UNIVERSIDADE FEDERAL DO AMAZONAS

Instituto de Ciências Exatas

Programa de Pós-Graduação em Informática

L-*MATCH*: MAPEAMENTO SEMÂNTICO ENTRE ONTOLOGIAS UTILIZANDO CLASSIFICAÇÃO SUPERVISIONADA

FABRÍCIO D’MORISON DA SILVA MARINHO

Manaus

2008

UNIVERSIDADE FEDERAL DO AMAZONAS

Instituto de Ciências Exatas

Programa de Pós-Graduação em Informática

FABRÍCIO D’MORISON DA SILVA MARINHO

L-MATCH: MAPEAMENTO SEMÂNTICO ENTRE ONTOLOGIAS UTILIZANDO CLASSIFICAÇÃO SUPERVISIONADA

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Informática, área de concentração Inteligência Artificial e Recuperação de Informação.

Manaus

2008

Todos me criticam por eu ser diferente, mas eu rio deles por serem todos iguais (...), loucos como eu vivem pouco, mas vivem como querem, portanto não me importa se não houver o amanhã, pois me deram a vida e não a eternidade!

Charles Chaplin

Acontecem tantas coisas ruins no mundo as quais não podemos evitar, mas nem por isso devemos nos fechar pra ele! Somos capazes de muito durante a vida tão curta de um humano: rir, chorar, odiar e até amar. Nada é em vão! Por isso, por mais longe que eu vá, a verdade é que sempre estarei aqui!

Fabrício D’Morison

RESUMO

Descobrir automaticamente relações semânticas entre ontologias é uma tarefa de grande importância para Integração de Informação. Conhecido como **mapeamento** ou **alinhamento semântico, este** problema consiste em estabelecer **axiomas ponte** entre conceitos de ontologias distintas, porém não é abordado pela maioria dos sistemas de alinhamento devido à dificuldade de tratar heterogeneidade semântica, dada por diferenças de implementação que surgem quando várias ontologias são criadas para modelar domínios sobrepostos. Esta sobreposição pode ser total, parcial ou, dependendo da visão de mundo, desconhecida ou inexistente, sendo medida e qualificada, *a priori,* por cinco axiomas simples: **equivalência**, **mais geral**, **menos geral**, **sobreposição** e **diferença***.* Adotando o formalismo OWL, este trabalho objetiva abordar o mapeamento semântico entre ontologias diferenciando-se uma vez que aplica Classificação Supervisionada: conceitos são alinhados, enquanto instâncias e propriedades participam como evidências. A idéia é descobrir **axiomas de pertinência** entre instâncias de uma ontologia com conceitos de outra ontologia e vice-versa, independentemente se as instâncias foram declaradas em ambas as ontologias ou não. A classificação possibilita obter valores de sobreposição (interseção) e, paliativamente, valores de similaridade entre conceitos a mapear. Uma das grandes observações é que não se deve perder a noção de direção na qual estes valores foram computados, o que possibilita relaxar regras taxonômicas para restringir e estabelecer os mapeamentos semânticos de maneira bastante intuitiva. Após uma **estratégia** **indutiva bottom-up** para propagação dos valores na taxonomia, segue-se uma **estratégia dedutiva top-down** que estabelece os mapeamentos finais na forma axiomática. Diferentes classificadores foram experimentados, sendo notório o desempenho do classificador **Naive Bayes with Shrinkage**(NB-Shrinkage) que produziu valores médios de precisão e revocação acima de 80%. Por fim, uma nova abordagem para seleção e desambiguação de sinônimos em vocabulário restrito foi criada com base em Teoria de Grafos e Teoria da Informação, objetivando incrementar a precisão da classificação e conseqüentemente do mapeamento.

**Palavras-Chave:** mapeamento entre ontologias, mapeamento semântico, classificação supervisionada, Naive Bayes Shrinkage.

ABSTRACT

Automatically discovering semantic relationships between ontologies plays an important role for Information Integration. Known as **semantic mapping**or**alignment**, this problem consists of establishing **bridge axioms** between concepts declared in distinct ontologies, however isn't handled by most alignment systems because of the difficulty of dealing with semantic heterogeneity, raised by implementation differences which naturally appear when many ontologies are created to model overlapping domains. Such overlap might be total, partial or, depending on the world assumption, unknown or nonexistent, being measured and qualified, a priori, by five simple axioms: **equivalence***,* **more general***,* **less general***,* **overlapping**and**mismatching**. Adopting the OWL formalism, this work aims to approach the semantic mapping between ontologies differently since it applies Supervised Classification: concepts are aligned, while instances and properties are used as evidences. The idea is to discover **pertinence axioms** between instances of an ontology with concepts of another ontology and vice-versa, independently if the instances have been declared in both ontologies or not. The classification allows obtaining overlap values (intersection) and, palliatively, similarity values between concepts being mapped. One of the major observations is that the direction notion of such computed values must not be lost, allowing the relaxation of taxonomical rules to constrain and establish the semantic mappings in a very intuitive fashion. After a **bottom-up inductive strategy** to propagate the obtained values up through the taxonomy, it follows a **deductive top-down strategy** for establishing the final mappings in its axiomatic form. Different classifiers have been tested, however the classifier **Naive Bayes with Shrinkage** (NB-Shrinkage) notoriously outperformed the others, obtaining average precision and recall over 80%. Finally, a novel approach for synonym selection and disambiguation inside restricted vocabularies has been created backed by Graph Theory and Information Theory to increase both the classification and mapping accuracy.

**Palavras-Chave:** ontology mapping, semantic mapping, supervised classification, Naive Bayes Shrinkage.

ÍNDICE DE ILUSTRAÇÕES

[Figura 2‑1: Entrada e Saída de um Mapeador 20](#_Toc204281431)

[Figura 2‑2: Axiomas utilizados como ponte para mapear conceitos 22](#_Toc204281432)

[Figura 2‑3: Conceito de Vértice e Conceito de Rótulo. 24](#_Toc204281433)

[Figura 2‑4: Exemplos de cliques máximos 25](#_Toc204281434)

[Figura 2‑5: Redes Bayesianas para (a) Independência Causal e (b) Recuperação de Informação 27](#_Toc204281435)

[Figura 2‑6: Entrada e Saída de um Classificador. 31](#_Toc204281436)

[Figura 2‑7: Distribuição (a) sem *shrinkage* e (b) com *shrinkage* 34](#_Toc204281437)

[Figura 2‑8: Conjuntos Ideal, Computado e Acerto 35](#_Toc204281438)

[Figura 4‑1: Arquitetura do L-*Match* 50](#_Toc204281439)

[Figura 4‑2: Texto básico gerado para a instância *WaterTemperature* 51](#_Toc204281440)

[Figura 4‑3: Texto tratado para a instância *WaterTemperature* 52](#_Toc204281441)

[Figura 4‑4: Palavra repetindo-se em diferentes partes do texto prejudica a qualidade 53](#_Toc204281442)

[Figura 4‑5: Adição de prefixos de escopo ao texto 54](#_Toc204281443)

[Figura 4‑6: Unidades de Significado: os cliques serão os novos *synsets* 56](#_Toc204281444)

[Figura 4‑7: Apenas *synsets* mais informativos são mantidos, eliminando a polissemia. 57](#_Toc204281445)

[Figura 4‑8: Descobrindo relações de pertinência entre ontologias 60](#_Toc204281446)

[Figura 4‑9: Classificação automática alternada das instâncias de um par de ontologias 61](#_Toc204281447)

[Figura 4‑10: Ranqueamento de conceitos por similaridade com *BaseUnit* e suas instâncias 62](#_Toc204281448)

[Figura 4‑11: Rede Bayesiana modelada para propagação de similaridade 64](#_Toc204281449)

[Figura 4‑12: Má distribuição de instâncias numa ramificação da taxonomia 74](#_Toc204281450)

[Figura 5‑1: Distribuição de instâncias na Ecolingua.owl 78](#_Toc204281451)

[Figura 5‑2: Distribuição de instâncias na Apes.owl 79](#_Toc204281452)

[Figura 5‑3: Mapeamentos pretos são inferidos a partir dos mapeamentos vermelhos 82](#_Toc204281453)

[Figura 5‑4: Avaliação dos *ranks* de similaridade para ontologias de doenças 84](#_Toc204281454)

[Figura 5‑5: Avaliação dos *ranks* de sobreposição para ontologias de doenças 84](#_Toc204281455)

[Figura 5‑6: Gráfico da Avaliação por Classificador do Mapeamento Ecolingua×Apes 86](#_Toc204281456)

[Figura 5‑7: Gráfico da 1° Iteração de Mapeamento Ecolingua×Apes com NB-Shrinkage 90](#_Toc204281457)

[Figura 5‑8: Gráfico da 3° Iteração de Mapeamento Ecolingua×Apes com NB-Shrinkage 92](#_Toc204281458)

[Figura 5‑9: Gráfico da 2° Iteração de Mapeamento Disease1×Disease2 com NB-Shrinkage 95](#_Toc204281459)

[Figura 7‑1: Exemplo de Mapeamento Semântico **Erro! Indicador não definido.**](#_Toc204281460)

ÍNDICE DE TABELAS

[Tabela 3‑1: Sistemas de Integração 47](#_Toc204337057)

[Tabela 5‑1: Informações sobre classes, instâncias e propriedades das ontologias utilizadas 79](#_Toc204337058)

[Tabela 5‑2: Regras de expansão de mapeamentos 81](#_Toc204337059)

[Tabela 5‑3: Limiares de Relaxamento 85](#_Toc204337060)

[Tabela 5‑4: Avaliação Geral do Mapeamento Ecolingua×Apes 86](#_Toc204337061)

[Tabela 5‑5: Avaliação da Sobreposição Ecolingua×Apes por Classificador 87](#_Toc204337062)

[Tabela 5‑6: Avaliação da Similaridade Ecolingua×Apes por Classificador 87](#_Toc204337063)

[Tabela 5‑7: Modelo de Avaliação do Mapeamento 89](#_Toc204337064)

[Tabela 5‑8: Avaliação da 1° Iteração de Mapeamento Ecolingua×Apes com NB-*Shrinkage* 90](#_Toc204337065)

[Tabela 5‑9: Avaliação da 2° Iteração de Mapeamento Ecolingua×Apes com NB-*Shrinkage* 91](#_Toc204337066)

[Tabela 5‑10: Avaliação da 3° Iteração de Mapeamento Ecolingua×Apes com NB-*Shrinkage* 92](#_Toc204337067)

[Tabela 5‑11: Avaliação por Classificador do Mapeamento Iterativo Ecolingua×Apes 93](#_Toc204337068)

[Tabela 5‑12: Ganho por Classificador do Mapeamento Iterativo Ecolingua×Apes 93](#_Toc204337069)

[Tabela 5‑13: Avaliação da 1° Iteração de Mapeamento Disease 1 e 2 com NB-Shrinkage 95](#_Toc204337070)

[Tabela 5‑14: Avaliação da 2° Iteração de Mapeamento Disease 1 e 2 com NB-*Shrinkage* 95](#_Toc204337071)

[Tabela 5‑15: Avaliação da 1° Iteração de Mapeamento Cornell×Washington com NB-*Shrinkage* 96](#_Toc204337072)

[Tabela 5‑16: Avaliação da 1° Iteração de Mapeamento Rússia 1 e 2 com NB-*Shrinkage* 97](#_Toc204337073)

[Tabela 5‑17: Avaliação da 1° Iteração de Mapeamento Rússia 1 e 2 com NB-*Shrinkage*+*NaiveBayes* 98](#_Toc204337074)

ÍNDICE DE EQUAÇÕES

[Equação 1: 28](#_Toc204281477)

[Equação 4:  29](#_Toc204281478)

[Equação 5:  32](#_Toc204281479)

[Equação 6:  33](#_Toc204281480)

[Equação 7:  35](#_Toc204281481)

[Equação 8:  35](#_Toc204281482)

[Equação 9:  35](#_Toc204281483)

[Equação 10:  58](#_Toc204281484)

[Equação 11:  59](#_Toc204281485)

[Equação 12:  59](#_Toc204281486)

[Equação 13:  59](#_Toc204281487)

[Equação 14:  59](#_Toc204281488)

[Equação 15:  63](#_Toc204281489)

[Equação 16:  65](#_Toc204281490)

[Equação 17:  65](#_Toc204281491)

[Equação 18:  65](#_Toc204281492)

[Equação 19:  65](#_Toc204281493)

[Equação 20:  66](#_Toc204281494)

[Equação 21:  69](#_Toc204281495)

[Equação 22:  69](#_Toc204281496)

[Equação 23:  69](#_Toc204281497)

[Equação 24:  70](#_Toc204281498)

[Equação 25:  70](#_Toc204281499)

SUMÁRIO

[1 INTRODUÇÃO 1](#_Toc204282009)

[1.1 MOTIVAÇÃO E OBJETIVOS 1](#_Toc204282010)

[1.2 CONTRIBUIÇÃO 11](#_Toc204282011)

[1.3 ORGANIZAÇÃO DA DISSERTAÇÃO 12](#_Toc204282012)

[2 CONCEITOS E TERMINOLOGIA 13](#_Toc204282013)

[2.1 ONTOLOGIAS 13](#_Toc204282014)

[2.1.1 OWL – Ontology Web Language 16](#_Toc204282015)

[2.1.2 Mapeamento entre Ontologias 19](#_Toc204282016)

[2.1.2.1 Mapeamento Sintático e Semântico 21](#_Toc204282017)

[2.2 PROBLEMA DA ENUMERAÇÃO DE CLIQUES MÁXIMOS 25](#_Toc204282018)

[2.3 MODELO *NOISY-OR* DE REDES BAYESIANAS 26](#_Toc204282019)

[2.4 ENTROPIA 28](#_Toc204282020)

[2.5 CLASSIFICAÇÃO SUPERVISIONADA 30](#_Toc204282021)

[2.5.1 Conjuntos de Teste e de Treino 31](#_Toc204282022)

[2.5.2 O Classificador NB-Shrinkage 32](#_Toc204282023)

[2.6 PRECISÃO, REVOCAÇÃO E MEDIDA-F 34](#_Toc204282024)

[3 TRABALHOS RELACIONADOS 38](#_Toc204282025)

[4 MAPEAMENTO SEMÂNTICO ENTRE ONTOLOGIAS 48](#_Toc204282026)

[4.1 EXTRAÇÃO 50](#_Toc204282027)

[4.1.1 Geração de Texto para Classificar 51](#_Toc204282028)

[4.1.2 Qualidade do Texto 53](#_Toc204282029)

[4.2 COMPUTANDO SOBREPOSIÇÃO E SIMILARIDADE ENTRE CONCEITOS 60](#_Toc204282030)

[4.2.1 Propagação Bottom-Up na Taxonomia 63](#_Toc204282031)

[4.2.2 Combinação de Classificadores 66](#_Toc204282032)

[4.3 COMPUTANDO MAPEAMENTOS SEMÂNTICOS ENTRE CONCEITOS 67](#_Toc204282033)

[4.3.1 Regras de Compatibilidade entre Conceitos 67](#_Toc204282034)

[4.3.2 Comparação Top-Down 71](#_Toc204282035)

[5 EXPERIMENTOS 76](#_Toc204282036)

[5.1 ONTOLOGIAS UTILIZADAS 77](#_Toc204282037)

[5.2 AVALIAÇÃO HIERÁRQUICA DA QUALIDADE DOS MAPEAMENTOS COMPUTADOS 80](#_Toc204282038)

[5.3 RESULTADOS 85](#_Toc204282039)

[5.3.1 Avaliação Geral do Mapeamento Ecolíngua × Apes 85](#_Toc204282040)

[5.3.2 Avaliação do Módulo de Similaridade para Ecolíngua × Apes 87](#_Toc204282041)

[5.3.3 Avaliação do Módulo de Mapeamento para Ecolíngua × Apes 88](#_Toc204282042)

[5.3.4 Avaliação do Módulo de Mapeamento para Outras Ontologias 95](#_Toc204282043)

[6 CONCLUSÃO E TRABALHOS FUTUROS 99](#_Toc204282044)

[7 REFERÊNCIAS 102](#_Toc204282045)

Capítulo 1

# INTRODUÇÃO

## MOTIVAÇÃO E OBJETIVOS

Nos anos recentes, é considerável a quantidade de sistemas de informação que têm surgido para atender às necessidades de guardar, organizar e recuperar as informações de diversas atividades, como Biologia, Medicina, Ecologia, Agricultura, Turismo, Comércio Eletrônico e etc. Contudo, paralelo a este crescimento também surgem problemas técnicos de organização, de redundância e de interoperabilidade semântica, pois estes sistemas guardam modelagens locais e próprias de conhecimento que pode pertencer a domínios compartilhados. Conseqüentemente, o problema é que muitas vezes estes sistemas acabam precisando colaborar e trocar informação entre si, o que se torna muito difícil. Este fenômeno se dá, antes de tudo, devido às facilidades atuais tanto de comunicação quanto de representação de conhecimento providas pelos avanços da Computação, aqui abordados através do uso e reuso de ontologias e dos mapeamentos entre elas. Portanto, a pesquisa de técnicas para a integração de informação é crucial.

Como tendem a se tornar cada vez maiores e mais elaborados, estes sistemas demandam maior rigor e homogeneidade de informação em suas bases de dados. O maior e mais conhecido exemplo deles é a WEB: tendo-se popularizado há pouco tempo (menos de duas décadas), a Web não para mais de crescer. Isso leva a uma quantidade quase infindável de informações heterogêneas e redundantes para um ser humano processar. Daí a necessidade de procedimentos automáticos para lidar com a situação.

Associada à automação, decorre a necessidade de representações de conhecimento que privilegiem a semântica dos dados oriundos de tão diferentes domínios, viabilizando a organização clara, objetiva e homogênea da informação em detrimento de ambientes heterogêneos, o que por sua vez facilita a atuação de processos automáticos, como agentes inteligentes e máquinas de busca, as quais ainda hoje ignoram informação semântica. Neste contexto, a utilização de ontologias como estruturas de dados para representação formal de conhecimento tem se tornado cada vez mais freqüente, popular e atrativa, por permitir a modelagem dos dados em alto nível (estrutura e semântica), na forma de conceitos, propriedades, restrições e instâncias (Udrea, et al., 2007).

Ontologias têm grande valor para a internet. Não é surpresa que a grande incentivadora delas seja a Web, em especial a Web Semântica. Particularmente, as características ímpares da Web, como o crescimento exponencial do seu já imenso volume de dados, a diversidade de assuntos e a grande quantidade de usuários, levarão a um desenvolvimento distribuído de inúmeras ontologias para amenizar (ou mesmo eliminar) a heterogeneidade dos dados. Disto resultarão informações multiplamente representadas, caracterizando um novo cenário de problemas onde a heterogeneidade passa a ser semântica e as pesquisas se concentram em viabilizar soluções automáticas que permitam a **comunicação** entre sistemas ao estabelecer mapeamentos semânticos entre os conceitos de suas ontologias. Esta é a motivação que torna a integração automática de ontologias cada vez mais importante e desejada.

Porém, o mapeamento entre entidades de ontologias diversas ainda é uma questão em aberto para a integração da informação. O problema se torna mais complexo (e útil de fato) quando se diz que o **mapeamento é semântico** (veja a Seção **Erro! Fonte de referência não encontrada.** para maiores definições), para o qual se direcionam os trabalhos desta Dissertação, representando por isso um grande diferencial. Frequentemente, problemas de integração são intitulados como **casamento**, **alinhamento**, **mapeamento** e **fusão**. O escopo desta Dissertação cobre apenas os três primeiros casos, entendendo-os genericamente e utilizando-os sem distinção: a verdade é que na literatura não existe consenso exato sobre diferença entre eles, o que existe de fato são preferências que cada autor tem para nomear o mesmo problema, cujo objetivo é tão somente descobrir mapeamentos entre ontologias sem alterá-las. A única diferença real surge quando existe alteração nas ontologias caracterizando uma fusão, na qual o produto final ou é uma nova ontologia ou é uma nova versão de cada ontologia inicial.

Infelizmente, abordagens atuais de integração semântica costumam ser fortemente dependentes de fontes de conhecimento extra. É o caso da utilização de uma ontologia global unificadora que, funcionando como filtro, tenta criar mapeamentos que nunca sejam *ad hoc*, isto é, que nunca sejam criados diretamente entre ontologias locais. Há também abordagens híbridas que misturam ontologias locais e globais: os mapeamentos são *ad hoc*, porém facilitados por um vocabulário comum (uma ontologia, um dicionário, etc.) que ainda assim é imposto às ontologias locais. Estas abordagens não favorecem características de liberdade de descrição semântica específicas de autor e de domínio, nem de manutenção de ontologias e mapeamentos no decorrer do tempo. É inviável depender sempre de um vocabulário global para mapear, dada a diversidade de domínios específicos e a necessidade de alterações evolutivas e corretivas nas ontologias. Sabe-se que estas alterações criam inconsistências entre mapeamentos já estabelecidos, logo quanto menos ontologias envolvidas no processo melhor. Portanto, para o bem da portabilidade de um método de mapeamento, deve-se diminuir a dependência de conhecimento externo, mas nem por isso desprezá-lo, utilizando-o apenas como Informação Auxiliar (Rahm, et al., 2001). Por esta razão, é objetivo considerar exclusivamente mapeamentos *ad hoc*, que podem ser computados simplesmente através do uso de algoritmos que independem de conhecimento externo, mas que podem utilizá-lo indiretamente como item opcional se disponível, sem representar dependências e muito menos imposições. Nestes termos, foi desenvolvida uma abordagem auxiliar para seleção de sinônimos em vocabulários restritos, apoiada por Informação Auxiliar oriunda do *WordNet* (Miller, 1995) e baseada em Teoria da Informação e Teoria de Grafos, visando apenas incrementar a precisão do mapeamento a partir de conhecimento léxico.

Para um mapeamento ser considerado semântico, é suficiente que o seu produto final seja expresso na forma de axiomas. Segundo a Lógica, axiomas são verdades auto-evidentes e que por isso não requerem prova, podem ser proposições assumidas conforme a conveniência ou podem ser regras e princípios universalmente aceitos. Além disso, o mapeamento é considerado mais elegante quando as evidências utilizadas também são axiomas e, portanto, semânticas. Este é o caso de (Giunchiglia, et al., 2005), onde uma taxonomia é transformada num conjunto de proposições lógicas submetidas a um raciocinador específico. Porém, por mais elegante e sucesso que tenha, não existe abordagem de mapeamento capaz de clamar que processa única e exclusivamente evidências semânticas. Num primeiro momento de mapeamento é impossível ignorar o uso de heurísticas de similaridade, pois raciocinadores lógicos de nada ajudarão neste ponto, já que nenhuma verdade *a priori* é conhecida entre as ontologias a mapear. Felizmente, raciocinadores estatísticos são bastante adequados para lidar com a informalidade da heterogeneidade inicial e propor heurísticas que assumam verdades preliminares, sem dependência e imposições de conhecimento externo. É o caso das heurísticas de classificação supervisionada exploradas neste trabalho, que além de consolidadas como tecnologia, desempenham muito bem a tarefa de encontrar relações de pertinência as quais se encontram desconhecidas. A partir daí sim, em um segundo momento, quando verdades preliminares possam ser assumidas com confiança, justificar-se-á alocar esforços concentrados para aplicação de raciocínio lógico (Udrea, et al., 2007) (Giunchiglia, et al., 2005) (Haase, et al., 2005) na expectativa de ajustar o mapeamento inicial, reparando erros e conseqüentemente aumentando a corretude e a completude do processo como um todo.

