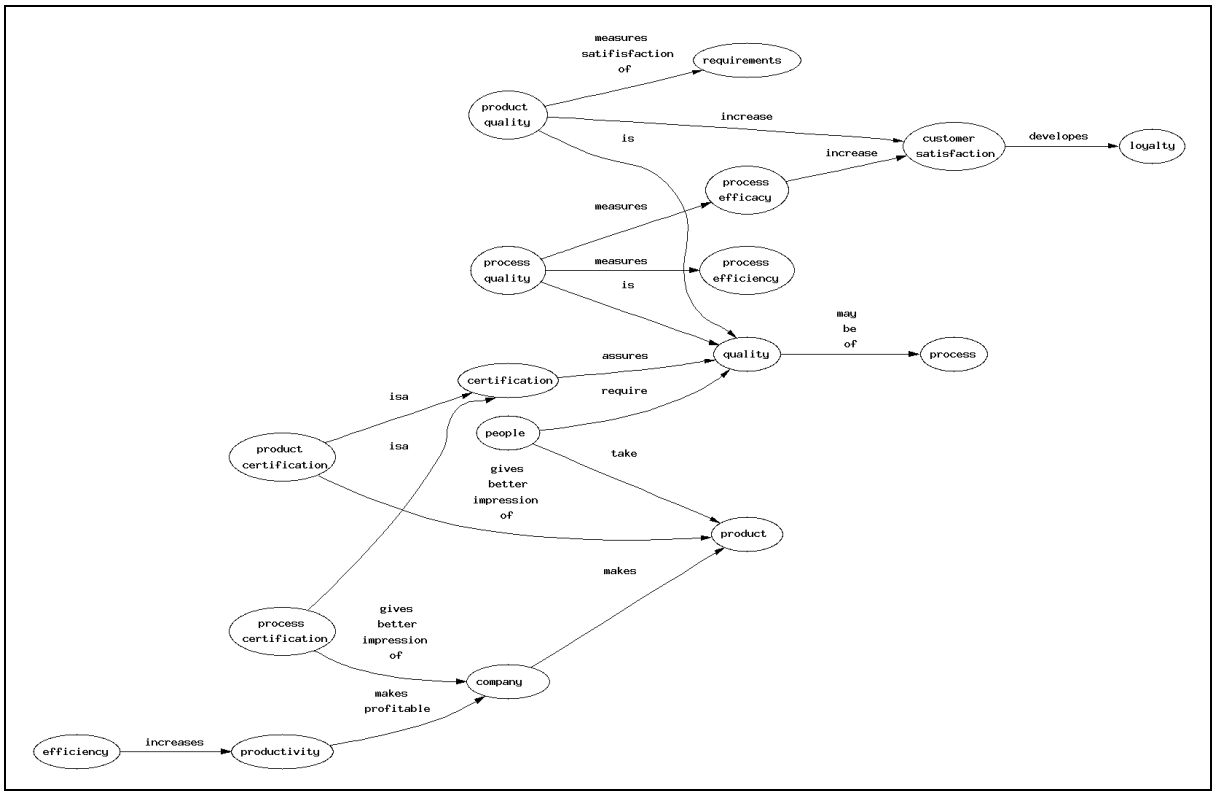
Etica


ethics
founds_on
because
person
moral_myself
have
chooses
freedom
project
judgement
coexists
imply
of
lives_in
responsibility
action
decision
face
imply
implies
other
face
conscience
obligation
live_in
society
obligate
generates
values
conflicts