Dadas tantas motivações, este trabalho começa a se definir ao tomar ontologias como objetos de estudo, mais precisamente seus conceitos unários, focando-se no mapeamento semântico entre eles. A principal estratégia utilizada para comparação de conceitos é extensional (veja a Seção **Erro! Fonte de referência não encontrada.** sobre estas estratégias): similar ao mapeador GLUE (AnHai, et al., 2004), as extensões investigadas são as instâncias dos conceitos as quais, caracterizadas por suas propriedades, permitem tratar o problema de mapeamento semântico como um problema de classificação supervisionada. Por fim, agora similar ao mapeador S-*Match* (Giunchiglia, et al., 2005), a saída do mapeador é sempre um axioma dentre **equivalência ()**, **mais geral (⊐)**, **menos geral (⊏)**, **sobreposto ()** e **diferença ()**.

Investigar instâncias é útil para mapeamento. Sem dúvida, o aprendizado de relações taxonômicas pode ser visto intuitivamente como uma tarefa de classificação (Cimiano, et al., 2004). Na maioria das vezes, a classificação em ontologias é feita por especialistas humanos e tende a capturar precisamente a relação entre sintaxe e semântica dos conceitos. Segundo (Korhonen, et al., 2004), esta classificação é atrativa por ter a propriedade de possibilitar, até certo ponto, a inferência da semântica de um conceito com base no seu comportamento sintático taxonômico e, mesmo não sendo suficiente para uma inferência semântica completa, a classificação é capaz de capturar generalizações de um grande número de propriedade, explícitas ou implícitas nas definições originais do autor da ontologia, sendo útil para resolver situações não cobertas por conhecimento léxico, como dicionários semânticos, os quais nunca são completos. Em (Resnik, 1990) é mostrado como taxonomias podem ser usadas para resolver ambigüidades sintáticas e semânticas e em (Sorokine, et al., 2005) é apresentado um estudo teórico e claro de como instâncias e propriedades são responsáveis por definir os limites de suas classes ou conceitos em uma hierarquia taxonômica.

Entender o significado que delimita os conceitos ontológicos é um problema chave do mapeamento. Não se deve contrariar a idéia e nem o entendimento que o autor da ontologia pretendia expressar no momento de criação de cada conceito. Por isso, mais uma vez, é importante evitar imposições e dependências globais. Um dos melhores recursos que o autor pode usar para evidenciar sua idéia é a utilização de exemplos dos conceitos por ele criados por meio de instanciação, um recurso didático que facilita aprender o objetivo ao qual é pressuposto cada conceito. Esta idéia é reforçada por ao afirmar, embasado em teoria, que as instâncias são boas evidências para relacionar conceitos que receberam interpretações dadas por comunidades diferentes. De fato, no que diz respeito a generalidade e especificidade, a abrangência de um conceito dada pela enumeração momentânea de instâncias pode apenas ser contrariada pela definição de uma restrição sobre o mesmo conceito. Porém, a utilização das restrições (ou axiomas de domínio) como evidências de mapeamento entre ontologias ainda é uma questão em aberto (Udrea, et al., 2007) na qual a literatura é pobre. Segundo (Fürst, et al., 2005), isso ocorre pelo seguinte:

“Certas evidências, como as propriedades algébricas, equivalências e disjunções, não são suficientemente usadas pela comunidade ao desenvolver ontologias para serem consideradas como material de similaridade. Já no caso dos axiomas, que incluem as regras e as restrições, não existe pesquisa nem suporte prático suficiente”.

A melhor estratégia para implementar estas idéias, de forma que instâncias sejam realmente úteis e tragam ganhos para o mapeamento, é aprender com elas. Para isso é necessário um sistema de classificação que seja capaz de classificar uma grande porção de instâncias a uma precisão aceitável. Isso é um desafio, pois é sabido que classificadores funcionam melhor conforme mais instâncias estão disponíveis para treiná-lo, sendo que a quantidade de instâncias nas ontologias não costuma ser grande. Contudo, esta situação é flexível: por outro lado existe vantagem, pois permite explorar técnicas de classificação mais eficazes que normalmente são ineficientes em bases de dados muito grandes, compensando a escassez de instâncias em muitas situações. O problema se agrava pois a literatura sobre mapeamento entre ontologias é muito pobre no que diz respeito à utilização de instâncias, classificação e demais técnicas de Aprendizado de Máquina. De maneira geral, a justificativa dada por muitos trabalhos para ignorar esta abordagem de pesquisa é que a maioria das ontologias hoje em dia não vem instanciada e por isso seria melhor investir em técnicas apenas no nível terminológico. Isso é muito contraditório, pois muitos dos mesmos trabalhos, cujo grande exemplo é , dizem que experimentam ontologias, mas exemplificam seus experimentos somente com taxonomias importadas de bancos de dados e de diretório da Web, ambientes onde as instâncias estão fartamente disponíveis. De qualquer forma, ontologias instanciadas realmente ainda são raras, mesmo porque ontologias como um todo ainda constituem uma tecnologia muito recente. Neste ponto, ainda é possível encontrar outra grande contradição: é um triste desperdício desenvolver uma ontologia e não utilizá-la para anotar dados, isto é, instanciá-la, pois isso sugere que a ontologia muito provavelmente será inútil por não atender à motivação de remover a heterogeneidade de dados e informações que precisam ser anotados. Isso é facilmente compreensível se citarmos o domínio biológico, que inclui subdomínios como agricultura, ecologia, medicina, piscicultura e genética, os quais constituem grandes exemplos de domínios onde as ontologias mais têm encontrado aceitação, investimento e pesquisa objetivando a anotação dos dados biológicos, antes em bases de dados comuns, heterogêneas e sem suporte à semântica. A verdade é que o mapeamento só é útil quando feito em ontologias que são usadas na prática e os dois grandes exemplos atuais são a *Gene Ontology* (GO) e a *Protein Ontology* (PO), vastamente utilizadas, anotadas e constituem as precursoras das ontologias futuras. Portanto, a atual escassez de ontologias instanciadas não constitui argumento para desconsiderar instâncias e classificação supervisionada durante mapeamento.

Para o nosso estudo, assumiremos que uma fração significativa das ontologias é instanciada. Sempre é interessante usar a maior quantidade de informação possível: uma vez que usamos apenas uma fonte de evidência ou apenas um algoritmo, ficamos muito dependentes deles, o que pode ser perigoso. A combinação das fontes de informação, objetivando que o mapeamento final seja semântico, de fato pode garantir que muito mais conceitos sejam corretamente mapeados. Por isso, apesar de nos concentrarmos numa abordagem específica, não defendemos utilizar exclusivamente a classificação das instâncias para mapear conceitos, mas sim investigar a melhor forma de explorá-las. De qualquer forma, a utilização de classificação tem a atrativa vantagem de ser flexível quanto à combinação de evidências e de algoritmos, por isso foram experimentados seis algoritmos de classificação, bem como combinações entre eles, incluindo algoritmos clássicos, no caso *k*NN (Shakhnarovich, et al., 2006), *Naive Bayes* (Rish, 2001) e SVM (Burges, 1998), e também algoritmos menos conhecidos, no caso *NB-Shrinkage* (Subseção 2.5.2), Máxima Entropia (Nigamy, et al., 1999) e TF-IDF (Joachims, 1997).

Um amplo corpo de ferramentas e trabalhos bem sucedidos, motivados por problemas da Web, tem trazido extensos benefícios a processos de classificação supervisionada, principalmente no que se refere a hipertexto e imagem, sugerindo fortemente a adaptação das técnicas existentes para problemas do domínio ontológico, a exemplo da anotação da *Gene Ontology* em e mapeamento em , o único trabalho realmente relevante que explorou classificação para fazer mapeamento entre conceitos. Essa realidade é preocupante, pois, no futuro, aplicar classificação automática para mapear ontologias da Web Semântica certamente será uma boa idéia. Por isso, o formalismo adotado para especificação das ontologias utilizadas no mapeamento é o *Ontology Web Language* ou simplesmente OWL (Seção 2.1.1), cuja popularidade é crescente. Em princípio, recursos específicos de OWL não foram explorados, como raciocínio envolvendo Lógica de Descrição em OWL-DL, contudo foi inevitável que características suas influenciassem o trabalho, especialmente porque este formalismo assume Visão de Mundo Aberto. Taxonomias em OWL permitem bastante sobreposição, dificultando a separação precisa das classes durante classificação automática e até mesmo manual em casos mais complexos, já que estas taxonomias suportam herança e instanciação múltiplas e são do tipo não-particionadas, por permitir sobreposição entre classes irmãs. Há também recursividade e as classes não podem ser consideradas disjuntas a não ser que a disjunção entre elas seja explicitamente declarada. Estas características criam um ambiente muito mais realista e flexível que trás várias facilidades, mas também dificulta a tomada de decisão, pois impede a utilização de certas restrições comuns em hierarquias taxonômicas mais tradicionais, constituindo um cenário onde modelar e processar classificação automática não é tão trivial. Felizmente, os resultados experimentais mostraram que os classificadores foram capazes de lidar com este ambiente.

Sendo assim, a falta de completa abordagem sobre mapeamento semântico e de satisfatória abordagem sobre classificação supervisionada aplicada para este fim, motivou os trabalhos desta Dissertação, cujo objetivo é investigar métricas, evidências e métodos de classificação capazes de estabelecer uma nova abordagem de boa qualidade para comparação e conseqüente alinhamento entre ontologias. Um protótipo do sistema proposto foi implementado, testado e chamado de L-*Match*, do inglês *Learning Match*. O L-*Match* é orientado à instanciação de conceitos e pode ser compreendido como um operador semântico *f* que toma duas ontologias A e B e sugere a **relação semântica mais forte** entre pares de conceitos como segue:



Os experimentos foram conduzidos principalmente sobre um par de ontologias de aplicação real, oriundas dos domínios Ecológico e Agricultura, os quais são sobrepostos. Estas ontologias são maiores, complexas e mais difíceis de mapear. Os resultados experimentais do nosso protótipo mostraram-se consideravelmente promissores principalmente no que diz respeito à precisão, porém satisfatoriamente equilibrados com revocação, o que é muito importante. Contudo, podemos considerá-los preliminares: primeiro porque otimizações de desempenho são necessárias, já que os experimentos foram orientados a eficácia (qualidade de resposta) e não totalmente a eficiência (velocidade de resposta); segundo porque os experimentos abriram uma ampla gama de possibilidades para novas experimentações que incluem combinações com outras técnicas de mapeamento; e, finalmente, porque é interessante dar suporte a outros formatos de ontologias além do OWL, visando portabilidade para experimentar novos casos de teste, facilitando inclusive comparação com outros sistemas.

## CONTRIBUIÇÃO

A principal contribuição deste trabalho é o desenvolvimento de um *framework* diferenciado por estudar fortemente técnicas de Classificação Supervisionada para mapear entre ontologias através de aprendizado, produzindo mapeamentos exclusivamente semânticos, o que ainda é novo e precisa de pesquisa, no que diz respeito a descobrir relações de **equivalência (), mais geral (⊐), menos geral (⊏), sobreposto () e diferença ()** entre conceitos de ontologias distintas.

Particularmente, uma grande diferença em relação a abordagens anteriores que utilizaram Aprendizado de Máquina é que não manipulamos apenas um algoritmo específico. Pelo contrário, investigamos e comparamos diferentes algoritmos de classificação, bem como combinações entre eles, em busca da mais eficaz forma de *classificar-para-mapear* semanticamente a partir de heterogêneas formas de evidências. O L-*Match* também não exige que conceitos mais genéricos sejam diretamente instanciados, identificando um método intuitivo e eficaz baseado em Redes Bayesianas para generalizar valores de similaridade e sobreposição através de uma propagação *bottom-up*, constituindo uma classificação hierárquica. Tal método pode ser reaproveitado e contribuir com qualquer outra abordagem que calcule similaridade entre conceitos sem utilizar classificação.

Além disso, vale ressaltar que a fase de pré-processamento das ontologias é de vital importância para as fases seguintes. Investimentos e atenção nesta fase são importantes. Inclusa nestes investimentos, desenvolveu-se uma abordagem mais prática para simplificar a identificação e desambiguação de sinônimos nos rótulos de conceitos, instâncias e propriedades. Considera-se que este é um vocabulário restrito (específico), onde sinônimos são cliques num grafo e a desambiguação é uma tarefa simples de selecionar sinônimos com mais informação e utilidade.

## ORGANIZAÇÃO DA DISSERTAÇÃO

Esta Dissertação está organizada em 7 capítulos a começar por esta introdução. No Capítulo 2 estão os principais termos e conceitos necessários para a boa compreensão deste trabalho. Seguindo ao Capítulo 3, encontra-se a revisão de literatura sobre os principais trabalhos correlatos. O Capítulo 4 é o mais importante, pois é nele que a abordagem desenvolvida ao longo deste trabalho é descrita, seguido pelo Capítulo 0, onde são apresentados os experimentos da abordagem e seus resultado comentados. No Capítulo 0, faz-se o fechamento do trabalho com a exposição de conclusões e sugestões para trabalhos futuros. Por fim, a enumeração de referências encontra-se no Capítulo 7.

Capítulo 2

# CONCEITOS E TERMINOLOGIA

Para este capítulo, foram selecionados, enumerados e revisados os principais conceitos necessários para a boa compreensão dos capítulos seguintes. Começando pela definição de ontologia, revisamos também aspectos principais sobre a linguagem OWL e sobre mapeamento entre ontologias com ênfase em mapeamento semântico. Não obstante, apresentamos brevemente o problema da enumeração de cliques máximos, prosseguindo diretamente para a discussão sobre combinação de evidências utilizando Redes Bayesianas e entropia. Na seqüência, apresentamos princípios básicos comuns a diferentes métodos de classificação supervisionada, finalizando o capítulo com conceitos sobre as métricas de avaliação de eficácia que aplicamos sobre os experimentos com o mapeador L-*Match*.

## ONTOLOGIAS

O termo **ontologia** vem do grego *ontos*+*logoi* que significa “conhecimento do ser”. Em Ciência da Computação e Ciência da Informação, diz respeito a uma tecnologia que visa estruturar dados e documentos em geral, atribuindo-lhes significado e contexto através de um modelo abstrato. Como pré-abordado no Capítulo 0, pode-se entender informalmente que ontologias são estruturas de dados especiais utilizadas para representação e anotação de conhecimento, funcionando como **bases de dados semânticas** ou **base de conhecimento**. Por esta razão, uma ontologia proporciona a modelagem formal do conhecimento acerca de um domínio (Genética, Turismo, Religião, etc.), relacionando entidades similares e agrupando-as em classes.

Popularizado pela comunidade de Inteligência Artificial, o termo ontologia é, contudo, originário da Filosofia e diz respeito ao estudo do que existe no mundo, ao estudo do “ser enquanto ser” e por isso compõe uma subárea da Metafísica, a parte da Filosofia preocupada com a existência. A seguir, mais algumas definições comumente dadas por outros autores:

“Uma especificação explícita e formal de uma conceituação compartilhada” (Gruber, 1995);

“Ontologia é a disciplina que lida com a natureza e a organização da realidade” (Guarino, et al., 1995);

“Uma ontologia identifica classes, cada qual caracterizada por propriedades que todos os elementos desta classe compartilham e as organiza hierarquicamente. Isto também inclui importantes relações entre classes e elementos, em um domínio de conhecimento específico” (Chandrasekaran, et al., 1999).

Representação formal de conhecimento tem suas motivações e aplicações. O grande objetivo é que a representação utilizada (uma ontologia, por exemplo) seja legível para computadores, viabilizando o processamento automático da semântica (informação contextual) que rege determinados objetos (como páginas da Web). As ontologias servem para compartilhar conhecimento entre sistemas, para classificar e aplicar inferências sobre os objetos de um domínio e para adicionar contexto nas buscas por informação, permitindo também que o sistema possa aprender conforme se agrega conhecimento na ontologia, manual ou automaticamente. Em relação a outras tecnologias, as ontologias estão em fase relativamente recente de pesquisa, seu uso é raro e quase que exclusivamente acadêmico, porém vem ganhando mais popularidade e aplicabilidade no decorrer do tempo.

Independentemente do formalismo utilizado para descrevê-las, ontologias são composta fundamentalmente pelas **primitivas conceituais** do domínio a ser modelado, que incluem os **conceitos** e as suas **propriedades**. Os conceitos são conjuntos abstratos de instâncias identificadas pelas mesmas propriedades. As propriedades, por sua vez, podem ser atributos elementares ou **relacionamentos** entre conceitos. Relacionamentos também são considerados conceitos, porém são conceitos compostos já que criam ligações específicas entre outros conceitos. Os tipos de relacionamento são muitos, sendo mais popularmente encontrados nas ontologias os relacionamentos taxonômicos (relação classe/subclasse) organizando os conceitos em hierarquias.

Ontologias podem também conter **regras lógicas** que criam restrições sobre conceitos e propriedades, aumentando o nível de formalidade da ontologia e permitindo a descrição mais fidedigna do domínio em questão. Além de conceitos, propriedades e regras, ontologias podem receber anotações de **instâncias**, também chamadasobjetos, exemplos ou indivíduos. As instâncias são os dados propriamente ditos, muitas vezes documentos. Elas são a manifestação, a personificação dos conceitos e propriedades concretizados no mundo real, constituindo por isso a base mais elementar da ontologia. Apesar disso, instâncias são consideradas **extensões**, pois não é obrigatório declarar explicitamente instâncias numa ontologia. Contudo, um dos grandes objetivos da Engenharia Ontológica é justamente a classificação e anotação de instâncias, mesmo que em um primeiro momento elas não estejam explicitamente declaradas nas ontologias.

Diferentes infra-estruturas e formalismos estão disponíveis hoje em dia voltados especificamente para a modelagem de ontologias. Dentre os formalismos, vale ressaltar *Lógica de Primeira Ordem* (Sowa, 2000), *Frames* (Minsky, 1975)e *Lógica de Descrição* (Baader, et al., 2002). Quanto à infra-estrutura, a linguagem que o W3C[[1]](#footnote-2) recomenda para especificação de ontologias é a *Ontology* *Web* *Language* (OWL), a qual vem se tornando um padrão cada vez mais forte, impulsionando especialmente pelo desenvolvimento da ferramenta *Protégé[[2]](#footnote-3)*.

### OWL – Ontology Web Language

Ontologias consideradas formais são especificadas por linguagens que envolvem algum tipo de lógica. Dos diferentes padrões disponíveis atualmente, a linguagem adotada no presente trabalho é a OWL, *Ontology Web Language*, pois é um padrão aberto. Isso é muito importante, pois seu uso é mais difundido e bem suportado por ferramenta gratuita e com código aberto, facilitando os trabalhos tanto dos desenvolvedores de ontologias quanto dos desenvolvedores de aplicações. Por isso, a OWL tem sido bastante utilizada para a construção de inúmeras ontologias espalhadas por repositórios da Web, contrastando com outras linguagens que são utilizadas apenas em projetos específicos, como ocorre com projeto conhecidos, a exemplo do Cyc[[3]](#footnote-4), definido como uma grande base de conhecimento provida por uma máquina de inferência própria, e o projeto OBO, cuja linguagem é utilizada basicamente apenas nas ontologias médicas desenvolvidas pelo próprio projeto.

OWL em inglês significa “coruja”, animal que ficou conhecido como símbolo grego da sabedoria, já que sempre acompanhava *Athena*, a deusa da sabedoria e da guerra justa. Uma das características marcantes de OWL é o fato de usar Visão de Mundo Aberto. Esta visão de mundo assume que não se sabe tudo que é possível saber sobre o mundo, por mais que se tenha uma grande base de conhecimento. Tecnicamente, se uma suposta verdade não pode ser provada apenas com o conhecimento disponível sobre o mundo, é impossível concluir que ela seja falsa e, portanto, permanece desconhecida. Em outras palavras, nenhuma verdade pode ser falsa a não ser que seja explicitamente declarada como falsa. É o contrário do que ocorre com outras linguagens, a exemplo de Prolog, que utilizam Visão de Mundo Fechado a qual prega que o conhecimento disponível sobre o mundo é suficiente para provar qualquer coisa como falso ou como verdadeiro.

A linguagem OWL visa aproximar ontologias e Web, contribuindo significativamente para o desenvolvimento da Web Semântica. A idéia é atribuir significado explícito aos dados publicados na rede mundial, que então se tornaria um ambiente facilitador da ação e da interoperabilidade entre aplicações inteligentes. Considerado um dos trabalhos pioneiros e ainda recentes acerca de recomendações do W3C relativas a padrões da Web Semântica, a linguagem OWL é o resultado mais atual dos trabalhos, estudos e lições aprendidas durante o desenvolvimento de tecnologias anteriores sobre modelagem, pois sua estrutura é implementada com bases em XML, RDF e RDF-*Schema*, bem como, na qualidade de linguagem de ontologia, é considerada uma revisão e evolução dos conceitos das linguagens OIL, DAML e DAML+OIL.

Facilitando-se destes modelos, OWL herda a capacidade genérica para rotular conteúdo através da sintaxe XML e herda a declaração de objetos/recursos e suas relações a partir de RDF, utilizando RDF-*Schema* para agrupar os objetos em classes e propriedades organizadas em hierarquias taxonômicas. Nesta pilha de modelos, OWL é a última camada que chega para adicionar mais detalhamento e riqueza semântica, através de um vocabulário que finalmente confere o nível de ontologia para o modelo como um todo. Resumindo, uma ontologia OWL é um conjunto de triplas RDF imbuídas de significado.

O *trade-off* entre **expressividade** e **decidibilidade** é abordado por OWL. Quanto mais expressiva for uma linguagem, maior a liberdade e facilidade para declarar conhecimento, em detrimento da completude e da decidibilidade, que é a complexidade dos algoritmos de racicinio. Para isso, OWL divide-se em três sub-linguagens que variam de formalismos mais fracos até mais fortes. A primeira é **OWL-*Lite***, cujo objetivo é ser mais simples e menos expressiva, sendo porém mais fácil de tratar computacionalmente, pois oferece suporte apenas a recursos básicos como taxonomias e restrições simples. A segunda linguagem é **OWL-DL** e o seu objetivo é o melhor equilíbrio. Com propriedades computacionais mais atrativas, adiciona maior formalismo às ontologias utilizando Lógica de Descrição, permitindo maior expressividade e garantindo a completude e decidibilidade do raciocínio. Finalmente, o máximo de expressividade e liberdade pode ser obtido apenas por **OWL-*Full***, ao custo da eliminação das garantias de completude e de decidibilidade.

Poucos sistemas consideram o uso de OWL-*Full* devido ao risco de intratabilidade e pela dificuldade de explorar restrições mais fortes. No caso dos sistemas de mapeamento entre ontologias, a maioria utiliza OWL-*Lite* e uma minoria tenta explorar OWL-DL. Apesar desta Dissertação atualmente não explorar recursos de Lógica de Descrição, o L-*Match* aceita ontologias especificadas tanto em OWL-*Lite* como em OWL-DL.

Por fim, maiores informação sobre OWL podem ser encontradas no site oficial[[4]](#footnote-5) mantido pela W3C.

### Mapeamento entre Ontologias

Toda atividade de integração de informação (relacionar, comunicar, transferir, interoperar, fundir, etc.) depende de alguma rotina inicial de mapeamento. Esta habilidade de combinar modelos automaticamente, relacionando e integrando suas estruturas, é uma área de pesquisa cada vez mais sólida sobre um problema comum a diferentes comunidades, como casamento de grafos (Klinger, et al., 2005), Visão Computacional (Pelillo, et al., 1998) e aplicações de bancos de dados (Rahm, et al., 2001), como gerenciamento de conhecimento científico, comércio eletrônico e *data* *warehousing*. No mais, a integração de ontologias tem uma tênue correlação a problemas de integração de esquemas de bancos relacionais (Udrea, et al., 2007), a diferença é que o compromisso com semântica é rigorosamente maior nas ontologias.

Como (Euzenat, et al., 2003), classificamos as abordagens de mapeamento em função das fontes de evidência e das técnicas de comparação:

* **Comparação Terminológica –** compara rótulos de entidades. Inclui técnicas de processamento de string (distância de edição, prefixo/sufixo e N-Gram), técnicas baseadas em linguagem (*tokenization*, *stemming* e *stopwords*) ou em recursos lingüísticos, como *thesaurus* e dicionários taxonômicos.
* **Comparação da Estrutura Interna –** compara evidências da estrutura interna de entidades (domínio, *range*, cardinalidade, etc.).
* **Comparação da Estrutura Externa** – compara as estruturas de relacionamento entre entidades, como a árvore taxonômica e o grafo cíclico formado por outras relações. Neste item, algoritmos para grafos são comuns.
* **Comparação Extensional –** compara extensões das entidades. Extensões são entidades que, em princípio, não precisam vir nas ontologias. Exemplos são comentários e instâncias. Inclui os algoritmos de classificação automática.
* **Comparação Semântica –** compara a interpretação (modelo lógico) das entidades. Inclui técnicas que processam axiomas, envolvendo inferência lógica.

**O L-Match utiliza comparação terminológica, da estrutura externa e extensional. Observe que mapeamento semântico não é o mesmo que comparação semântica (vide subseção 2.1.2.1), podendo andar juntos ou não.**

Definimos genericamente um **mapeador de ontologias** como um sistema em função de sua **entrada** e **saída**. Sejam então as ontologias *A*,com *m* conceitos, e *B*,com *n* conceitos:



Figura ‑: Entrada e Saída de um Mapeador

* O mapeador é uma função ;
* A entrada são as duas ontologias *A* e *B*;
* A saída é o mapeamento potencial entre cada par de conceitos na forma de triplas .

Onde são conceitos oriundos das ontologias *A* e *B*, respectivamente, e é o **qualificador** do mapeamento, podendo aparecer sob a forma de uma relação axiomática, um valor de similaridade na escala [0,1], ou ambas as coisas.

#### Mapeamento Sintático e Semântico

Diz-se que o **mapeamento é sintático** quando o seu qualificador *q* é composto apenas por um valor numérico de similaridade. A tarefa consiste em encontrar os pares de conceitos mais similares, isto é, mapeamentos 1-1 (um para um). Contudo, este tipo de mapeamento é bastante limitado, pois se restringe a estabelecer apenas relações de equivalência ou de diferença. O qualificador *q* ideal para mapeamentos entre ontologias recai na noção básica de **axioma ponte**. Axiomas qualificam mapeamentos mais expressivos, imbuídos de maior riqueza semântica e permitem discriminar melhor quais as propriedades que se aplicam simultaneamente a elementos de ontologias distintas, permitindo mapeamentos do tipo *N-N* (muitos para muitos).

De maneira genérica, entende-se por **mapeamento semântico** a tarefa de estabelecer axiomas ponte entre conceitos de ontologias distintas, de forma que os mapeamentos sejam discretos e permitam inferência lógica. Cinco axiomas serão utilizados como pontes de mapeamento neste trabalho, ordenados a seguir do mais forte até o mais fraco: **equivalência ()**, **mais geral (⊐)**, **menos geral (⊏)**, **sobreposto ()** e **diferença ()**. Estes axiomas são operadores de Lógica de Descrição, representam as relações mais básicas e possíveis de estabelecer entre conjuntos e oferecem um ponto conveniente entre expressividade e tratabilidade para a atual maturidade deste trabalho. As relações expressas por estes axiomas são melhor definidas como interseções entre dois conceitos fictícios A e B em diagramas de Vem, como na figura seguinte:

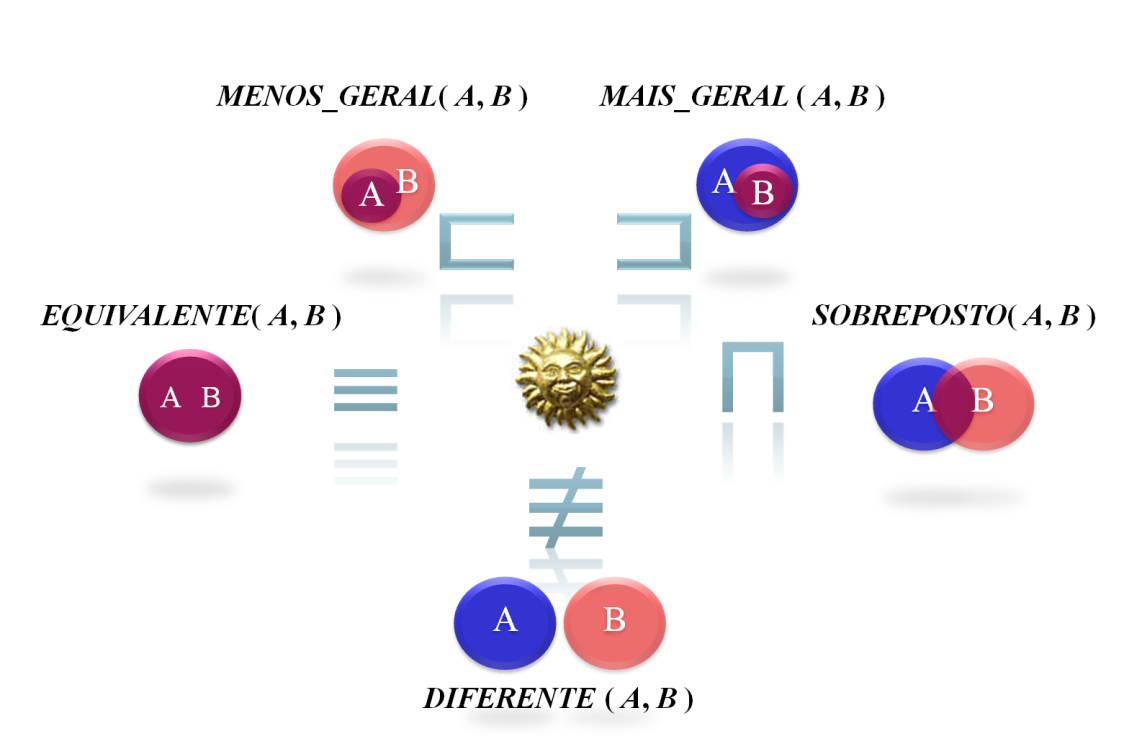


Figura ‑: Axiomas utilizados como ponte para mapear conceitos

Quanto maior a interseção, mais forte será a relação. A *equivalência* é a relação mais forte por representar uma interseção total entre A e B, o que significa que suas propriedades e instâncias são as mesmas. A equivalência também é a relação mais importante, pois é utilizada quando a mesma entidade é encontrada em duas ontologias. As relações de *mais geral* e *menos geral* são relações inversas e não são tão fortes quanto a equivalência, pois representam a interseção total de apenas um dos conceitos: A é menos geral que B se todas as instâncias de A também pertencem a B, porém a recíproca não é verdadeira; logo, B é mais geral que A. Estas relações de generalidade fornecem informação de herança (superclasse/subclasse), permitindo identificar quando um conceito contém ou é contido por outro. Contudo, diferente de (Giunchiglia, et al., 2005), assumimos que a relação de mais geral é mais forte que a de menos geral. Na seqüência, temos a relação de *sobreposição* que é a mais fraca dentre as relações que envolvem alguma interseção: representa a interseção parcial de ambos os conjuntos, nem total e nem nula. A sobreposição é a relação menos importante, pois, a princípio, nenhuma informação relevante pode ser concluída a partir dela. Por fim, resta a relação *diferença*, aplicada quando os conceitos são disjuntos, isto é, não existe qualquer interseção entre eles, indicando que o complemento de A contém B, e vice versa. Por ser o oposto da equivalência, a diferença também é bastante importante, pois assumir a mínima interseção entre conjuntos disjuntos configura erro grave.

Excepcionalmente, pode-se ainda pensar na relação **desconhecido (?)** a ser aplicada como uma relação adicional ou substituta da diferença. Por exemplo, em uma linguagem que utiliza Visão de Mundo Aberto como o OWL, pode não ser muito adequado estabelecer automaticamente relações de diferença sem uma boa justificativa. Por outro lado, no S-*Match* a relação desconhecido não substitui a diferença e é utilizada adicionalmente.

Os mapeamentos semânticos podem também ser qualificados por **axiomas complexos**, formado a partir da composição de mais a **negação(¬)** e eventualmente a **união()**, como observado e exemplificado por (Bouquet, et al., 2003). Este seria o caso de utilizar todos os recursos de inferência para Lógica de Descrição. Porém, a composição de axiomas complexos está além dos trabalhos desta Dissertação.

Ao apresentar o S-*Match*, (Giunchiglia, et al., 2005) também não usa axiomas complexos e faz considerações interessantes sobre mapeamento semântico, deixando clara a diferença entre **conceito de vértice** e **conceito de rótulo**, o que normalmente causa bastante confusão e discussões desnecessárias entre especialistas: não se deve esquecer que o mapeamento é entre os conceitos atribuídos aos vértices, e não aos rótulos. Veja:

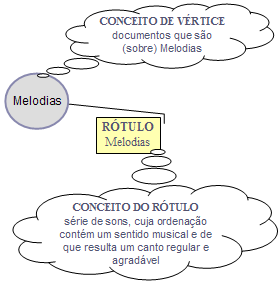


Figura ‑: Conceito de Vértice e Conceito de Rótulo.

Neste caso, a idéia de conceito dentro da ontologia vai mais longe, pois o conceito passa a ser divisível e composto por duas entidades: um vértice e um rótulo. No exemplo, temos um vértice rotulado como Melodia,maso conceito do rótulo não necessariamente é igual ao conceito do vértice, apesar de que estão relacionados. A observação chave é que rótulos em hierarquias de classificação são usados para definir o conjunto de documentos a serem classificados sob a classe que é identificada pelo rótulo. Por isso, quando falamos de *Melodias*, não necessariamente queremos definir ou referir-se exatamente a “*uma série de sons, cuja ordenação contém um sentido musical e de que resulta um canto regular e agradável*”, mas sim a “*documentos que são (sobre) Melodias*”.

O rótulo é apenas um identificador humano, uma facilidade, uma palavra conveniente que deve ser simples e pequena o suficiente para desempenhar sua função que é tão somente terminológica. Ao ser lido por uma pessoa, o rótulo remete à semântica mais comum que ele receberia no mundo real. Porém, o conceito de vértice não é determinado pela terminologia, ele é determinado e regido pela lógica, através das propriedades e regras declaradas numa ontologia.

## PROBLEMA DA ENUMERAÇÃO DE CLIQUES MÁXIMOS

Problemas de cliquesão clássicos na área de Otimização Combinatória, sendo muito importantes para diversas áreas, como Biologia (especialmente Genética) e Recuperação de Informação, como é o caso desta Dissertação e de trabalhos com *links* das páginas Web. Por definição, um clique é um subgrafo completo (existe aresta de todo vértice para todo vértice), veja o exemplo:

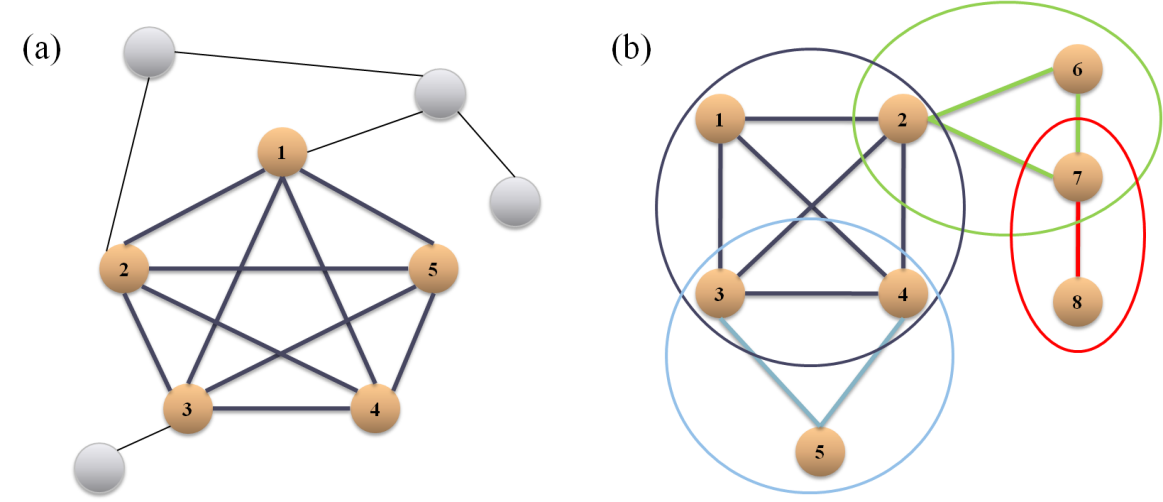


Figura ‑: Exemplos de cliques máximos

No grafo (a) existem vários cliques menores em destaque e todos estão contidos num clique maior de tamanho cinco. Este é um **clique máximo** (Du, et al., 2007), definido como o maior clique que nãoestá contido em outro clique. Contudo, neste trabalho estamos interessados num problema mais abrangente que consiste na **enumeração de todos os cliques máximos** (Bron, et al., 1973) como no grafo (b) que contém quatro cliques máximos: observe que nenhum clique máximo em está contido em outro.

Por fim, o problema de encontrar um clique máximo é *NP*-completo, mas o problema de enumerá-los é *NP*-difícil. Para tal, pode-se usar estratégias heurísticas e/ou de força bruta, dependendo do tamanho do espaço de busca do problema específico.

## MODELO *NOISY-OR* DE REDES BAYESIANAS

Em Inteligência Artificial, abordagens que lidam com raciocínio podem ser divididas em duas grandes áreas que compreendem **Raciocínio Lógico** e **Raciocínio Probabilístico** (ou Estatístico). No contexto probabilístico, o emprego de Redes Bayesianas (Jensen, 2001) constitui umas das principais técnicas de raciocínio (como lógica *fuzzy*), especialmente útil quando se quer diminuir/sintetizar o total de probabilidades a cerca de um sistema. Exemplos de aplicação incluem classificação automática, reconhecimento de voz, análise de texto em geral, processamento de imagem, suporte a decisão, seqüências de proteínas, etc.

Também conhecidas como redes de crença, redes de opinião, redes causais e etc., as Redes Bayesianas associam Teoria das Probabilidades com Teoria dos Grafos para modelar a idéia de relacionamento probabilístico entre entidades. Tecnicamente, uma Rede Bayesiana é um modelo de árvore, isto é, um grafo acíclico direcionado (DAG) no qual os vértices representam variáveis ou entidades de qualquer natureza e os arcos representam relacionamentos entre as entidades. Os relacionamentos são condicionalmente independentes, isto é, o estado de uma entidade (causa) influencia o estado de outra entidade (conseqüência), contudo as causas são independentes umas das outras, garantindo que o grafo seja acíclico.

Apesar do poder de Raciocínio Lógico ser inegável, as Redes Bayesianas são consideradas melhores para tratar **raciocínio com** **incerteza**, situação na qual não se sabe tudo a respeito do domínio, ou seja, falta informação (Ignorância Teórica):

“A principal vantagem de raciocínio probabilístico sobre raciocínio lógico é fato de que agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará” (Charniak, 1991).

A Figura 2‑5 exemplifica Redes Bayesianas combinando evidências: se uma proposição representada por um nó *x* implica em uma proposição *y*, então uma aresta é colocada de *x* para *y* significando . O modelo em (a) representa a estrutura geral de **Independência Causal** em Redes Bayesianas (Lamma, et al., 1999), uma árvore que expressa a idéia de que um conjunto de causas independentes entre si influenciam um efeito comum através de causas intermediárias utilizando uma função determinística *Operador*. Em (b) a rede foi modelada especificamente para recuperar documentos em uma máquina de busca: *dj* são os documentos, *q* é a consulta e *ki* são os termos do vocabulário da coleção. Neste exemplo, os termos são evidências (causas) que, quando combinadas, representam documentos e consulta (conseqüências ou efeitos).

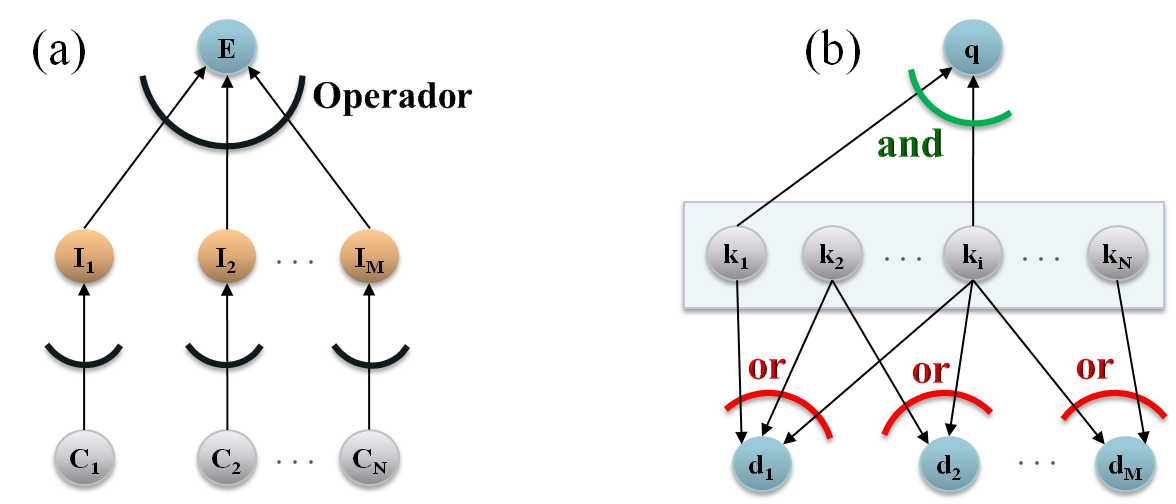


Figura ‑: Redes Bayesianas para (a) Independência Causal e (b) Recuperação de Informação

Como em (Vale, et al., 2001), o exemplo acima demonstra a utilização de dois operadores diferentes: um conjuntivo (AND) combinando evidências para a consulta, e um disjuntivo (OR) combinando evidências para os documentos. Estes operadores representam dois típicos modelos para construir Redes Bayesianas: o modelo *Noisy-Or* e o modelo *Noisy-And*. Excepcionalmente, existem ainda os modelos *Noisy-Max* e *Noisy-Min*.

Consideramos apenas o modelo *Noisy-Or* para propagação *bottom-up* em árvores de generalização ou taxonomias (Subseção 4.2.1). Esteé um dos modelos mais populares para construir Redes Bayesianas em diversas situações práticas e tem a mesma estrutura matemática de funções de combinação (Lamma, et al., 1999). O *Noisy-Or* é uma adaptação probabilística equivalente ao OU da Lógica, e por isso representa tão somente uma disjunção de preposições. Seja um efeito que depende das causas , o *Noisy-Or* é definido genericamente pelo produtório do complemento das causas, veja:

Equação :

Por exemplo, se o efeito *E* dependesse das probabilidades causais *a*, *b* e *c*, a fórmula acima seria assim instanciada: .

## ENTROPIA

Entropia é uma métrica popular e eficaz para medir uniformidade de distribuições de probabilidade. Sua teoria é bastante intuitiva, podendo ser entendida informalmente como o ato de *medir a informação* *de um* *conjunto de resultados*, independente de qual seja a semântica específica destes resultados.

É utilizada em várias partes deste trabalho, sendo associada como peso durante a combinação de evidências probabilísticas feita com *noisy-or* ou sendo aplicada para medir a quantidade de informação em *synsets* com palavras polissêmicas, visando a desambiguação destas palavras. O princípio chave da utilização de entropia durante a combinação de evidências, o qual muitas vezes é esquecido, é que, tendo-se à disposição várias distribuições de probabilidade, deve-se certamente preferir a distribuição que seja mais uniforme: este é o conceito da Máxima Entropia. Um dos mais interessantes exemplos de aplicação foi feito no trabalho de (Nigamy, et al., 1999) que, seguindo este princípio, desenvolveu o precursor dos classificadores de texto baseados na Máxima Entropia.

O temo entropia vem do grego “**εντροπία”,** mais precisamente *em + trope* que significa “*em transformação*”: quanto maior a entropia maior a probabilidade de que um dado evento ocorra. Originalmente empregado na termodinâmica (Física), foi agregado à moderna *Teoria da Informação* como *Entropia Informacional* desde a publicação de (Shannon, 1948) em 1948, na tentativa de modelar matematicamente a comunicação como um problema estatístico para benefício do trabalho de engenheiros elétricos.

A Entropia Informacional é usada para medir a quantidade de informação contida em um conjunto de símbolos que formam uma mensagem. No entanto, a teoria de entropia não considera a semântica dos dados, mas sim o nível de caos ou desequilíbrio entre eles: quanto maior o caos, menor a quantidade de informação. A entropia informacional *H* de um conjunto de símbolos de tamanho *n* é definida como a seguir:

Equação : 

*  é a probabilidade de ocorrência de cada i-ésimo símbolo dentro do conjunto de *n* símbolos, logo temos que ;
*  é o grau de caoticidade;
* , logo podemos normalizar *H*: 

Como se vê, o valor máximo de entropia ou valor máximo de informação ocorre quando a distribuição das probabilidades  é totalmente uniforme. Por exemplo, a probabilidade  poderia representar a probabilidade de um termo ocorrer dentro de um documento de texto e seria dada pela razão entre freqüência do termo no documento e a soma das freqüências de todos os termos que ocorrem no mesmo documento.

## CLASSIFICAÇÃO SUPERVISIONADA

Classificação Supervisionada trata de um subconjunto de problemas de Aprendizado de Máquina vastamente utilizados para Recuperação de Informação (R.I.). Encontra grande aplicação na Web, como em filtros de spam e em diretórios de máquinas de busca, onde a divisão de documentos é feita por categoria, classificados taxonomicamente por assunto. Bastante pesquisa sobre classificação supervisionada tem sido feito no decorrer das últimas décadas, sugerindo fortemente a adaptação das técnicas existentes para o domínio ontológico. Apesar de pouco explorado neste domínio, o uso de Aprendizado de Máquina em ontologias felizmente não é novidade, como em para fazer mapeamento entre ontologias e em onde classificadores e algoritmos genéticos são combinados para instanciar a *Gene Ontology* (GO). Este exemplo mostra a importância de combinar evidências e algoritmos, o que é fácil utilizando classificação supervisionada.

Denomina-se **classificador automático** todo algoritmo capaz de atribuir categorias a instâncias sem categoria conhecida (objetos de teste), a partir de um conjunto de categorias conhecidas e finitas (objetos de treino). A classificação é **supervisionada** quando utiliza uma base de conhecimento com exemplos previamente classificados (amostras de treinamento) nas categorias conhecidas. Quando isso não ocorre, a classificação é **não-supervisionada**. Em estatística, os classificadores são conhecidos como **funções discriminantes**.

Uteis por substituir o trabalho manual, as máquinas de classificação tem a difícil tarefa de identificar, grosso modo, o “tipo” ou “qualidade” de cada objeto submetido a elas, como imagens, sons, páginas html, e-*mails* e outros tipos de documentos, mídia e dados, utilizando-se de passos matemáticos, sintáticos e semânticos. Quando feita por especialistas humanos, a classificação é considerada mais segura. No entanto, mesmo os especialistas podem ter dúvidas e errar ao classificar um objeto, por isso não se deve esperar que mesmo a melhor máquina de classificação acerte sempre 100% a classificação dos objetos que recebe.

### Conjuntos de Teste e de Treino

Independente de como um classificador supervisionado for implementado, ele sempre trabalha com dois conjuntos: o *conjunto de teste* e o *conjunto de treino*. Veja:



Figura ‑: Entrada e Saída de um Classificador.

O *conjunto de teste* é um grupo de instâncias cuja classificação ainda é desconhecida, ou pelo menos, assume-se que é desconhecida no momento em que eles são submetidos ao classificador. A expressão “*testar a instância*” é o mesmo que afirmar que tal instância vai ser submetida à classificação.