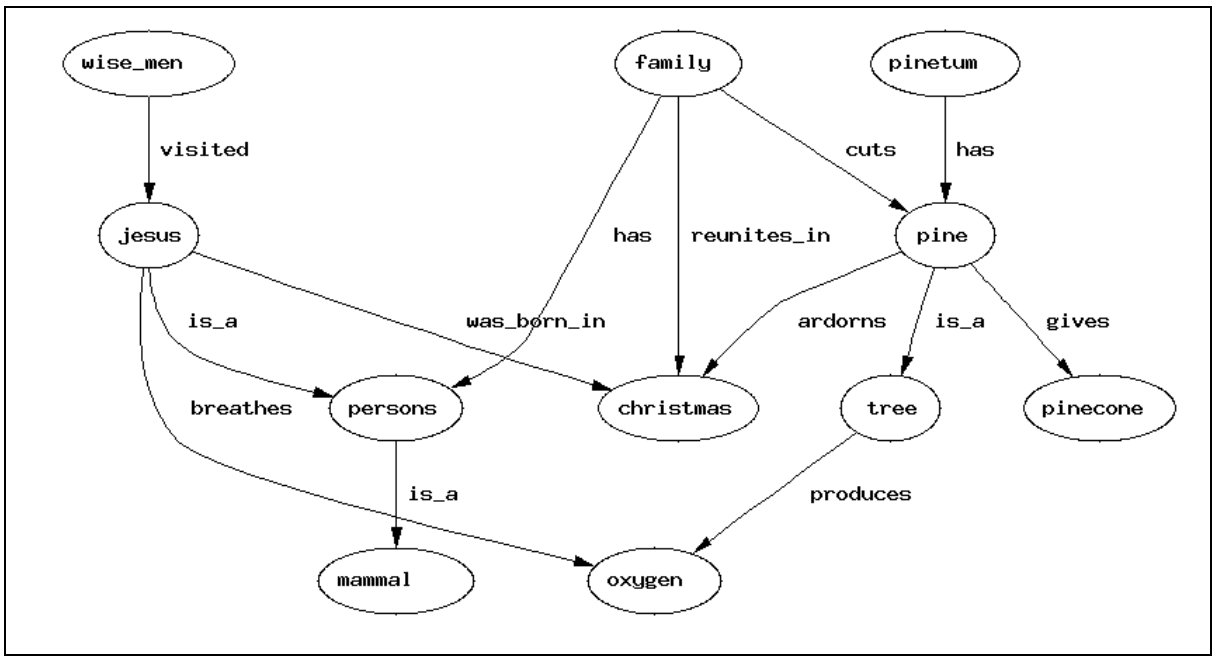
Gestão da Qualidade(1)



Gestão da Qualidade (2)

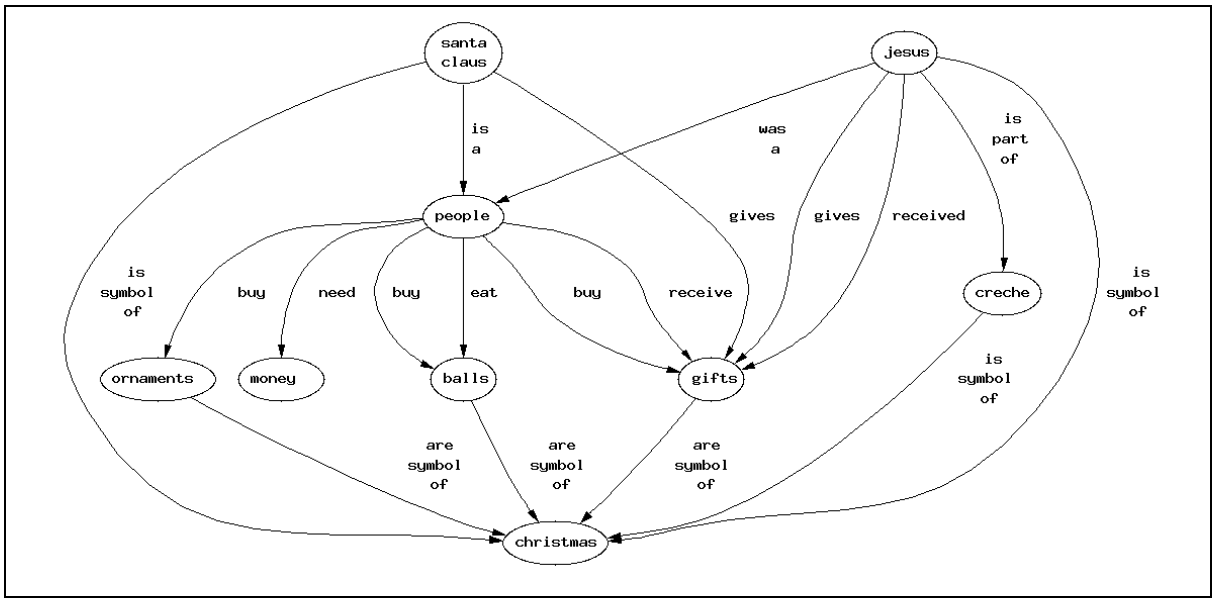


Natal(1)



Anexo 2. Mapas Conceptuais Utilizados

Natal(2)



A.3. Código Desenvolvido

Em anexo é fornecido um *cd-rom* com todo o código desenvolvido assim como os testes apresentados nesta Tese.