*O conjunto de treino* é um grupo de instâncias cuja classificação é conhecida, isto é, instâncias pré-classificadas que servirão de exemplo para o classificador, formando sua base de conhecimento. Assume-se que a pré-classificação do conjunto de treino é confiável e indiscutível. Normalmente, para assegurar a confiabilidade do conjunto de treino, a pré-classificação é feita por especialistas humanos. Porém, em casos mais específicos, é aceito que pré-classificação seja feito por procedimento automático considerado confiável e seguro.

### O Classificador NB-*Shrinkage*

Dos diferentes algoritmos de classificação utilizados para mapeamento no L-*Match*, o *Naive Bayes with Shrinkage*, ou simplesmente NB-*Shrinkage*, foi evidentemente melhor. Por isso, esta Seção foi criada para destacar os conceitos gerais associados a este classificador e abordar especificamente a técnica de *shrinkage*.

A denominação *Naive Bayes* remete a uma família de classificadores estatísticos bem fundamentados em Probabilidade Bayesiana. Neste caso, “*naive*” significa que estes classificadores utilizam o Teorema de Bayes simplificado, assim definido em Estatística:

Equação : 

A notação equivale ao *Modus Pones* de lógicas discretas, como Lógica Proposicional, e significa “*o que se pode inferir sobre A, dado que tudo que se sabe é B?*”. Cada parte tem nome próprio: é a probabilidade posterior, é a probabilidade anterior de A, é a probabilidade condicional de B dado A, e é utilizada como constante de normalização. No caso de classificação supervisionada, seja uma classe C descrita por um conjunto de características , o Teorema de Bayes é reescrito:

Equação : 

Por sua simplicidade e eficiência, estes classificadores são bastante utilizados em diferentes problemas (notoriamente em categorização de texto), também porque costumam ser bastante eficientes, pois são treinados em tempo linear. Por esta razão, são considerados classificadores de bom custo benefício entre eficácia e eficiência.

A forma mais simples do *Naive Bayes* não é tão robusta e sofre com alguns problemas. Contudo, muitas melhorias foram propostas, culminando no surgimento de algoritmos que associam outros princípios ao *Naive Bayes* original, como *smoothing* (Juan, et al., 2002), maximização de entropia (Nigamy, et al., 1999) e *shrinkage* (McCallum, et al., 1998).

Na Estatística, s*hrinkage* é uma técnica conhecida e eficaz usada para melhorar estimativas de probabilidades e pode ser associada paralelamente com outras técnicas de *smoothing*. Intimamente relacionada à inferência Baiyesiana, a técnica de *shrinkage* foi adaptada para classificação, dando origem ao classificador NB-*Shrinkage* que incrementa significativamente a eficácia do *Naive Bayes*, como mostram experimentos de classificação de texto em (McCallum, et al., 1998) nos quais o *accuracy* do Naive Bayes foi aumentado em até 29% ao utilizar *shrinkage*. Por exemplo, considerando a média de uma distribuição Gaussiana e a variância em torno da média, basicamente o que ocorre é o seguinte:

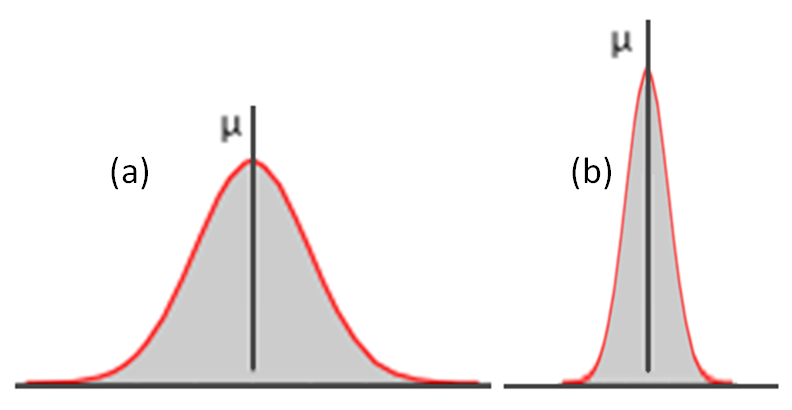


Figura ‑: Distribuição (a) sem *shrinkage* e (b) com *shrinkage*

Com a utilização de *shrinkage*, os pontos da distribuição ficaram mais próximos da média, diminuindo a variância. Essa é a idéia chave para melhorar as estimativas.

*Shrinkage* é uma técnica especialmente utilizada em problemas de classificação hierárquica (McCallum, et al., 1998): explora os caminhos de uma hierarquia taxonômica tentando melhorar as estimativas dos descendentes em função de dados dos ancestrais através de combinação linear (Chakrabarti, 2002). Talvez isso explique o desempenho do NB-*Shinkage* ao ser aplicado para mapear conceitos de ontologias, naturalmente organizados em taxonomia, o que ficou claro nos experimentos desta Dissertação (Capítulo 0). Apesar da nossa abordagem não ter implementado exatamente uma classificação hierárquica e sim uma propagação hierárquica da classificação, ao que parece, o NB-*Shinkage* beneficiou-se da hierarquia assim mesmo. Disso, pode-se concluir que s*hrinkage* se comporta bem ao ser aplicado em ontologias, e pode ser explorado mais profundamente em trabalhos futuros.

## PRECISÃO, REVOCAÇÃO E MEDIDA-F

Esta seção trata das métricas de avaliação utilizadas neste trabalho para mensuração da eficácia ou qualidade de resposta do sistema de mapeamento. Um mapeamento entre conceitos eé classificado nos conjuntos:



Figura ‑: Conjuntos Ideal, Computado e Acerto

O **conjunto Ideal** contém mapeamentos estabelecidos por especialistas humanos, por isso mapeamentos deste conjunto são considerados confiáveis e assumidos como verdade. O **conjunto Computado** é equivalente à saída do mapeador a ser avaliado. Da interseção entre estes conjuntos, surge um terceiro: o **conjunto Acerto** contendo mapeamentos corretamente computados. Uma vez que mapeamentos sugeridos automaticamente podem não coincidir com a realidade, a idéia é confrontar os conjuntos Ideal e Computado e esperar que a interseção entre eles seja a maior possível: quanto maior o conjunto Acerto, maior a qualidade da resposta do sistema.

Para quantificar a relevância de um conjunto Computado resultante de experimentos de Recuperação de Informação (R.I.), mensura-se a **corretude** e a **completude** do sistema. Para isso, utilizam-se três métricas principais variando dentro da escala [0,1]: **precisão**, **revocação**e **medida-F**, definidas como segue:

|  |  |
| --- | --- |
| Equação : | Equação : |
| Equação : | |

A precisão é a métrica de corretude. Serve para medir a quantidade de acertos computados em relação ao total de respostas. O desejável é que a quantidade de respostas ideais supere a quantidade de respostas erradas, que funcionam como “lixo” afetando negativamente a precisão. No melhor caso da precisão, o lixo é inexistente.

A revocação é a métrica de completude ou abrangência da resposta. Serve para medir a quantidade de acertos computados em relação ao total de acertos possíveis: mesmo que venha lixo na resposta, deseja-se que todas as respostas ideais sejam encontradas, o que seria o melhor caso da revocação.

Sendo assim, para obter valores máximos de precisão e revocação, os resultados computados devem conter **apenas mapeamentos corretos** e **todos os mapeamentos corretos**, quando então os conjuntos Ideal, Computado e Acerto serão equivalentes. Infelizmente, isso dificilmente acontece, pois é comum encontrar valores de precisão e revocação desequilibrados: precisão alta e revocação baixa não é sinal de bom resultado e vice-versa. Para avaliar o equilíbrio entre as duas métricas, costuma-se combiná-las numa terceira métrica: a medida-F, que é a média harmônica entre precisão e revocação.

Quando temos valores parciais de precisão, revocação e medida-F computados para vários objetos, mas queremos um valor global, precisamos pensar na **macro-avaliação** e na **micro-avaliação** (Chakrabarti, 2002). Seja um conjunto composto por objetos a avaliar, a macro-precisão e a macro-revocação são as médias das avaliações parciais, sendo que a macro-medida-F é a média harmônica destas duas métricas:

Equação : 

Equação : 

Por outro lado, a micro-precisão e a micro-revocação são calculadas normalmente em função do somatório de acertos, de respostas ideais e de respostas computadas para cada objeto individual, sendo a micro-medida-F a média harmônica destas duas métricas.

Equação : 

Equação : 

No caso do mapeamento semântico aqui desenvolvido, utilizamos a macro-avaliação para obter a avaliação global dos mapeamentos produzidos pelo L-*Match* (Seção 5.3), em função das avaliações parciais de cada tipo de mapeamento, uma vez que existe um conjunto Ideal, um Computado e um Acerto para cada relação *R* utilizada como axioma ponte:

* ***Ideal*** é constituído pelos mapeamentos idealmente qualificados por *R*;
* ***Computado*** é constituído pelos mapeamentos automaticamente qualificados por *R*;

Espera-se que o mapeador acerte a relação *R* para cada par de conceitos, pois quando erra, diminui a precisão da relação computada (pois insere um mapeamento errado) e diminui a revocação da relação ideal (pois não foi capaz de encontrá-la).

Capítulo 3

# TRABALHOS RELACIONADOS

Neste capítulo, serão revisadas as idéias principais das referências que mais contribuíram com a presente Dissertação, focando nos trabalhos mais relevantes voltados a integração de informação no que diz respeito a formalismos, utilização de aprendizado de máquina e a mapeamento semântico.

Sabe-se que o problema de mapeamento semântico consiste em estabelecer regras axiomáticas entre entidades conceituais declaradas em diferentes estruturas semânticas. Contudo, antes de pensar em técnicas para resolver o problema, existe a necessidade de organizar, modelar e formalizar os conceitos sobre mapeamento. Nesta área, (Rahm, et al., 2001) propõe uma taxonomia para classificar a maioria das abordagens sobre mapeamento de esquema até então desenvolvidas, numa primeira tentativa de padronizá-las. Nesta taxonomia, as abordagens são classificadas como estando no nível de esquema ou no nível de instância, no nível de estrutura ou no nível de elemento, podendo ser baseadas em linguagem natural e em restrições. Seguindo esta taxonomia, (Euzenat, et al., 2003) propõe classificar as abordagens de acordo com a fonte de evidência e a estratégia utilizadas para comparar conceitos, distinguindo-se cinco classes: comparação terminológica, comparação da estrutura interna, comparação da estrutura externa, comparação extensional e comparação semântica. Esta classificação é mais adequada para o estado da arte e foi adotada nesta Dissertação e em vários outros trabalhos como (Euzenat, et al., 2004) e (Fürst, et al., 2005).

Uma nova e diferente abordagem para mapeamento, incluindo formalismo e algoritmo, começa a aparecer nos trabalhos de (Bouquet, et al., 2003) e (Magnini, et al., 2003) com respeito à *coordenação semântica*. Observando a falta de infra-estrutura para especificação de mapeamentos semânticos, (Bouquet, et al., 2003) apresenta uma linguagem chamada *Context* OWL(C-OWL) que estende a sintaxe e a semântica do OWL convencional para coibir certas operações e permitir a especificação de mapeamentos semânticos entre ontologias distintas, o que atualmente não é suportado. Porém, a sua maior contribuição diz respeito à formalização das tarefas de comparação semântica e de mapeamento semântico. Sugere-se mapear cada par de conceitos aplicando um axioma ponte que seja capaz de expressar a relação semântica mais forte entre os conceitos, sendo possíveis relações triviais de **equivalência (), mais geral (⊐), menos geral (⊏), sobreposto () e diferença ()**. Para comparar os conceitos, um raciocinador lógico é consultado através de expressões pré-definidas especialmente para cada relação, introduzindo o uso de raciocinadores baseados no problema da Satisfabilidade (SAT) para resolver o problema de mapeamento. Para montar as consultas, as ontologias são primeiramente contextualizadas a fim de permitir a atuação do raciocinador: cada conceito *C* é convertido em uma conjunção lógica que representa o contexto de *C* na ontologia, o qual serve como fonte de evidência. Porém, estes contextos ainda são limitados, pois são extraídos apenas da taxonomia, como a maioria dos trabalhos, e não incluem outras regras e restrições. Contudo, esta abordagem representa um grande avanço por permitir que, pela primeira vez, o problema de integração de ontologias produza mapeamentos semânticos e possa ser tratado como um problema de raciocínio lógico. Mais tarde, (Stuckenschmidt, et al., 2004) testou a linguagem C-OWL durante o alinhamento e a fusão de ontologias médicas, concluindo que este é um formalismo adequado para o suporte de mapeamentos complexos.

A partir destas idéias, um grande corpo de trabalho vem sendo desenvolvido, sobrepujando em qualidade várias outras abordagens correntes que praticam tão somente mapeamento sintático. A primeira experimentação envolvendo geração de contexto e raciocinadores SAT culminou com o desenvolvimento do mapeador semântico chamado de *Context Match* ou Ctx-*Match*, descrito e avaliado nos trabalhos de(Bouquet, et al., 2003) e (Magnini, et al., 2004) através de mapeamentos entre hierarquias de classes dos diretórios do Google e do Yahoo. As avaliações do Ctx-*Match* incluem apenas as relações de **equivalência (), mais geral (⊐)** e **menos geral (⊏)**. O Ctx-*Match* é muito mais preciso do que completo, notavelmente desequilibrado, pois produziu valores de precisão geralmente na casa dos 80%, alguns ultrapassando 90%, para relações de mais e menos geral, ao custo de valores de revocação baixos, geralmente ao redor de 50%. Para a relação de equivalência observou-se revocação extremamente baixa, variando entre 4% e 7% até 10% e 13%. A relação de equivalência também não obteve os melhores resultados de precisão, variando entre 27% e 33% até 71% e 78%. Estes resultados são ruins, pois para encontrar equivalência, que é a relação mais importante, mapeadores sintáticos funcionam muito melhor. Em média, a relação de menos geral foi a que demonstrou os melhores resultados em todos os experimentos, o que pareceu ser uma tendência, seguida pela relação de mais geral. A mesma tendência é observada nos experimentos do L-*Match* (Seção **Erro! Fonte de referência não encontrada.**), que nem por isso é desequilibrado ou se comporta mal com as equivalências.

Apesar dos resultados ruins, a abordagem anterior é promissora. Sendo assim, (Giunchiglia, et al., 2005) deu andamento às idéias primeiramente experimentadas no Ctx-*Match*, implementando um novo mapeador semântico batizado de *Semantic Match* ou S-*Match*. Diferentemente do seu precursor, o S-*Match* conseguiu obter qualidade de resposta notável, inclusive quando comparado com três atuais sistemas de mapeamento, no caso Cupid (Madhavan, et al., 2001), COMA (Do, et al., 2002) e Rondo (Melnik, et al., 2003). Isso ocorreu em grande parte devido a várias melhorias, incluindo otimizações e uma extensiva biblioteca de estratégias paliativas de mapeamento combinadas. Só então a abordagem ganhou reconhecimento, sendo considerada a abordagem com maior sucesso até agora.

Apesar das metodologias de comparação entre conceitos do S-*Match* e do nosso L-*Match* serem diferentes, o S-*Match* constitui uma das principais referências desta Dissertação, pois o mapeamento produzido pelos dois sistemas é igualmente semântico e qualificado pelas mesmas cinco relações axiomáticas. A classe de sistemas representada pelo S-*Match* faz comparação e mapeamento apenas no nível de esquema, pois não considera informações contidas nas instâncias, ao contrário do L-*Match*, que atualmente inicia as comparações no nível de instância para criar mapeamentos no nível de esquema. Como na maioria dos trabalhos, S-*Match* e L-*Match* são restritos à taxonomia por não considerar ainda as regras e restrições. Esta restrição pode não ser tão grave: como a fundamentação do S-*Match* e do OWL é em Lógica de Descrição, podemos futuramente generalizar a taxonomia, entendendo que conceitos nomeados também são subclasses de conceitos anônimos decorrentes de regras e restrições.

Porém, para trabalhar apenas no nível conceitual, o S-*Match* acaba tornando-se dependente do *WordNet*, que funciona como uma ontologia global impondo mapeamentos locais. O S-*Match* pode ser dividido em duas grandes etapas de mapeamento: a primeira etapa gera mapeamentos iniciais extraídos integralmente do *WordNet*, enquanto que a segunda etapa aplica raciocínio sobre os mapeamentos iniciais, reacertando-os através de solucionadores SAT. Sabe-se que o problema da SAT é NP-Completo, e que portanto as soluções não são ótimas e sim heurísticas, podendo ser muito demoradas, prejudicando o desempenho do S-*Match*. De qualquer forma, a qualidade das conclusões deste tipo de raciocínio é interessante, podendo futuramente ser agregado ao L-*Match* para fins de experimentação, criando um híbrido entre o L-*Match* e o S-*Match* onde não existe a imposição global do *WordNet*.

É importante corrigir afirmações feitas nos trabalhos envolvendo Ctx-*Match* e S-*Match* a respeito de mapeamento utilizando classificação automática de instâncias. (Bouquet, et al., 2003) critica a utilização de instâncias por sua atual escassez nas ontologias, esquecendo-se que sua abordagem é extremamente dependente do *WordNet* , já que não apresentou alternativas ao dicionário, tendo por isso grandes problemas de portabilidade, já que atualmente o *WordNet* só está disponível para língua inglesa, além de ser um conhecimento incompleto, estático e que impõe sua visão de mundo nos mapeamentos locais, o que não é bem aceito por grande parte da comunidade de Recuperação de Informação. Por sua vez, (Magnini, et al., 2004) afirma que classificadores são inviáveis quando a natureza das instâncias não é textual, como é o caso de imagens. Isso não é verdade, primeiro porque instâncias convencionais de ontologias podem facilmente ser renderizadas em texto (Seção 4.1) e entendidas como documentos virtuais (Qu, et al., 2006). Segundo porque uma das principais aplicações de classificação automática encontra-se justamente em Recuperação de Informação em imagens, Processamento Digital de Imagem e Visão Computacional, incluindo *Optical Character Recognition* (OCR), filtros de pornografia, processamento de imagens de satélite e radar, etc. Alguns exemplos são (Shiba, et al., 2005) que avalia cinco classificadores aplicados sobre imagens de radar, (Braga, et al., 2006) e (Ribeiro, et al., 2005) que comparam classificadores supervisionados e não-supervisionados sobre imagens relativas ao mapeamento do uso de solo e à cobertura terrestre de solo e (Kumar, et al., 2006) que classifica objetos de uma imagem usando informação de cor e textura como evidências. Além disso, ainda é possível citar a classificação de recursos da Web, a qual utiliza *hyperlinks* associados ou não com texto ou imagem, a exemplo de (Calado, et al., 2003).

Quanto à metodologia de utilização de instâncias, o trabalho que merece maior destaque em relação ao nosso é o mapeador GLUE (AnHai, et al., 2004), que é o sistema mais relevante com respeito à aplicação de classificação supervisionada para mapeamentos entre conceitos. O GLUE também tem duas grandes etapas: a primeira utiliza o classificador *Naive Bayes* para descobrir valores absolutos de sobreposição entre conceitos, que então são transformados em valores de similaridade através da métrica de *Jaccard*; na segunda etapa os mapeamentos são estabelecidos através de *relaxation labelling*, uma técnica utilizada no GLUE para otimizar e relaxar regras taxonômicas, entre outras. Os resultados experimentais mostraram que a precisão do GLUE é alta, entre 70% a 90%, quando combina todas as evidências disponíveis e utiliza o *relaxation labelling*.

Certamente, os resultados obtidos com o GLUE justificam a utilização de Aprendizado de Máquina. Contudo, a sua abordagem nesse ramo foi feita de maneira bastante trivial, sem explorar diversas outras possibilidades. A eficácia do GLUE é muito dependente do *Naive Bayes*, o único algoritmo de classificação considerado, enquanto existem alternativas possivelmente melhores a serem investigadas e combinadas. Além disso, a metodologia de classificação do GLUE é dispendiosa, pois o classificador precisa ser re-treinado várias vezes, na tentativa de tratar o problema como uma *classificação binária*: para cada conceito *C*, o classificador é re-treinado em duas classes compostas por *C* e pela *classe complemento de C*, que é o conjunto de todos os outros conceitos declarados na ontologia que não são *C*. Este tipo de classificação costuma ser mais fácil de fazer e ter bons resultados, mas apenas quando o problema realmente é de classificação binária, como em detecção de spam e não-spam, imagem pornográfica e não-pornográfica, imagem de desmatamento e não-desmatamento, etc. Ontologias naturalmente possuem várias classes, sendo mais apropriada uma *classificação multiclasse* onde o classificador é treinado apenas uma vez, como normalmente é feito. Além de perdas no desempenho, a classe complemento facilmente se torna muito maior que o conceito em questão, abrindo margem para problemas de desbalanceamento de classes, um dos piores que podem ocorrer durante um processo de classificação supervisionada.

Em relação à estrutura, a abordagem de classificação do GLUE não facilita explorar a taxonomia dos conceitos, não sendo surpresa que classificação hierárquica não é feita e nem induzida. Apesar de utilizar classificação e *relaxation labeler* como ferramentas de Raciocínio Probabilístico, o GLUE não é semântico (como se diz ser) e sim sintático, pois não qualifica os mapeamentos através de axiomas ou relações, e sim através de um coeficiente numérico na escala [0,1]. Apesar de *relaxation labeler* ser eficiente, funcionar relativamente bem e poder gerenciar um grande número de regras, não se conhece muito bem como ocorre a sua convergência até as soluções, podendo convergir pra máximos locais. Isso impede maior domínio sobre o que está acontecendo, dificultando entender por que alguns mapeamentos funcionam e outros não, o que permitiria manutenções evolutivas no algoritmo.

Em (Ichise, et al., 2001) e (Ichise, et al., 2003) foi proposto o mapeador HICAL que também explora a sobreposição de instâncias entre duas taxonomias para determinar mapeamentos, utilizando o método *k-statistics* (ao invés de *Jaccard*) para mensurar similaridade entre conceitos, obtendo bons resultados. Como no L-*Match*, o HICAL determina seus mapeamentos finais utilizando uma estratégia *top-down* na taxonomia e, apesar de não ter implementado, previu a necessidade de alguma estratégia *bottom-up*, o que existe no L-*Match*. Infelizmente, o HICAL dependente muito que duas taxonomias compartilhassem exatamente as mesmas instâncias e em grande quantidade, o que, na prática, acontece raramente. Isso acontece porque a classificação é não-supervisionada, ignorando completamente o conteúdo das instâncias, inclusive suas propriedades.

Tanto o GLUE quanto o HICAL são limitados a encontrar mapeamentos sintáticos de equivalência, também ditos como 1-1 (um-para-um). Ambos admitem que quando a relação mais forte entre um par de conceitos for uma relação de generalização, não serão capazes de encontrá-la. Outro ponto fraco é em relação à perda de informação. Suas idéias preliminares assemelham-se à do L-*Match*: classificar (GLUE) ou identificar (HICAL) instâncias de uma ontologia *A* nos conceitos de outra ontologia *B* e vice-versa, isto é, classificar em dois direções. Porém, após a classificação, os valores bidirecionais de similaridade são combinados num valor só, perdendo-se então a noção de direção e, com isso, contrariando a idéia de fluxo de informação. Além disso, esta similaridade não é a similaridade natural fornecida por algoritmos de classificação, ela é calculada através de *Jaccard* (GLUE) ou *k-statistics* (HICAL) em função apenas dos valores absolutos de sobreposição, os quais logo em seguida são simplesmente descartados.

Ao contrário dos dois últimos sistemas, o L-*Match* propõe várias modificações: classificação multiclasse, investigação de alternativas ao classificador Naive Bayes (incluindo combinações de classificadores), utilização dos valores de similaridade produzidos pelo classificador bem como dos valores absolutos de sobreposição, manutenção destes valores na direção em que foram calculados, extração mais elaborada de características das instâncias, uma alternativa mais controlada ao *relaxation labeler* e, por fim, mapeamentos do tipo *n-n* (muitos pra muitos) qualificados por axiomas e não por coeficientes numéricos.

Outros trabalhos também utilizaram instâncias e/ou técnicas de Aprendizado de Máquina para resolver diferentes problemas envolvendo taxonomias e ontologias, especialmente na área de Aprendizado de Ontologia. Em (Buitelaar, et al., 2005) apresenta-se o estado da arte sobre aquisição automática de conhecimento na forma de conceitos, taxonomias, relações e regras. Exemplos são (Snow, et al., 2006) e (Cimiano, et al., 2004) que criam ou evoluem taxonomia. Mas o grande exemplo é (Huang, et al., 2008) que utiliza instâncias de proteínas anotadas na *Gene Ontology* pra treinar o classificador SVM e um algoritmo genético. Como a *Gene Ontology* é grande e complexa, o objetivo é auxiliar especialistas que fazem anotações na ontologia, sugerindo automaticamente a localização subcelular de novas proteínas recentemente descobertas e que também precisam ser anotadas.

Por fim, existem vários outros sistemas de Integração de Informação bastante conhecidos, como o *Anchor*-PROMPT (Noy, et al., 2001), o *Chimerae* (McGuinness, et al., 2000), o FCA-*Merge* (Stumme, et al., 2001), o IF-*Map* (Kalfoglou, et al., 2003) e o COMA++ (Aumueller, et al., 2005) que é a versão mais atual do COMA. Há também bastante trabalho relacionado à (Euzenat, et al., 2003) que culminou com desenvolvimento do sistema OLA (Euzenat, et al., 2004) específico para mapear ontologias em OWL, mas que não utiliza todo recurso de inferência possível em OWL, como Lógica de Descrição, similar ao que ocorre com o L-*Match* nesta Dissertação. Atualmente, similar às idéias do S-*Match*, a tendência dos novos trabalhos é adicionar etapas de raciocínio Lógico ao processo de integração de ontologias, como no sistema ILIADS desenvolvido nos recentes trabalhos de (Udrea, et al., 2007) que combina inferência estatística e lógica para fundir ontologias. O ILIADS não apresentou resultados muitso melhores ao ser comparado com outros sistemas de alinhamento, que justifique a sua inferência lógica incompleta, seu processamento demorado e o exagero de ter 12 limiares a otimizar.

Na tabela a seguir, estão resumidos alguns dos sistemas citados neste capítulo, incluindo o L-*Match.* A tabela foi montada seguindo critérios do tutorial de (Shvaiko, et al., 2005). Observe que o fato de um sistema utilizar instâncias não significa que o mesmo tenha aplicado Aprendizado de Máquina em sua abordagem. Por isso, os sistemas que associam instância a Aprendizado de Máquina possuem marcação extra feita com um asterisco (\*).

Tabela ‑: Sistemas de Integração

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SISTEMA | | NÍVEL DE ELEMENTO | | NÍVEL DE ESTRUTURA | | INSTÂNCIA | ESQUEMA |
| **Sintaxe** | **Conhecimento Externo** | **Sintaxe** | **Semântica** |
| 2003 | **HICAL** | - | - | Classificação  Não-Supervisionada,  k-statistics | - | **√\*** | - |
| 2004 | **GLUE** | Language-based | - | Naive Bayes, Jaccard, Relaxation Labeller | - | **√\*** | - |
| 2004 | **OLA** | String-Based, Language-Based, Datatypes | WordNet | Iterative Fix-Point Computation, Neighborhood Mathing | - | **√** | **√** |
| 2002  e  2005 | **COMA e**  **COMA++** | String-Based, Language-Based, Datatypes | Auxiliado por Dicionário Global | Diversas Estratégias | - | **√** | **√** |
| 2004 e 2005 | **Ctx-M*atch* e S-*Match*** | String-Based, Language-Based | Dependente de Dicionário Global (WordNet) | - | Propositional SAT Solvers | - | **√** |
| 2008 | L-*Match* | Language-Based, Entropia,  Grafos | Auxiliado por Thesaurus (WordNet) | Classificação Supervisionada Propagada Bottom-Up e Combinada utilizando  NoisyOr+Entropia | Dedução Top-Down, dada por Sobreposição | √\* | - |

Capítulo 4

# MAPEAMENTO SEMÂNTICO ENTRE ONTOLOGIAS

A denominação L-*Match* vem do inglês *Learning Match* e foi dada ao sistema desenvolvido neste trabalho, pois seu funcionamento é centrado em Classificação Supervisionada, uma subárea de Aprendizado de Máquina. Neste contexto, propomos uma abordagem para a identificação de mapeamentos semânticos e locais entre conceitos de ontologias distintas com base na mensuração da interseção entre os conceitos estimada por classificadores.

Investigamos características de instâncias de ontologias OWL para então descobrir informações sobre conceitos e finalmente, utilizando a estrutura taxonômica como guia, estabelecemos os mapeamentos. Instâncias são ricas em evidências e funcionam na prática como canalizadores da informação do esquema da ontologia: obtemos características como conceitos, propriedades, generalizações, além de tipos e valores de propriedades em cada instância. As instâncias são então convertidas em texto em função de suas características que são separadas em escopos textuais, estabelecendo com isso uma entrada genérica para diferentes classificadores supervisionados, cuja saída é utilizada no mapeamento.

O interesse de estimar a interseção ou sobreposição entre conceitos surgiu principalmente pela seguinte razão: a atual eficácia dos métodos de classificação torna viável a implementação da **relação de compatibilidade entre contextos locais definida pela Semântica de Modelos Locais** (Ghidini, et al., 2001)**, veja a** Seção 4.3. Já o interesse em mapeamento semântico é relevante, uma vez que o único mapeador realmente semântico desenvolvido até hoje é o S-*Match* (Giunchiglia, et al., 2005), e antes dele todos os mapeadores eram sintáticos. O L*-Match* surge como abordagem alternativa ao S-*Match*, estimulando e continuando pesquisas sobre mapeamento semântico.

O S-*Match* e L-*Match* são estruturalmente restritos à taxonomia por não considerar ainda regras e restrições genéricas. Esta restrição pode não ser tão grave, pois S-*Match* e OWL são fundamentados em Lógica de Descrição, por isso podemos futuramente generalizar a taxonomia, entendendo que regras e restrições criam superconceitos anônimos contendo conceitos nomeados. Como o S-*Match*, o L-*Match* também tem dependências: como todo sistema de aprendizado de máquina supervisionado, depende da quantidade de instâncias no conjunto de treinamento, atualmente escassas em ontologias OWL. Contudo, esta escassez se deve em muito ao fato de ontologias constituírem ainda uma tecnologia nova, cujo uso ainda não é tão difundido quanto o de diretórios Web e bancos de dados relacionais. Mas isso não é justificativa para ignorar as instâncias, afinal são evidências que representam conhecimento especialista local e, se for o caso, a escassez pode ser contornada por abordagens de instanciação automática que mineram a Web.

Ao todo, foram utilizadas três estratégias de comparação (Subseção **Erro! Fonte de referência não encontrada.**): comparação terminológica, extensional e da estrutura externa. Portanto, a arquitetura geral do L-*Match* foi estabelecida contendo três módulos: Extração, Similaridade e Mapeamento. Além disso, a abordagem é iterativa, pois o sistema pode reutilizar os mapeamentos computados como conhecimento adquirido. Veja na figura a seguir:

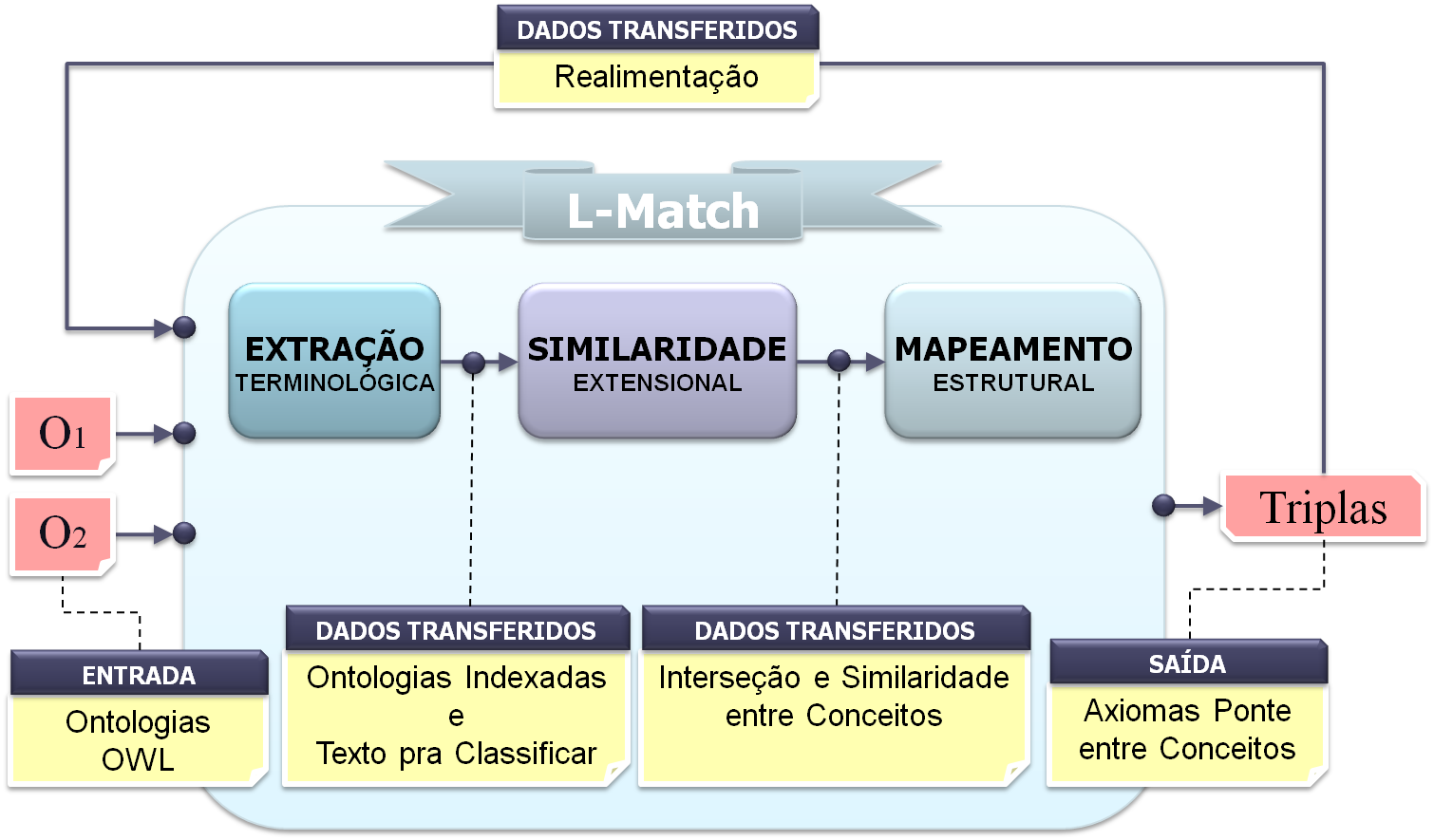


Figura ‑: Arquitetura do L-*Match*

Este capítulo descreve a abordagem proposta nesta Dissertação. Portanto, nas seções seguintes, os três módulos acima serão detalhadas: será explicado como gerar e melhorar a qualidade do texto para cada instância, como manipular e combinar classificadores para obter similaridade e interseção entre conceitos, como propagar estes valores numa taxonomia e, por fim, como obter os mapeamentos finais na forma axiomática.

## EXTRAÇÃO

Este módulo é responsável por pré-processar as ontologias a mapear. Ontologias são formadas basicamente pelas mesmas entidades (conceitos, propriedades, instâncias e regras) e por isso nós as padronizamos num formato interno normal que independe do formato inicial (apenas OWL por enquanto), facilitando o mapeamento. As entidades são indexadas porque precisarão ser tabeladas nos módulos seguintes, mas a principal funcionalidade deste módulo é gerar texto para classificar, o que iremos discutir nas subseções seguintes.

### Geração de Texto para Classificar

Instâncias em OWL (como em outros formatos de ontologia) não são documentos de texto e sim estruturas em XML/RDF. Por isso, primeiramente são renderizadas num conteúdo textual que é submetido aos classificadores de texto do módulo seguinte. Ao contrário do GLUE (Doan, et al., 2002), não consideramos conteúdo textual de páginas Web relacionadas a ontologias, consideramos apenas instâncias fornecidas com as próprias ontologias, o que nos dá certa autonomia ao gerar texto para classificar.

Ontologias fornecem informações precisas sobre suas instâncias: nome, conceitos, super conceitos e propriedades com valor e tipo definidos. Por exemplo, em uma de nossas ontologias, ***WaterTemperature*** instância diretamente o conceito *TemperatureOf* e indiretamente os seus super-conceitos (*EcologicalData*, *Quantity* e *TemperatureQuantity*), relacionando-se com *celsius* (instância direta de *ScaleOfTemperature*) e *water* (instância direta de *EcologicalEntity*)por meio das propriedades *hasTemperatureScale* e *hasEntity*, respectivamente. Na configuração atual do L-*Match*, as características foram renderizadas e discriminadas em quatro blocos dependendo de sua origem, veja:

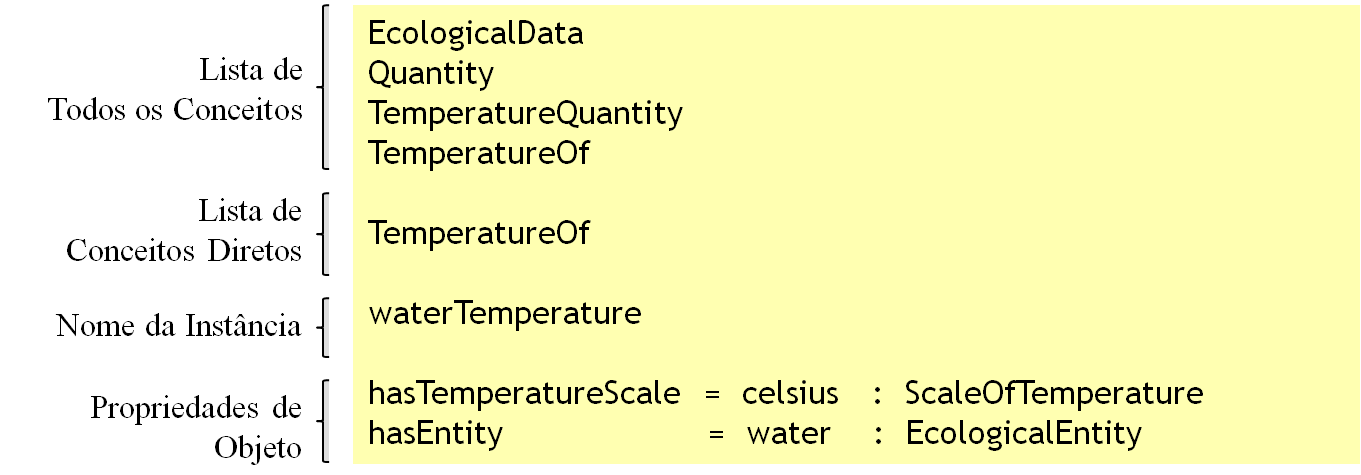


Figura ‑: Texto básico gerado para a instância *WaterTemperature*

O texto gerado é pequeno e preciso, pois é composto por palavras chave. Observe que conceitos diretos são discriminados, porém super conceitos não: essa configuração foi escolhida empiricamente porque propiciou os melhores valores de avaliação nos experimentos. Mas o texto precisa ser tratado: os nomes são convertidos para minúsculo, separados em *tokens* e remove-se as *stopwords* (*to*, *has*, *of*, *the*, etc.) deixando apenas as *keywords* (palavras-chave). O resultado é o seguinte:

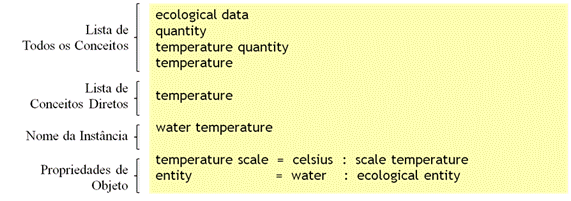


Figura ‑: Texto tratado para a instância *WaterTemperature*

A realimentação iterativa do L-*Match* é responsável por uma das mais importantes alterações feitas no texto: assumindo que o L-*Match* é capaz de gerar mapeamentos precisos, a idéia é retreiná-lo utilizando conhecimento compartilhado descoberto a partir dos mapeamentos computados. Neste caso, a cada iteração de mapeamento, aproveitamos os mapeamentos de equivalência, mais e menos geral para atribuir conceitos de uma ontologia às instâncias de outra ontologia, e vice-versa, para então regerar e atualizar o texto de cada instância. Essa não é toda a informação que pode ser reutilizada, contudo é a mais óbvia.

Por fim, instâncias como *WaterTemperature* estão quase pronta para serem classificadas, falta ainda ver estratégias mais elaboradas para melhorar a qualidade do texto na subseção seguinte.

### Qualidade do Texto

Deve-se lembrar que o resultado do Módulo de Extração será a entrada de classificadores de texto: por melhor que seja o algoritmo de classificação, ele não funcionará satisfatoriamente se os dados submetidos a ele forem de má qualidade. Portanto, o Módulo de Extração é muito importante, uma vez que influenciará diretamente a precisão dos classificadores e conseqüentemente do mapeamento final. Esta subseção descreve as estratégias que consideramos essenciais para qualidade: a utilização de escopos textuais associada à identificação de sinônimos ajuda a obter boa qualidade textual. Primeiramente, observe a palavra ***temperature*** ocorrendo em todos os blocos principais do texto seguinte:

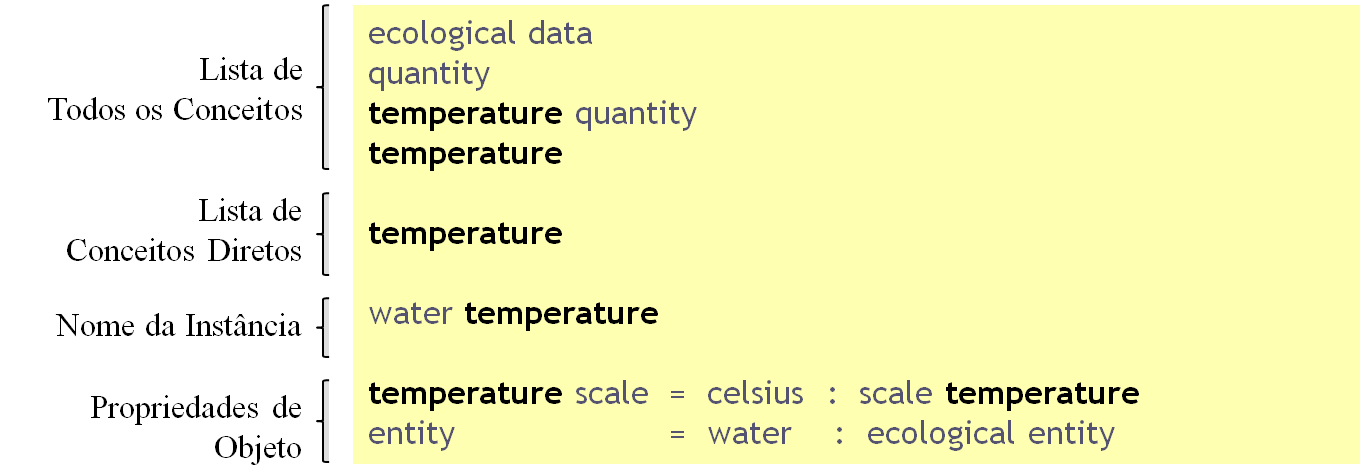


Figura ‑: Palavra repetindo-se em diferentes partes do texto prejudica a qualidade

Blocos têm semântica específica, mas isso não está explícito, causando um efeito negativo. Por exemplo, ***temperature*** ocorrendo como conceito que contém *WaterTemperature* não se diferencia de ***temperature*** ocorrendo como tipo de propriedade: esta instânciapertence aos conceitos *TemperatureQuantity* e *TemperatureOf*, mas não a *ScaleOfTemperature*. Para esta situação não confundir os classificadores, adotamos uma solução simples e eficaz: como classificadores comparam *strings*, prefixamos as palavras originais, gerando novas palavras que discriminam **escopos textuais**, mesmo que originalmente as palavras sejam idênticas:

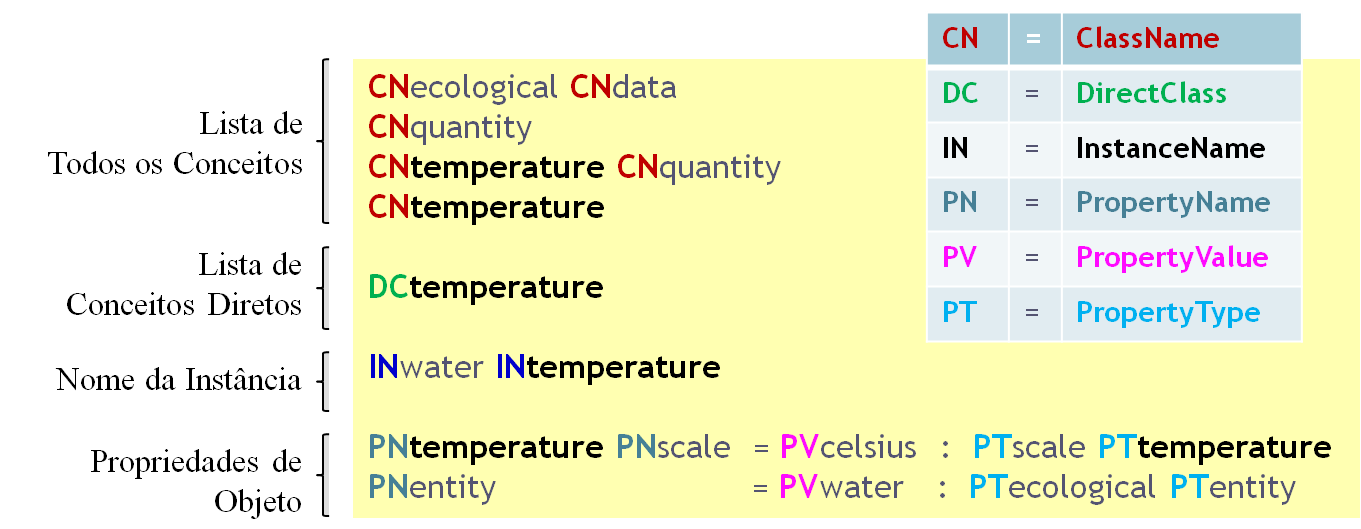


Figura ‑: Adição de prefixos de escopo ao texto

Agora, toda palavra recebe um prefixo diferente dependendo do escopo textual em que aparece, explicitando a semântica do escopo. Para um classificador de texto, esta é uma forma indireta de passar significado. No exemplo, *temperature* não é mais a mesma palavra morfologicamente: CNtemperature, DCtemperature, INtemperature, PNtemperature e PTtemperature. Portanto, os escopos textuais foram utilizados para indicar o contexto de cada característica da instância, ou seja, de que parte da ontologia cada característica é oriunda, evitando com isso maiores ambigüidades.

Agora discutiremos sobre identificação de sinônimos e desambiguação de palavras. Acreditamos que reduzir palavras a uma única forma padrão é útil, pois ajuda a resolver o problema de nomear conceitos equivalentes com palavras diferentes nas ontologias. Experimentos preliminares de *hold out* (reclassificação de instâncias na mesma ontologia) obtiveram altos valores de avaliação. Isso ocorre porque o vocabulário de uma ontologia é restrito e bem formado (específico, controlado, etc.), sem tanta “impureza” como, por exemplo, em hypertexto. Porém, vocabulários distintos apresentam muitas diferenças quando são mapeados, logo é interessante reduzir estas diferenças na medida do possível para, mais uma vez, facilitar a ação dos classificadores, os quais comparam *strings*.

A técnica terminológica aplicada durante a extração visa amenizar diferenças entre os vocabulários de um par de ontologias, estabelecendo um terceiro vocabulário comum que seja o maior possível. Palavras são reduzidas morfologicamente através de *stemming* e semanticamente através de *synsets* (grupos de sinônimos) para uma forma padrão e única. Descobrir *synsets* automaticamente seria ideal, porém é mais prático utilizar conhecimento externo via *thesaurus*. O algoritmo de *stemming* é o *Porter Stemmer[[5]](#footnote-6)* (Porter, 1997) e o serviço de *thesaurus* é o *WordNet* (Fellbaum, 1998). Veja exemplos de reduções:

Porém, o *WordNet* é incompleto porque não inclui todo sinônimo: isso acontece por exemplo com *Primary* e *Fundamental* que obviamente são sinônimos. Felizmente, o *WordNet* interliga palavras similares não consideradas sinônimas, possibilitando descobrir o caminho *Fundamental→Important→Primary* e vice-versa, com o qual geramos um novo *synset*. Contudo, caminhar por *links* de similaridade pode levar muito longe, gerando *synsets* grandes e genéricos em vez de equivalentes, o que é indesejável. Veja:

Para explorar *links* de similaridade do *WordNet*, controlar a abrangência do caminhamento e gerar *synsets* mais equivalentes, procuramos por **cliques** no grafo de similaridade: entendemos num *synset* existe caminho de similaridade entre todo par de palavras, ou seja, as palavras concordam que são sinônimas quando formam um clique entre si. Com isso introduzimos a idéia de **Unidade de Significado ou Clique** (Venant, 2006) como a menor unidade semântica possível, o que tem sido aplicado em trabalhos sobre modelos geométricos de palavras organizadas em coordenada semânticas, como é o caso do Atlas Semântico[[6]](#footnote-7), um dos melhores *thesauri* disponíveis na Web.

O vocabulário das ontologias é transformado num grafo direcionado : palavras são vértices em , sendo *E* composto por *links* de similaridade importados do *WordNet*. O passo seguinte é **enumerar todos os cliques máximos** (Seção ), um problema *NP*-Difícil. Utilizamos um algoritmo exato como em (Bron, et al., 1973), pois o espaço de busca da nossa aplicação é pequeno por natureza: é restrito ao tamanho do vocabulário das ontologias, *synsets* nunca são grandes e palavras similares formam subgrafos pequenos e isolados. Veja o exemplo:

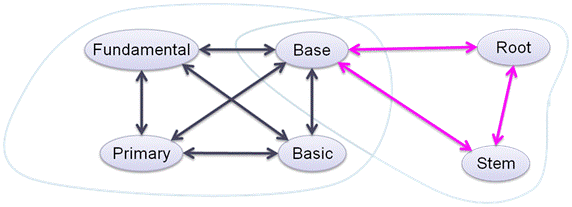


Figura ‑: Unidades de Significado: os cliques serão os novos *synsets*

Na Figura 4‑6 temos um exemplo de **polissemia ou ambigüidade:** a palavra *Base* aparece em dois *synsets* distintos, logo possui dois significados. Isso contrasta severamente com a idéia de reduzir todas as palavras a uma única forma padrão, pois neste caso existem duas possibilidades de redução para *Base*. A solução é eliminar polissemia garantindo que cada palavra esteja em um único *synset*, isto é, gerar um *thesaurus* monossêmico.

O problema de escolher o mais apropriado *synset,* significado ou sentido para uma palavra dentro de um contexto é chamado de **Desambiguação de Sentido de Palavra** ou *Word Sense Disambiguation* (WSD). Trabalhos sobre **Desambiguação Suave** (*Soft Disambiquation*) permitem que palavras continuem tendo mais de um significado no mesmo contexto e criticam a atribuição de apenas um *synset* por palavra. Contudo, as necessidades nesta Dissertação são diferentes, o que nos leva a concordar com (Buitelaar, et al., 2001) sobre que no contexto de um domínio restrito (como é o caso das ontologias) muitos termos ambíguos certamente terão forte preferência por apenas um de seus *synsets*, o qual é específico e maximiza a similaridade entre os domínios das ontologias a mapear. Por exemplo, nos nossos experimentos gostaríamos que os *synsets* da Figura 4‑6: Unidades de Significado: os cliques serão os novos *synsets*fossem desambiguados como segue:

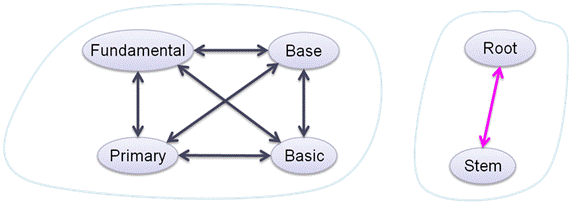


Figura ‑: Apenas *synsets* mais informativos são mantidos, eliminando a polissemia.

Com base em Teoria da Informação, elaboramos uma técnica simples e eficaz para desambiguação, visando eliminar polissemia por completo. Diferente do tradicional, esta desambiguação não depende do contexto taxonômico e sim do contexto do vocabulário do domínio, isto é, apenas os ***synsets*** **mais importantes e informativos** para o domínio continuarão existindo. Os *synset* são ranqueados do menos para o mais importante, quando então se aplica uma estratégia gulosa: percorrer o *rank* do inicio ao fim, penalizando cada *synset* removendo dele as palavras ainda polissêmicas. Ao atingir o final do *rank*, todas as palavras serão monossêmica, muitos *synsets* do inicio do *rank* deixaram de existir e as palavras tenderam a se concentrar nos *synsets* mais importantes do final do *rank*.

Agora precisamos criar nossa **definição de importância**: um *synset* importante possui informação valiosa para o domínio, já um *synset* sobreposto a muitos outros é polissêmico e não traz informação. Intuitivamente, utilizamos a idéia de **entropia** (Seção 2.4) para mensurar a importância de um *synset* dada pela quantidade de informação que ele contém: quando mais polissemia o *synset* possuir, menos informação eleterá, constituindo um grande candidato a eliminação. Para isso, a importância de um synset foi definida em função (a) do total de palavras polissêmicas no *synset* e (b) do total de significados que cada palavra possui, isto é, total de *synsets* que se sobrepõe ao *synset* original.

Sejam dois vetores *A* e *B* criados para um *synset S* com *n* palavras, onde *A* é um vetor de inteiros que contabiliza a quantidade de *synsets* para cada palavra de *S*, e onde *B* é um vetor booleano que marca cada palavra de *S* como polissêmica (*true*) ou monossêmica (*false*). A quantidade de informação em *S* é dada pela combinação *NoisyOr* (Seção *2.3*) da informação contida nos dois vetores *A* e *B*:

Equação 12: 

Perceba que quantificar informação como grandeza oposta à polissemia é intuitivo e ajuda a eliminar *synsets* genéricos que pouco ou nada ajudarão. Análogo à definição de entropia, pode-se mensurar a importância do mesmo *synset* com palavras como segue:

Equação : 

Equação : 

Equação : 

Estas equações alcançam o maior valor de importância “1” quando não existe palavra polissêmica em . A Equação 13 atinge o menor valor de importância “0” quanto toda palavra em é polissêmica, a Equação 14 tende a zero quanto mais significados cada palavra em possuir e a Equação 15 combina as duas equações anteriores. Finalmente, a importância de um *synset* ficou definida nesta Dissertação como segue**:**

Equação : 

Durante os experimentos, a importância de definida apenas pelo cálculo de informação (Equação 12) funcionou muito melhor do que quando definida apenas pelo cálculo do fator (Equação 15). Portanto, a definição de entropia é satisfatoriamente suficiente. Porém, as duas equações funcionam levemente melhor quando combinadas na Equação 16.

Todos os aspectos abordados nesta subseção contribuíram com a precisão da solução deste trabalho. Escopos textuais são essenciais para boa qualidade do mapeamento, enquanto a identificação de sinônimos potencializa sistemas que processam texto, como é o caso do nosso L-*Match*, um mapeador baseado em algoritmos de classificação de texto.

## COMPUTANDO SOBREPOSIÇÃO E SIMILARIDADE ENTRE CONCEITOS

O Módulo de Similaridade do L-*Match* estabelece analogias iniciais entre conceitos de ontologias distintas através de algoritmos de classificação supervisionada, inserindo nesta Dissertação abordagens de Aprendizado de Máquina e de Raciocínio Probabilístico. Nesta seção, discutiremos como manipular classificadores para descobrir dois tipos de analogias entre conceitos: **sobreposição** e **similaridade**, que em seguida sofrem propagação *bottom-up* na taxonomia. Finalmente, a combinação de classificadores também será abordada.

Inicialmente, nada se conhece sobre o que há em comum entre ontologias distintas. Queremos começar a eliminar esta ignorância classificando instâncias de uma ontologia nos conceitos de outra ontologia e vice-versa, ou seja, fazer **mapeamentos de pertinência *instância-conceito* (∈)**, como esquematizado a seguir:

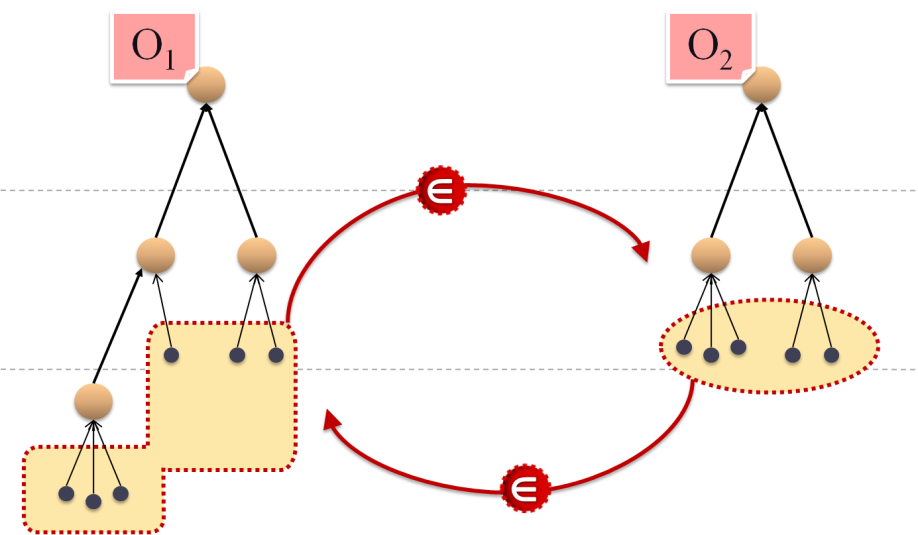


Figura ‑: Descobrindo relações de pertinência entre ontologias

Nosso interesse em classificação automática decorre da sua habilidade de relacionar instâncias desconhecidas (conjunto de teste) com categorias conhecidas (conjunto de treino), com certa precisão e utilizando apenas informação heterogênea: precisamos então adaptar os classificadores para trabalhar com heterogeneidade semântica. Entendemos que ao descobrir relações de pertinência desconhecidas (com ajuda de classificadores) será possível comparar conceitos de ontologias, descobrindo mapeamentos semânticos também desconhecidos até então. Para isso, definimos a arquitetura do **Módulo de Similaridade** do L-*Match* em duas etapas que revezam as ontologias hora como conjunto de treino e hora como conjunto de teste, de forma que as instâncias de uma sejam classificadas nos conceitos da outra:

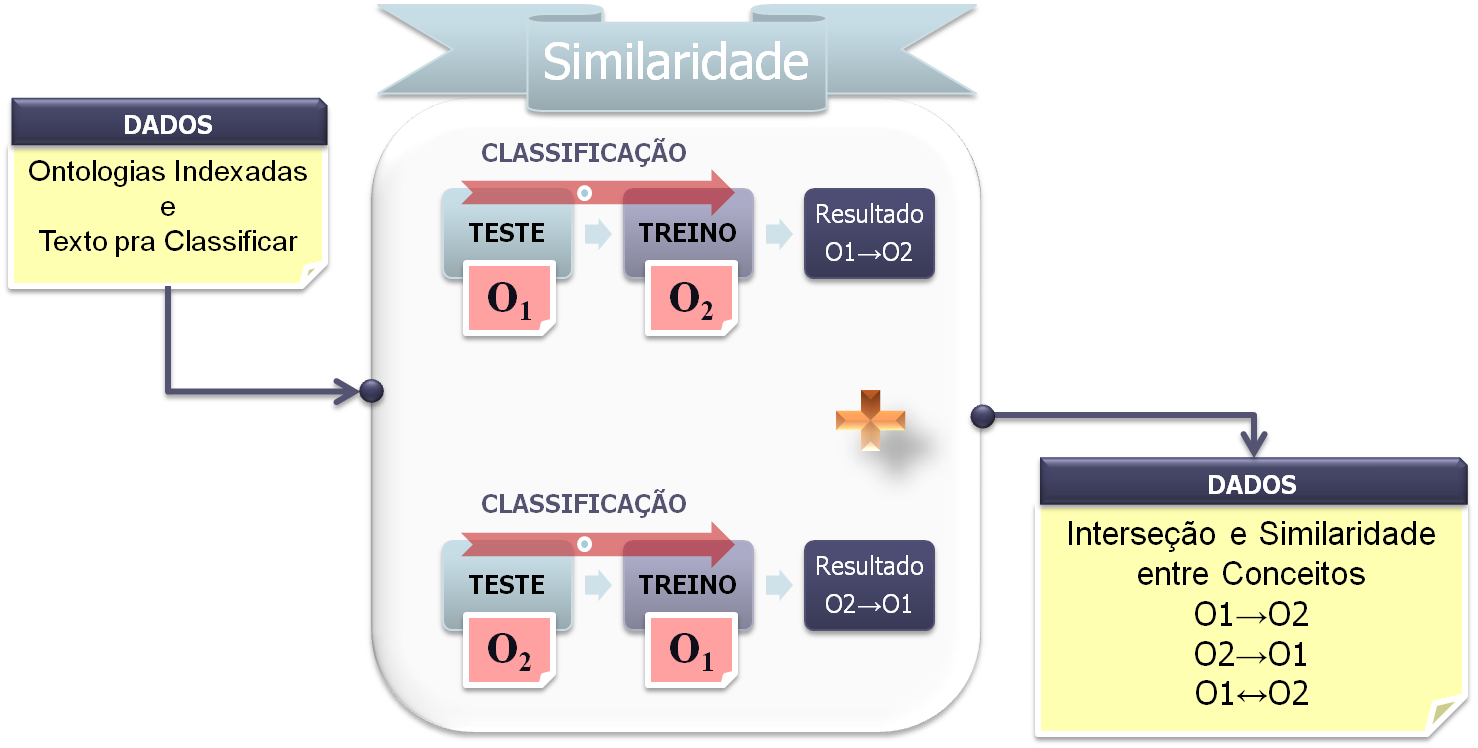


Figura ‑: Classificação automática alternada das instâncias de um par de ontologias

Quando uma instância é submetida a um classificador, o resultado é um *rank* de conceitos ordenados por similaridade decrescente. Por exemplo, na Figura 4‑10 as instâncias do conceito *BaseUnit* foram comparadas por classificadores a conceitos de outra ontologia onde *BaseUnit* não foi declarado:

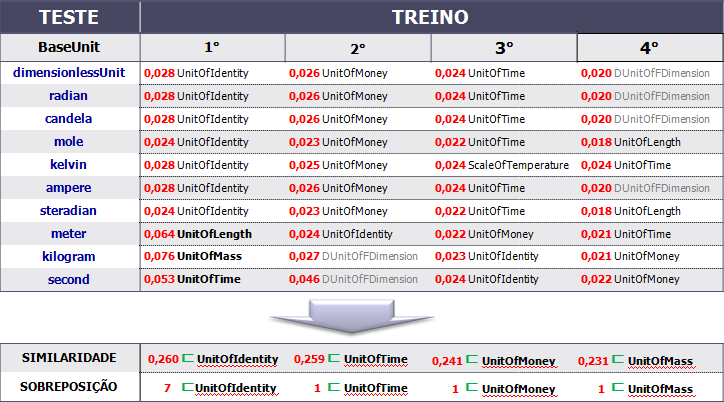


Figura ‑: Ranqueamento de conceitos por similaridade com *BaseUnit* e suas instâncias

A similaridade é computada entre toda instância e todo conceito de treino e depende do algoritmo de classificação. Mas a classificação só ocorre de fato ao atribuir pertinência entre cada instância com alguns conceitos. Para simplificar, adotamos um limiar *Top-k* com *k*=1 que classifica instâncias apenas na primeira classe do *rank* correspondente.

Pertinência e similaridade são relações ***instância-conceito*** inúteis quando sozinhas, contudo podemos combiná-las para descobrir relações ***conceito-conceito***, descrevendo conceitos em função de suas instâncias. Sejam então dois conceitos e , queremos descobrir possíveis relações e para (a) contabilizar as instâncias comuns e descobrir a **sobreposição absoluta**  e (b) combinar as similaridades *instância-conceito* e descobrir a **similaridade *conceito-conceito*** . As combinações são implementadas com *NoisyOr* e entropia, a ser visto na subseção seguinte.

Pela regra , e são estimados nos sentidos (contexto da ontologia de *A*) e (contexto da ontologia de *B*), sendo que é uma combinação que elimina a noção de sentido: trabalhos sobre mapeamento sintático (como o GLUE) perdem informação uma vez que mantêm apenas valores . Por esta e outras razões (Seção 4.3), e são mantidos nos três sentidos.

### Propagação *Bottom-Up* na Taxonomia

Freqüentemente, e são estimados apenas em conceitos específicos: quanto mais específico, maior a chance de um conceito ser diretamente instanciado e, na prática, a maioria das instâncias está nas folhas, impedindo que classificadores comuns processem conceitos genéricos instanciados indiretamente. Mas a semântica taxonômica permite generalizar/propagar e : um conceito é a união () de suas **subsunções** (**instâncias e subconceitos)**.

Como a união equivale à operação ***OR*** e é uma probabilidade, utilizamos o **Modelo *NoisyOr* de Redes Bayesianas** para **propagação *bottom-up*** **de** , desde instâncias e conceitos específicos até o topo genérico de uma taxonomia: as similaridades *instância-conceito,* obtidas por meio de classificadores,são combinadasna similaridade *conceito-conceito* que então é propagada para conceitos mais genéricos. A definição recursiva de no sentido é:

Equação : 

A Equação 17 otimiza o *NoisyOr*: o primeiro produtório combina as similaridades entre B e apenas instâncias diretas de A e o segundo produtório combina as similaridades entre B e os subconceitos de A. Essa configuração evita recombinar as similaridades entre B e as instâncias dos subconceitos de A, que também são instâncias de A.

Por exemplo, a Rede Bayesiana da figura abaixo é utilizada para propagação de similaridade numa taxonomia. Nela, conceitos são representados por esferas maiores e instâncias por esferas menores. Deseja-se então calcular a similaridade entre um conceito qualquer *Q* com todos os conceitos de uma taxonomia, dado que tudo que se sabe é a similaridade entre *Q* e as instâncias da taxonomia:

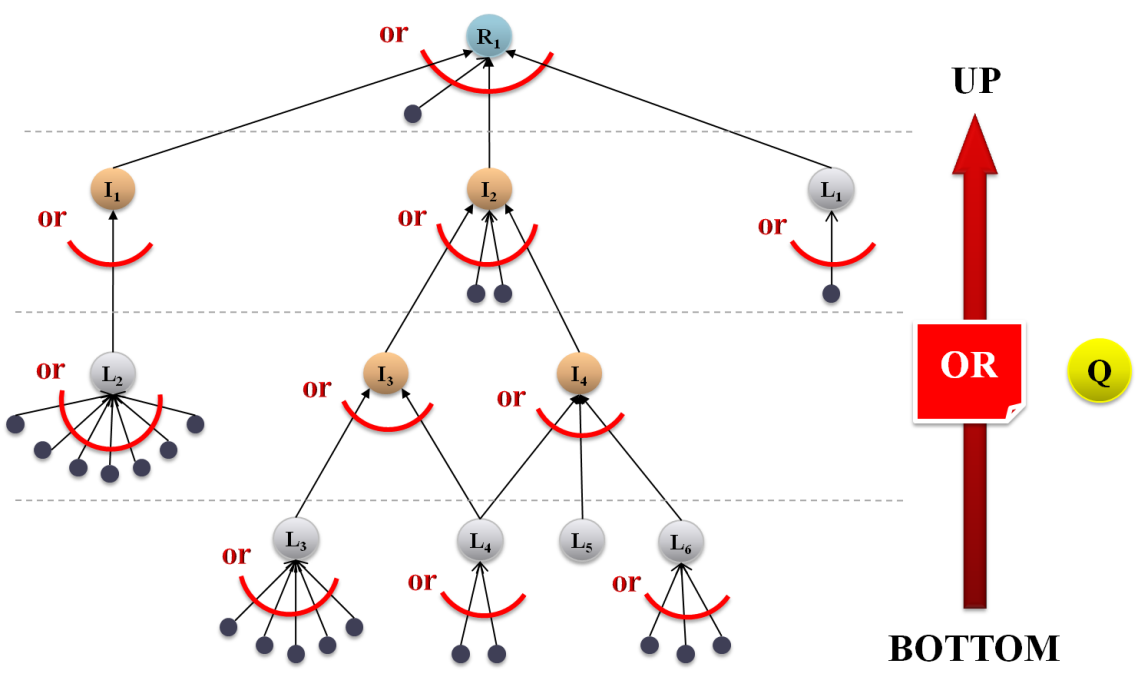


Figura ‑: Rede Bayesiana modelada para propagação de similaridade

Contudo, a propagação de da Equação 17 não está completa, pois precisa agregar conceitos de Teoria da Informação. Propagar é equivalente a sintetizar distribuições de probabilidade: conforme a síntese ocorre através de *NoisyOr*, a probabilidade aumenta (maior disponibilidade de instâncias) mas a quantidade de informação diminui (perda de propriedades específicas); portanto, quanto mais genérico um conceito, menor a quantidade de informação (Resnik, 1990). Recorremos novamente à definição de entropia para alterar a definição recursiva de no sentido :

Equação : 

* Onde *V* é a enumeração de , para todo ou .
* A entropia penaliza a similaridade conforme diminui a informação em .

A perda de informação também ocorre devido a erros de classificação: por exemplo, mesmo que uma instância de um conceito A erroneamente receba grande similaridade com um conceito disjunto B, espera-se que o mesmo não ocorra com a maioria das instâncias de A; certamente a similaridade entre A e B será penalizada pela entropia, para não implicar, por exemplo, numa falsa equivalência entre A e B.

De maneira similar, a **propagação *bottom-up* de**  é feita através da **soma (+)**, que também é análoga à união. A definição recursiva de no sentido é:

Equação : 

Por fim, falta ainda definir as combinações dos valores de similaridade e sobreposição. Para a similaridade, a combinação que teve a melhor avaliação foi a média harmônica, em comparação com *NoisyOr*, média aritmética e *Máx*. Veja:

|  |  |
| --- | --- |
| Equação : | Equação : |

### Combinação de Classificadores

Diferentes algoritmos de classificação foram investigados e comparados em busca da melhor forma de *classificar para mapear*, incluindo *k*NN, *Naive Bayes*, SVM, Máxima Entropia, NB-*Shrinkage* e TF-IDF, que e foram testados individualmente e combinados.

Nossa motivação inicial era o SVM, pois esperávamos que sobrepujasse os outros algoritmos, como comumente ocorre. Mas os experimentos nos surpreenderam com o péssimo desempenho do SVM, enquanto os melhores resultados vieram de onde menos esperávamos: do NB-*Shrinkage*, que inicialmente nem ao menos havia sido cogitado.

*Ensemble* ou *meta-learner* é a denominação de um classificador resultante da combinação de outros. Combina-se as decisões dos algoritmos e não os algoritmos propriamente ditos (embutindo um no outro). Apesar de não haver garantias, espera-se que a eficácia melhore se entendermos a combinação como um processo democrático: se um classificador individual erra, espera-se que a maioria dos classificadores não o faça, discordando do primeiro. Contudo, o processo demora mais à medida que mais classificadores são combinados. O ideal seria que um classificador individual produzisse a melhor eficácia, mas isso nem sempre ocorre, sendo válido investigar *ensembles*.

A combinação convencional de classificadores é feita com *NoisyOr*. Como na Equação 18, preferimos combinar *NoisyOr* com entropia para penalizar respostas controversas quando classificadores discordam muito entre si. Então, seja uma instância de teste, um conceito de treino e um vetor de respostas dadas por classificadores atribuindo diferentes similaridades entre e , temos que:

Equação : 

## COMPUTANDO MAPEAMENTOS SEMÂNTICOS ENTRE CONCEITOS

Os módulos de Extração e Similaridade fornecem satisfatoriamente valores de similaridade e sobreposição entre conceitos de ontologias distintas, valores que na seqüência são utilizados para alimentar o Módulo de Mapeamento, onde finalmente os mapeamentos são estabelecidos em sua forma axiomática e que por isso são semânticos.

### Regras de Compatibilidade entre Conceitos

Apesar de obtermos valores de similaridade e sobreposição, a similaridade apenas ajuda saber o quão equivalentes/diferentes dois conceitos são. Por isso, o que realmente importa para o mapeamento semântico é a sobreposição, pois conceitos são delimitados pela enumeração de suas instâncias e/ou pela definição de suas propriedades.

Podemos mais uma vez diferenciar L-*Match* e S-*Match*, pois o primeiro raciocina sobre a enumeração (estatisticamente e localmente) enquanto o segundo raciocina sobre propriedades taxonômicas (simbolicamente e globalmente) para chegar num objetivo comum: relacionar conceitos através de ***equivalência* (), *mais geral (⊐)*, *menos geral (⊏)*, *sobreposto* ()** e ***diferença* (). Teoricamente, concordamos que este problema remete à relação de compatibilidade entre contextos locais definida pela Semântica de Modelos Locais** (Ghidini, et al., 2001) e que inspirou o **S-*Match*.** Na prática, isso é Teoria de Conjuntos utilizada para regrar interpretações de contextos locais, tornando-os compatíveis.

Sendo assim, a semântica formal dos axiomas **,** , **,** e  **é dada pela compatibilidade entre a classificação de documentos** (Bouquet, et al., 2003)**:** uma função de mapeamento será extensionalmente correta se as condições abaixo se aplicarem às instâncias de dois conceitos arbitrários *A* e *B* oriundos de hierarquias taxonômicas**:**

Onde o símbolo ↓ indica que as instâncias indiretas também são consideradas. Podemos traduzir as regras acima para um algoritmo:

Esta idéia é simples, porém chave, pois através dos valores de sobreposição, obtidos com a ajuda de classificadores supervisionados, implementaremos as regras acima. Contudo, estas regras são muito fortes dado que ocorrem erros de classificação, não permitindo esperar, por exemplo, que dois conceitos compartilhem 100% de suas instâncias para concluirmos que são equivalentes. Portanto, as regras de compatibilidade serão relaxadas através de limiares.

Não utilizamos a popular similaridade de *Jaccard* sobre os valores de sobreposição, pois perderíamos informação valiosa, impossibilitando o mapeamento semântico: os axiomas e são direcionados, assimétricos e inversos o que remete à **direcionalidade do Fluxo da Informação** (Bouquet, et al., 2003),área intimamente ligada à semântica da classificação de instâncias (Kent, 2000)**.** Logo, deve-se guardar a **ontologia origem** e a **ontologia destino** ao estimar similaridade e sobreposição, mantendo-so nas direções , e .

foi estrategicamente escolhida para ser implementada primeiro: considera-se A menos geral que B quando toda instância de A também pertence a B, logo temos a seguinte equação booleana:

Equação : 

* Onde é um limiar maior que 0.5 e próximo de 1;
* é a sobreposição na direção , pois inclui apenas instâncias de classificadas em , independente das instâncias de classificadas em ;
* Utiliza-se e não para podermos implementar a equivalência a seguir.

e são definidas em função de , pois e são inversos e o ocorre quando dois conceitos estão contidos um no outro:

Equação : 

Equação : 

A Equação 23 fica mais precisa conforme aumenta; um valor na casa dos *80%* funciona bem na prática. Contudo, ainda é possível melhorar esta equação, observando-se que quando um conceito *A* está contido em outro conceito *B*:

1. espera-se que  seja 1: toda instância de A teoricamente está em B. Podemos esperar que esta similaridade esteja próxima de 1, para o que estabelecemos um limiar próximo de 1:



1. espera-se que  tenda a 0: nem toda instância de B teoricamente está em A. Podemos esperar o mínimo de sobreposição ou pelo menos similaridade de , para o que estabelecemos um limiar próximo de 0:



Considerando as observações nos itens **a** e **b**, alteramos então a Equação 23 para:

Equação : 

A não depende de e ocorre quando ainda existe alguma interseção entre os conceitos, mas nenhuma relação mais forte se aplica. Na prática, assumimos um limiar próximo de 0 que diferencie sobreposição e diferença:

Equação : 

A é o que resta quando as relações anteriores falharem. Temos agora como consultar qual a relação mais apropriada para um par de conceitos e , iniciando da relação mais forte até a mais fraca, como a seguir no método ***consultaRelacao***:

### Comparação *Top-Down*

De posse do método *consultaRelacao*, o passo seguinte poderia ser comparar todos os conceitos entre si ao custo de uma complexidade quadrática. Contudo, essa abordagem não é interessante uma vez que, devido a erros oriundos da classificação das instâncias, observamos no decorrer dos experimentos algumas relações fortes sendo estabelecidas entre conceitos que, além de diferentes, estavam localizados em regiões taxonômicas também diferentes.

Na tentativa de contornar os erros que naturalmente ocorrem durante a classificação, exploramos a estrutura taxonômica das ontologias para evitar erros grosseiros de mapeamento. Basicamente, a idéia é percorrer as taxonomias de cima para baixo comparando seus conceitos, ou em outras palavras, utilizar uma **estratégia dedutiva** ***top-down*** para mapear primeiramente conceitos mais genéricos, trançando o contexto taxonômico (conjunção de ancestrais) de conceitos mais específicos para só então mapeá-los usando o método *consultaRelacao*. Esta estratégia tem caráter dedutivo porque, como toda dedução, envolve inferência partindo de princípios gerais, isto é, do universal para o particular.

Intuitivamente, como conceitos similares contêm subconceitos também similares, a comparação *top-down* usa as taxonomias como guias, evitando comparações desnecessárias para reduzir o espaço de busca. Outra vantagem é que a hierarquia taxonômica melhora a estimativa dos parâmetros de entrada (Resnik, 1990), no nosso caso similaridade e sobreposição: devido à propagação *bottom-up*, classes genéricas (outrora problemáticas por falta de instâncias diretas) passam a ter mais instâncias do que classes específicas. Logo, parâmetros estimados próximo ao topo taxonômico são mais confiáveis. Espera-se então que esta estratégia repare estimativas menos confiáveis nas classes específicas, evitando falsos mapeamentos fortes entre conceitos distintos de regiões taxonômicas de fato distintas.

A comparação *top-down* é simples, recursiva e foi implementada **basicamente** como no método que recebe dois conceitos *A* e *B*:

Diferente de , as chamadas recursivas nas linhas 5 a 11 recebem vetores. Deve-se assumir estes casos como iterações implícitas que chamam novamente , seguindo a ordenação decrescente de , o que chamamos de **alinhamento horizontal**.

A recursão é diferente dependendo da relação estimada. Ao estimar , devemos comparar apenas seus subconceitos, evitando comparar um conceito com os subconceitos do outro. Ao estimar , devemos comparar aos subconceitos de , sendo desnecessário comparar com os subconceitos de . Analogamente, o mesmo é feito ao estimar . Ao estimar , significa que pode haver uma sobreposição mais específica entre subconceitos, ou mesmo diretamente entre A ou B com subconceito um do outro. Finalmente, ao estimar , podemos encontrar alguma relação que, em classes mais específicas, passe nos testes de limiar durante a relaxação das regras de compatibilidade.

O método apresentado é básico e precisa de adaptações. Por exemplo, evitar a recursão ao estimar tendo-se que e fazer um **alinhamento vertical** ao estimar , pois quando detectamos uma equivalência no topo das taxonomias, não sabemos se a equivalência realmente ocorre no topo ou mais abaixo, pois devemos considerar a possibilidade de erro escassez e má distribuição de instâncias concentradas em poucos conceitos mais específicos. Veja:

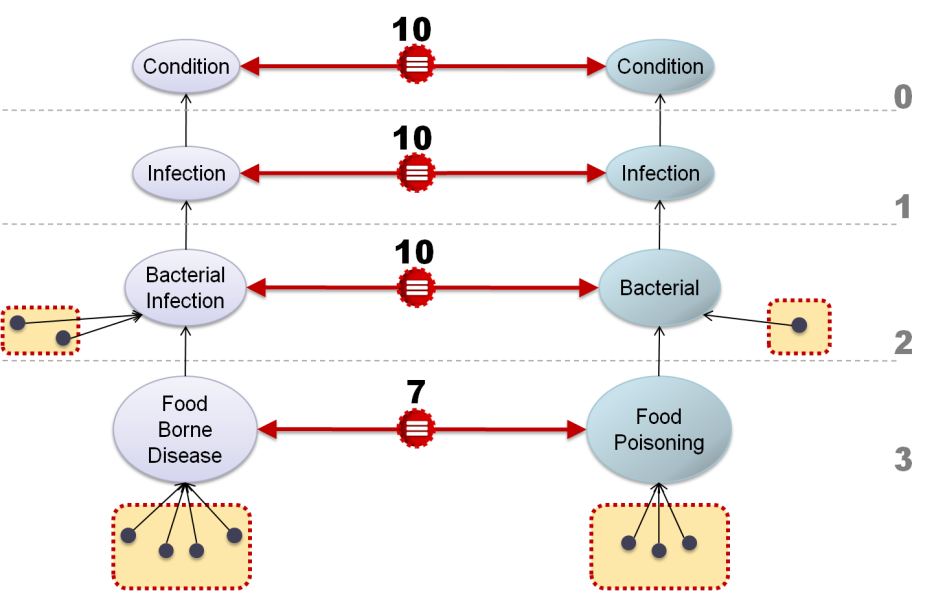


Figura ‑: Má distribuição de instâncias numa ramificação da taxonomia

No exemplo, todas as classes são equivalentes, o que nem sempre acontece. Contudo, perceba que nos níveis 0 e 1 não existem instâncias senão aquelas oriundas de ***BacterialInfection*** e ***Bacterial*** (10 instâncias no total): o ideal seria que *Infection* e *Condition* possuíssem outros subconceitos instanciados ou mesmo instâncias diretas. Portanto, apesar dos pares e serem equivalentes, a sobreposição estimada não é suficiente para concluir isso em casos como este em que faltam instâncias. Esta situação também pode ocorrer devido a uma regra ou restrição (axioma de domínio) definida em um dos conceitos, tornando-o mais geral que o outro, apesar da enumeração de instâncias em ambos sugerir que sejam equivalentes.

Portanto, na dúvida se uma equivalência está no topo ou no fundo das taxonomias, o alinhamento vertical foi implementado dando preferência para o fundo: um superconceito nunca tem menos instâncias que qualquer de seus subconceitos, portanto se a sobreposição absoluta for igual no próximo nível mais específico, a equivalência desce um nível na taxonomia, exceto se alguma outra condição se opuser a isso. Sejam então os conceitos *A*, *B* e *B’ tal que B’⊂ B*, definimos formalmente o alinhamento vertical como segue:

Deve-se então chamar e . Neste caso, utilizamos o valor de sobreposição combinada, pois estamos no contexto de uma equivalência, onde já não há noção de direção. A função deve ser uma técnica paliativa não baseada nos valores de sobreposição para determinar se *A* e *B* são equivalentes. Atualmente, o L-*Match* compara os nomes dos conceitos.

Capítulo 5

# EXPERIMENTOS

Neste capítulo, será discutida e apresentada a avaliação da qualidade dos resultados da abordagem de mapeamento semântico proposta no presente trabalho. Primeiramente serão apresentadas as ontologias experimentadas juntamente com os recursos computacionais utilizados, seguidos pela discussão do método de avaliação hierárquico e, finalizando o capítulo, os resultados de avaliação obtidos serão apresentados e comentados.

Apesar de ser considerada importante, a eficiência (velocidade) não é tratada rigorosamente neste trabalho, uma vez que ontologias normalmente têm poucos conceitos (menos que 1000) e também poucas instâncias. Como nosso objetivo é criar uma nova abordagem, precisamos primeiramente nos certificar da eficácia (qualidade) da abordagem e saber se ela é viável e promissora, para só depois pensar em otimizações de desempenho, se for o caso.

No contexto de mapeamento entre ontologias e esquemas, existem iniciativas[[7]](#footnote-8) de avaliação que tentam trazer algum consenso para o tema, porém seus padrões são limitados, mais voltados a mapeamento sintático e não fornecem *baselines* realmente úteis, seja devido a *links* quebrados, ontologias corrompidas, ontologias triviais ou, o que mais acontece, ontologias não instanciadas que definitivamente não servem para avaliar mapeadores baseados em instâncias como o L-*Match*. Na prática, muitos trabalhos não utilizam estas iniciativas de padronização, pois ainda são precárias e não atendem a demanda. Por isso estabelecemos nossa própria abordagem de avaliação e também escolhemos as ontologias a experimentar, como veremos na seção a seguir.

Para a realização dos experimentos, todos os códigos foram feitos em classes implementadas em JAVA e compiladas com J2SDK 1.6. Os experimentos foram feitos em ambiente Ubuntu Linux 10.4, numa máquina HP Intel(R) Core(TM) 2 Duo T5250 1.50GHz com 2GB RAM. Utilizamos a API JENA[[8]](#footnote-9) para ler ontologias em OWL, a *Rainbow[[9]](#footnote-10)* como serviço de classificação e o *WordNet* 2.1[[10]](#footnote-11) como serviço de thesaurus.

## ONTOLOGIAS UTILIZADAS

Bases oficiais de referência são úteis para trabalhos subseqüentes, mas quando a demanda por estas bases consiste em ontologias instanciadas, a dificuldade para encontrá-las é imensa. Na falta destas referências oficiais, descreveremos a seguir as ontologias que utilizamos.

As ontologias Ecolíngua (Brilhante, 2004) e Apes (Athanasiadis, et al., 2006) possuem domínios sobrepostos e foram utilizadas para conduzir nossos experimentos. A Ecolíngua é uma ontologia do domínio ecológico, originalmente implementada em Prolog e posteriormente em OWL, instanciada com exemplos do *PondSystem* (Grant, et al., 1997) (Ford, 1999) (Haefner, 1996). Por sua vez, a Apes é uma ontologia do domínio agrícola desenvolvida pelo Projeto SEAMLESS-IF[[11]](#footnote-12), relacionado a problemas de modelagem ambiental e agrícola. Este é o par de ontologias mais importante para o presente trabalho, pois contém ontologias cuidadosamente elaboradas, criadas para aplicações reais e desenvolvidas por grupos de trabalho distintos que não sabiam da existência um do outros, ou seja, as ontologias não foram criadas artificialmente para avaliar mapeamentos. Conseqüentemente, apesar de serem correlatas, Apes e Ecolíngua tiveram implementações bem diferentes o que, somado ao fato de possuírem muitos conceitos, as torna muito atraentes para avaliar o desempenho do L-*Match*.

Mesmo sendo instanciadas, Apes e Ecolíngua não possuem uma quantidade ideal de instâncias, criando um desafio ainda maior para serem mapeadas pelo L-*Match*. Os gráficos a seguir apresentam a distribuição de instâncias por classe concreta nestas ontologias:

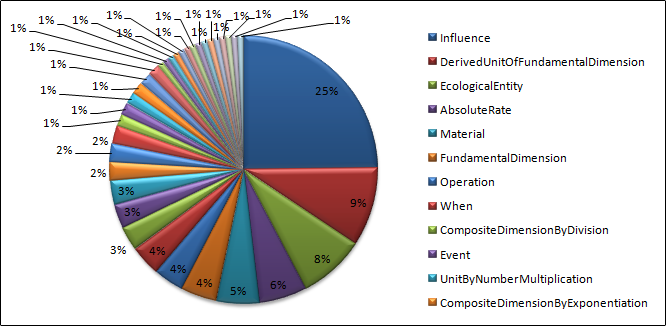


Figura ‑: Distribuição de instâncias na Ecolingua.owl

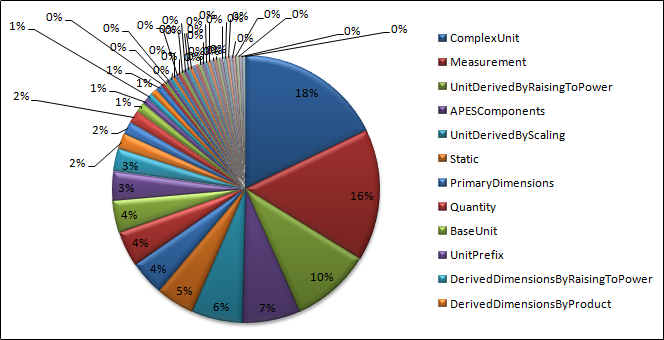


Figura ‑: Distribuição de instâncias na Apes.owl

A distribuição não é homogênea, mas não houve problema de desbalanceamento de classes durante a classificação das instâncias. Acreditamos que isso ocorra porque o conteúdo das instâncias ontológicas é bastante preciso, permitindo separar bem as classes.

Outros pares de ontologias também foram utilizados, cujas informações aparecem resumidas na tabeladas a seguir, incluindo conceitos concretos (diretamente instanciados), abstratos (instanciados indiretamente apenas) e vazios, propriedades de objeto (owl:ObjectProperty) e elementares (owl:DataTypeProperty) e dados sobre instâncias:

Tabela ‑: Informações sobre classes, instâncias e propriedades das ontologias utilizadas

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PAR | ONTOLOGIA | CONCEITOS | | | | PROPRIEDADES | | INSTÂNCIAS | | |
| Total | Concreto | Abstrato | Vazio | Objeto | Elementar | Total | Por conceito | Por conceito concreto |
| 1 | Apes | 157 | 47 | 25 | 85 | 68 | 70 | 258 | 1.64 | 5.48 |
| Ecolíngua | 65 | 35 | 25 | 5 | 32 | 6 | 136 | 2.09 | 3.88 |
| 2 | Disease1 | 9 | 4 | 3 | 2 | 3 | 1 | 8 | 0.88 | 2.0 |
| Disease2 | 7 | 4 | 3 | 0 | 4 | 0 | 12 | 1.71 | 3.0 |
| 3 | Cornell | 21 | 17 | 4 | 0 | 0 | 0 | 398 | 18.95 | 23.41 |
| Washington | 31 | 25 | 6 | 0 | 0 | 0 | 721 | 23.25 | 28.84 |
| 4 | Russia1 | 162 | 75 | 32 | 55 | 61 | 19 | 239 | 1.47 | 3.18 |
| Russia2 | 151 | 51 | 39 | 61 | 59 | 16 | 157 | 1.03 | 3.07 |

O segundo par é composto por ontologias simples e pequenas sobre doenças, por isso tem função mais ilustrativa e foram implementadas exatamente como foram encontradas em (Udrea, et al., 2007). O terceiro par e composto por uma amostra dos catálogos de cursos da Universidade de Cornell[[12]](#footnote-13) (Austrália) e da Universidade de Washington[[13]](#footnote-14) (Estados Unidos), idêntica à amostra utilizada pelo S-*Match*. Estes catálogos são freqüentemente utilizados como caso de teste em trabalhos de integração de ontologias, como S-*Match* e GLUE, o que permite fazer alguma comparação, mas não são ontologias originalmente, e sim esquemas formados por cursos, departamentos, colégios e escolas que funcionam como conceitos organizados hierarquicamente que podem ser instanciados pelas disciplinas semestrais. Por fim, o quarto par de ontologias é sobre a Rússia, é publicamente conhecido assim como seus mapeamentos, é conhecido por ser muito difícil de mapear e foi testado por vários sistemas durante a I3CON[[14]](#footnote-15), *The Information Interpretation and Integration Conference*, em 2004.

## AVALIAÇÃO HIERÁRQUICA DA QUALIDADE DOS MAPEAMENTOS COMPUTADOS

Problemas mais elaborados por envolver classes organizadas em hierarquia taxonômica inserem maior dificuldade na solução e na avaliação do problema, pois desafiam a capacidade computacional de fazer generalizações. Exemplos de problemas assim são mapeamento semântico e classificação hierárquica.

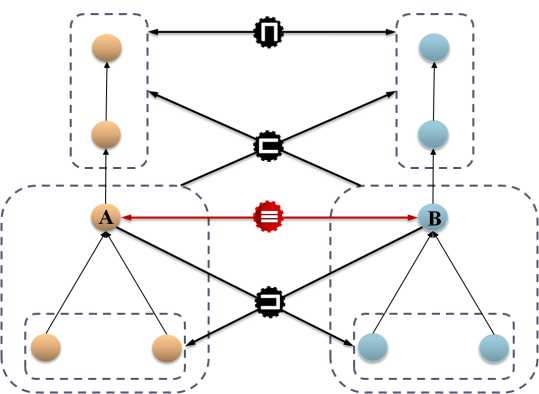
Apesar de não existir uma abordagem padrão para avaliar estes problemas, sabemos intuitivamente que devemos julgar os resultados computados comparando-os com resultados ideais dados por especialistas humanos. Porém, mesmo os especialistas podem não ter certeza sobre o mapeamento ideal quando a subjetividade é grande, o que exige certa flexibilidade ao estabelecê-los, permitindo que mais de um mapeamento seja aceito em casos particulares.

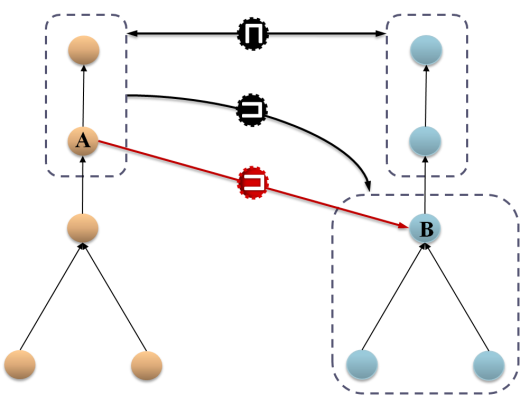
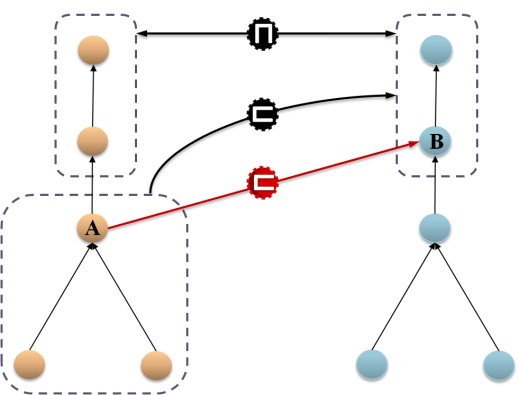
Baseamos nossa avaliação nas métricas de precisão, revocação e medida-F. Contudo, se utilizadas de maneira convencional num ambiente hierárquico, estas métricas não capturariam detalhes importantes e considerariam nossos resultados completamente errados: por exemplo, quando uma equivalência entre dois conceitos *A* e *B* é esperada, mas é erroneamente computada entre *A* e algum *Ancestral(B)*, não podemos dizer que o erro foi total, pois ainda é possível inferir que *A* é ancestral de todo *Descendente(B)* e que *A* descende de algum *Ancestral(B)*.

Para resolver este impasse, nos inspiramos na idéia de **escopo expandido** (Ding, et al., 2000) para primeiramente inferir mapeamentos implícitos para só então avaliá-los: seja *M* um mapeamento entre dois conceitos *A* e *B*, temos então que é o conjunto formado por *M* e por todo mapeamento inferível a partir de *M*. No nosso caso, a expansão ou inferência precisa ocorrer de maneira diferenciada dependendo do axioma utilizado no mapeamento, o que acontece seguindo as regras abaixo:

Tabela ‑: Regras de expansão de mapeamentos

|  |  |
| --- | --- |
|  | *A* e *B* compartilham os mesmos ancestrais e descendentes. |
|  | *A* é mais geral que todo descendente de *B*  *B* é menos geral que todo ancestral de *A*  *A* é sobreposto a todo ancestral de *B* |
|  | *B* é mais geral que todo descendente de *A*  *A* é menos geral que todo ancestral de *B*  *B* é sobreposto a todo ancestral de *A* |
|  | *A* e todo ancestral de *A* são sobrepostos a todo ancestral de *B*.  *B* e todo ancestral de *B* são sobrepostos a todo ancestral de *A*. |
|  | *A* e *B* não possuem descendentes em comum. |



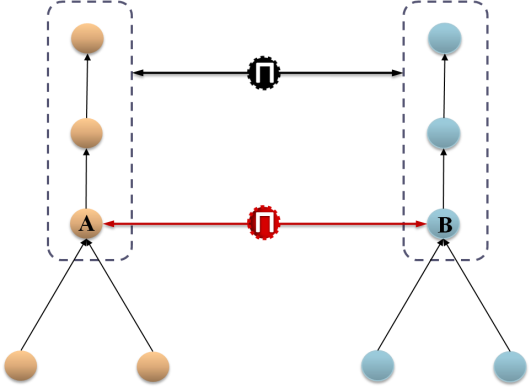
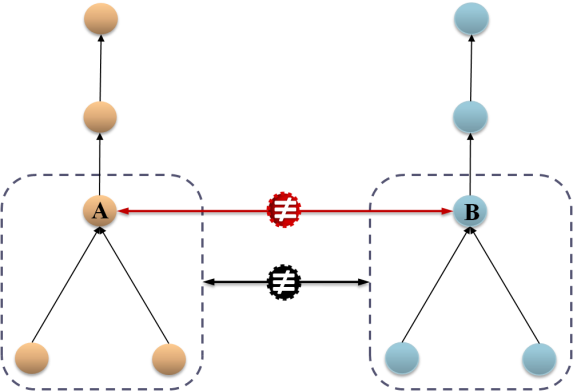
 

Figura ‑: Mapeamentos pretos são inferidos a partir dos mapeamentos vermelhos

As regras acima são implementadas numa **Tabela de Mapeamentos** que é o produto cartesiano , onde *O1* e O2 são ontologias. Para a avaliação, são criadas duas tabelas: uma com mapeamentos ideais e outra com mapeamentos computados. Particularmente, a tabela ideal é um complemento à abordagem de mapeamento semântico discutida na Seção 4.3. As posições destas tabelas são então comparadas na expectativa de que os mapeamentos atribuídos para ambas sejam iguais, indicando um acerto. Os acertos são então contabilizados para cada axioma de mapeamento, por isso cada axioma recebe um valor de precisão, revocação e medida-F.

O Módulo de Similaridade do L-*Match* também é avaliado com base na tabela ideal. Como os resultados deste módulo são valores de sobreposição e similaridade (e não axiomas de mapeamento), não há como utilizar a tabela computada. Contudo traçamos uma estratégia de avaliação: a idéia é observar a saída do Módulo de Similaridade e mensurar sua utilidade para o cálculo dos mapeamentos semânticos.

Após a propagação *bottom-up*, os conceitos de treino são ranqueados pela ordem decrescentes dos valores de similaridade e sobreposição com cada conceito de teste (veja a Figura 4‑10). Utilizamos um **limiar relativo** para “limpar” o *rank* de similaridade, isto é, remover conceitos pouco similares; este limiar mantém no *rank* apenas conceitos com similaridade maior ou igual a *x*% da similaridade observada na primeira posição do *rank*. Para os experimentos, utilizamos sempre um limiar relativo a 70%.

É fácil avaliar o *rank* de sobreposição: basta manter conceitos com sobreposição maior que zero, o que automaticamente inclui superconceitos. Mas se um conceito é mantido no *rank* de similaridade, não significa que seus superconceitos também serão e vice-versa. Portanto, a taxonomia de treino não sofre propagação *bottom-up* antes da avaliação do *rank* de similaridade, de forma que este *rank* contenha apenas conceitos concretos (diretamente instanciados). Em seguida, o *rank* é expandido e passa a incluir outros conceitos, seguindo as regras da Tabela 5‑2 em função do mapeamento ideal mais forte esperado para o conceito de teste correspondente.

As figuras seguintes exemplificam a avaliação de sobreposição e similaridade, mostrando conceitos de teste na primeira coluna, seguidos pela precisão (P), revocação (R), medida-F (F) e pelo *rank* com classes de treino; os mapeamentos idealmente esperados estão indicados ao lado dos conceitos de treino, a cardinalidade dos conceitos de teste aparecem entre parênteses e os *ranks* de similaridade incluem apenas conceitos concretos:

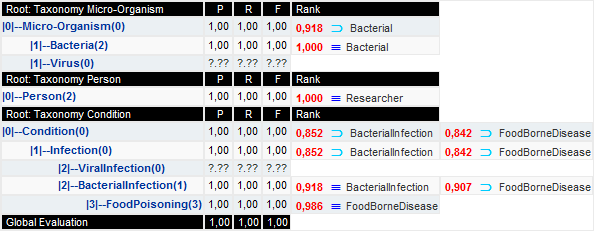


Figura ‑: Avaliação dos *ranks* de similaridade para ontologias de doenças

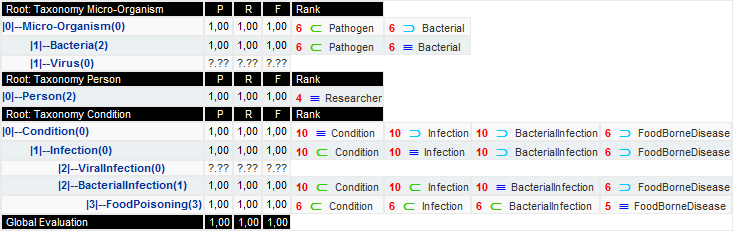


Figura ‑: Avaliação dos *ranks* de sobreposição para ontologias de doenças

Sendo assim, espera-se que apenas conceitos relevantes permaneçam após estas “limpezas” nos *ranks*. A avaliação global (*Global Evaluation*) do Módulo de Similaridade não é a macro-avaliação (média das avaliações locais), e sim a micro-avaliação calculada em função do somatório de acertos, conceitos ideais e conceitos computados em cada *rank*.

## RESULTADOS

Nesta seção, os resultados dos experimentos são apresentados e comentados dando ênfase à eficácia do método. Os seguintes limiares de relaxamento foram fixados para todos os pares de ontologias utilizados:

Tabela ‑: Limiares de Relaxamento

|  |  |
| --- | --- |
|  | 0.85 |
|  | 0.90 |
|  | 0.10 |
|  | 0.30 |

Além destes, foi fixado um limiar relativo a 70% para avaliar os valores de similaridade. Nas subseções a seguir veremos os resultados dos experimentos com Ecolingua e Apes, seguidos pelos experimentos com outras ontologias.

### Avaliação Geral do Mapeamento Ecolíngua  Apes

Iniciamos os experimentos sem realimentar o L-*Match*, pois queremos uma idéia inicial sobre o comportamento dos módulos de Similaridade e de Mapeamento e também dos algoritmos de classificação. A avaliação do Módulo de Similaridade é válida porque sua eficácia influencia diretamente a do Módulo de Mapeamento. As conclusões tiradas com Ecolingua e Apes serão então usadas para direcionar os experimentos com outros pares de ontologias.

A tabela seguinte apresenta uma avaliação geral da primeira iteração de mapeamento por algoritmo de classificação utilizado junto ao L-*Match*, a fim de compará-los. Incluímos a micro-avaliacão dos valores combinados de similaridade e sobreposição, a macro-avaliação do mapeamento e a mensuração de tempo em segundos. Em seguida é plotado um gráfico a partir da macro-avaliação do mapeamento e da mensuração de tempo:

Tabela ‑: Avaliação Geral do Mapeamento Ecolingua×Apes

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CLASSIFICADOR | *SOBREPOSIÇÃO* | | | *SIMILARIDADE* | | | *MAPEAMENTO* | | | TEMPO |
| **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| k-NN | 0.38 | 0.64 | 0.48 | 0.80 | 0.41 | 0.54 | 0.60 | 0.47 | 0.53 | 6.79s |
| Naive Bayes | **0.61** | 0.76 | **0.68** | 0.73 | 0.17 | 0.27 | 0.80 | 0.63 | 0.70 | 19.83s |
| NB-Shrinkage | 0.48 | **0.77** | 0.59 | **0.90** | **0.61** | **0.73** | **0.82** | **0.76** | **0.79** | 19.19s |
| SVM | 0.38 | 0.75 | 0.50 | 0.62 | 0.53 | 0.57 | 0.40 | 0.44 | 0.42 | 20.22s |
| Máxima Entropia | 0.42 | 0.70 | 0.53 | 0.71 | 0.57 | 0.63 | 0.68 | 0.66 | 0.67 | 20.40s |
| TF-IDF | 0.34 | 0.68 | 0.46 | 0.65 | 0.47 | 0.55 | 0.60 | 0.59 | 0.60 | **5.84s** |

Figura ‑: Gráfico da Avaliação por Classificador do Mapeamento Ecolingua×Apes

Em geral, os resultados foram bastante equilibrados, mas a vantagem qualitativa é do NB-*Shrinkage*, desde a avaliação de sobreposição e similaridade, consolidando-se na avaliação do mapeamento com valores próximos a 80%. Bons resultados também foram obtidos por *Naive Bayes* e Máxima Entropia, a avaliação dos mapeamentos ultrapassa 50% em quase todos os casos e a maior velocidade foi obtida por TF-IDF, o que não é vantagem, pois mapeamentos muito melhores foram obtidos em menos de apenas 20 segundos pelo NB-*Shrinkage*.

Observamos também que a avaliação de sobreposição não foi tão notória quanto a de similaridade, o que não prejudicou gravemente o mapeamento e nos leva a concluir que a técnica de mapeamento guiada pela taxonomia contornou muitos dos erros de classificação, com era esperado. Dada a boa avaliação da similaridade computada entre conceitos, podemos supor também que, ignorando valores de sobreposição computados entre conceitos pouco similares, podemos futuramente melhorar a precisão dos valores de sobreposição e, conseqüentemente, do mapeamento.

### Avaliação do Módulo de Similaridade para Ecolíngua  Apes

As tabelas seguintes exibem nas colunas e a avaliação da sobreposição e similaridade parciais (com noção de direção) entre conceitos, re-exibindo na coluna os valores combinados da Tabela 5‑4 (sem noção de direção). Seja então o mapeamento Ecolingua× Apes:

Tabela ‑: Avaliação da Sobreposição Ecolingua×Apes por Classificador

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CLASSIFICADOR |  | | |  | | |  | | |
| **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| k-NN | 0.66 | 0.54 | 0.59 | 0.25 | 0.32 | 0.28 | 0.38 | 0.64 | 0.48 |
| Naive Bayes | 0.69 | 0.47 | 0.56 | **0.57** | 0.51 | **0.54** | **0.61** | 0.76 | **0.68** |
| NB-Shrinkage | **0.81** | **0.68** | **0.74** | 0.43 | **0.69** | 0.53 | 0.48 | **0.77** | 0.59 |
| SVM | 0.68 | 0.61 | 0.64 | 0.38 | 0.64 | 0.48 | 0.38 | 0.75 | 0.50 |
| Máxima Entropia | 0.64 | 0.62 | 0.63 | 0.38 | 0.57 | 0.46 | 0.42 | 0.70 | 0.53 |
| TF-IDF | 0.64 | 0.55 | 0.59 | 0.30 | 0.49 | 0.37 | 0.34 | 0.68 | 0.46 |

Tabela ‑: Avaliação da Similaridade Ecolingua×Apes por Classificador

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CLASSIFICADOR |  | | |  | | |  | | |
| **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| k-NN | 0.66 | 0.59 | 0.62 | 0.35 | 0.50 | 0.41 | 0.80 | 0.41 | 0.54 |
| Naive Bayes | 0.69 | 0.47 | 0.56 | **0.61** | 0.51 | 0.55 | 0.73 | 0.17 | 0.27 |
| NB-Shrinkage | **0.81** | 0.61 | **0.70** | 0.50 | **0.71** | **0.58** | **0.90** | **0.61** | **0.73** |
| SVM | 0.58 | **0.66** | 0.62 | 0.37 | 0.67 | 0.47 | 0.62 | 0.53 | 0.57 |
| Máxima Entropia | 0.72 | 0.58 | 0.64 | 0.46 | 0.63 | 0.54 | 0.71 | 0.57 | 0.63 |
| TF-IDF | 0.65 | 0.62 | 0.63 | 0.34 | 0.58 | 0.43 | 0.65 | 0.47 | 0.55 |

Os resultados nas colunas são muito melhores que aqueles da coluna , tendo o NB-*Shrinkage* alcançado até 81% de precisão e cerca de 70% de medida-F. Isso significa que houve maior dificuldade de treinar na Ecolíngua no momento de classificar as instâncias da Apes, talvez porque a Ecolíngua possui menos instâncias: apenas 3.88 instâncias por conceito concreto, contra 5.48 na Apes. De qualquer forma, esta diferença é natural, pois também foi observada em outros pares de ontologias.

Os melhores valores se concentram entre *Naive Bayes* e NB-*Shrinkage*. Apesar de não ter sido unânime, a vantagem é do NB-*Shrinkage*, pois é sempre superior ou próximo ao *Naive Bayes*. Esta superioridade é mais clara na avaliação de similaridade do que na avaliação de sobreposição. Isso ocorre porque o cálculo de similaridade entre conceitos é mais abrangente, pois não se resume à primeira posição do *rank* retornado pelos classificadores para cada instância: Por outro lado, o cálculo da sobreposição espera que o conceito correto de toda instância venha na primeira posição deste *rank* (utilizamos um limiar *Top-k*=1), o que é provável mas não garantido, pois às vezes pode ocorrer em outras posições iniciais. Contudo, a investigação de valores maiores para *Top-k* não trouxeram melhoria.

Isso mostra também que, ao comparar conceitos em função de instâncias, é melhor utilizar a similaridade computada pelos algoritmos de classificação, os quais capturam nuances maiores, do que utilizar a similaridade de *Jaccard* em função dos valores absolutos de sobreposição.

### Avaliação do Módulo de Mapeamento para Ecolíngua  Apes

Esta subseção detalha a avaliação mais importante que é a dos mapeamentos. Apresentaremos a avaliação por relação (axioma ponte) e realimentaremos o L-*Match* com sua própria saída para mostrar que a abordagem iterativa é capaz de melhorar a qualidade do mapeamento quando associada ao NB-*Shrinkage*, classificador utilizado para guiar os experimentos devido ao seu bom desempenho, como vimos anteriormente.

Basicamente, a avaliação do mapeamento será tabelada seguindo o modelo abaixo. Os tamanhos dos conjuntos Ideal, Computado e Acerto sao exibidos por relação, bem como precisão, revocação e medida-F, que chamaremos de medidas parciais. A linha de totais sumariza os tamanhos dos conjuntos e apresenta as medidas de macro-precisão, macro-revocação e macro-medida-F, que são as médias das parciais:

Tabela ‑: Modelo de Avaliação de Mapeamento

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | REVOCAÇÃO | MEDIDA-F |
| relação 1 | i1 | c1 | a1 | P1 | R1 | F1 |
| ... | ... | ... | ... | ... | ... | ... |
| relação N | iN | cN | aN | PN | RN | FN |
| Total |  |  |  | PMACRO | RMACRO | FMACRO |

Utilizamos a macro-avaliação porque, neste caso, a micro-avaliação iguala precisão, revocação e medida-F; logo, obeteríamos sempre o *accuracy*. Isso acontece porque os conjuntos Ideal e Computado tem o mesmo tamanho, isto é, .

Fizemos apenas três iterações de mapeamento, pois observamos um fenômeno interessante: após um pequeno número de iterações, o mapeamento tende a convergir para um resultado que não se altera mais ou então se altera muito pouco. No caso do mapeamento Ecolingua×Apes utilizando NB-*Shrinkage*, os resultados se tornam constantes a partir da terceira iteração.

A tabela seguinte detalha a avaliação do mapeamento da 1° iteração. Observe que as macro-medidas são as mesmas apresentadas na Tabela 5‑4 para o NB-*Shrinkage*. Em seguida, as medidas parciais são utilizadas para plotar o gráfico da Figura 5‑7, veja:

Tabela ‑: Avaliação da 1° Iteração de Mapeamento Ecolingua×Apes com NB-*Shrinkage*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | REVOCAÇÃO | MEDIDA-F |
| Diferença | 9756 | 9801 | 9736 | 0.99 | 1.00 | 1.00 |
| Sobreposição | 38 | 46 | 18 | 0.39 | 0.47 | 0.43 |
| Menos Geral | 157 | 154 | 147 | 0.95 | 0.94 | 0.95 |
| Mais Geral | 240 | 193 | 180 | 0.93 | 0.75 | 0.83 |
| Equivalência | 14 | 11 | 9 | 0.82 | 0.64 | 0.72 |
| Total | 10205 | 10205 | 10090 | 0.82 | 0.76 | 0.79 |

Figura ‑: Gráfico da 1° Iteração de Mapeamento Ecolingua×Apes com NB-*Shrinkage*

Não houve problema para identificar diferenças, o que demonstra a capacidade da estratégia *Top-Down* de não associar conceitos extremamente diferentes. Contudo, muitos erros recaem sobre a sobreposição, que teve a pior avaliação: sua baixa precisão indica que relações mais fortes identificadas incorretamente acabaram sendo computadas, no mínimo, como sobreposição, por erro da abordagem ou por falta de instâncias. Porém, a sobreposição é a relação que menos importa, já que praticamente não tem utilidade.

Como desejávamos, depois da diferença, os melhores valores foram obtidos para as relações mais importantes (equivalência, mais geral e menos geral), especialmente em termos de precisão e medida-F. Isso significa que tivemos sucesso ao descobrir relações de herança e também desempenhamos bem ao descobrir equivalências com alta precisão (82%) e revocação mais baixa a 64%, que pelo menos está acima de 50%.

De maneira geral, relações de herança também foram descobertas com boa precisão nos mapeamentos semânticos de (Bouquet, et al., 2003) e (Magnini, et al., 2004), trabalhos que avaliaram o Ctx-*Match*. Contudo, a medida-F do Ctx-*Match* mostra grande desequilíbrio, devido à revocações muito mais baixas que precisões. Além disso, o Ctx-*Match* teve a pior avaliação para a equivalência, onde o desequilíbrio é gritante, hora com baixa precisão (33%) e revocação (4%), hora com alta precisão (78%) e baixa revocação (13%). Por sua vez, o S-*Match* não apresentou a avaliação por relação e, apesar de aparentemente interessante, suas avaliações foram feitas sobre esquemas pequenos, sem grandes diferenças estruturais e terminológicas, portanto fáceis de mapear.

Agora apresentaremos a avaliação da estratégia iterativa do L-*Match*: certamente obtivemos bons valores para a equivalência com o L-*Match*, mas isso precisa melhorar já que esta relação é a mais importante. Para isso realimentamos o L-*Match* com seus próprios mapeamentos e avaliamos a segunda iteração de mapeamentos:

Tabela ‑: Avaliação da 2° Iteração de Mapeamento Ecolingua×Apes com NB-*Shrinkage*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | REVOCAÇÃO | MEDIDA-F |
| Diferença | 9756 | 9764 | 9717 | 1.00 | 1.00 | 1.00 |
| Sobreposição | 38 | 21 | 16 | 0.76 | 0.42 | 0.54 |
| Menos Geral | 157 | 145 | 137 | 0.94 | 0.87 | 0.91 |
| Mais Geral | 240 | 263 | 223 | 0.85 | 0.93 | 0.89 |
| Equivalência | 14 | 12 | 11 | 0.92 | 0.79 | 0.85 |
| Total | 10205 | 10205 | 10104 | 0.89 | 0.80 | 0.84 |

O total de acertos foi maior que o da iteração anterior (+14), com ganhos ocorrendo em todas as relações, com pequenas perdas na precisão de mais geral (-8%) e na revocação de menos geral (-7%) e da sobreposição (-5%). Os ganhos para equivalência são evidentes (+10% de precisão e +15% de revocação), bem como o ganho na precisão da sobreposição (+37%). Há também queda no número de falsas diferenças. Estes ganhos elevam a avaliação total para a casa dos 80%, com precisão de quase 90%. Estes resultados melhoraram ainda mais após a terceira iteração, quando então convergiram, veja:

Tabela ‑: Avaliação da 3° Iteração de Mapeamento Ecolingua×Apes com NB-*Shrinkage*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | REVOCAÇÃO | MEDIDA-F |
| Diferença | 9756 | 9762 | 9717 | 1.00 | 1.00 | 1.00 |
| Sobreposição | 38 | 19 | 16 | 0.84 | 0.42 | 0.56 |
| Menos Geral | 157 | 143 | 137 | 0.96 | 0.87 | 0.91 |
| Mais Geral | 240 | 266 | 227 | 0.85 | 0.95 | 0.90 |
| Equivalência | 14 | 15 | 14 | 0.93 | 1.00 | 0.97 |
| Total | 10205 | 10205 | 10111 | 0.92 | 0.85 | 0.88 |

Figura ‑: Gráfico da 3° Iteração de Mapeamento Ecolingua×Apes com NB-Shrinkage

O aumento de acertos (+5) nesta iteração não é o grande responsável pelos ganhos, mas sim a distribuição dos acertos: ganhos ocorreram em todas as relações, sem nenhuma perda de valor; exceto pela revocação e medida-F da sobreposição, todas as outras medidas estão acima de 83%, algumas ultrapassando 90%. O resultado mais satisfatório foi o da equivalência, que alcançou revocação máxima (100%) e altíssima precisão (93%), o que surpreende porque Ecolíngua e Apes possuem mapeamentos bastante subjetivos.

Portanto, concluímos que a abordagem iterativa foi muito importante porque propiciou ganhos muito satisfatórios. Contudo, estes resultados foram obtidos utilizando apenas o NB-*Shrinkage*, por isso experimentamos fazer o mesmo com outros algoritmos de classificação. As duas tabelas seguintes são equivalente e apresentam resultados por classificador em cada iteração, mas a primeira apresenta a macro-avaliação e a segunda apresenta os ganhos/perdas de avaliação em relação à primeira iteração:

Tabela ‑: Avaliação por Classificador do Mapeamento Iterativo Ecolingua×Apes

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFICADOR** | *1**Iteração* | | | *2**Iteração* | | | *3**Iteração* | | |
| **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| k-NN | 0.60 | 0.47 | 0.53 | 0.57 | 0.37 | 0.45 | 0.53 | 0.40 | 0.45 |
| Naive Bayes | 0.80 | 0.63 | 0.70 | 0.75 | 0.46 | 0.57 | 0.62 | 0.42 | 0.50 |
| NB-Shrinkage | **0.82** | **0.76** | **0.79** | **0.89** | **0.80** | **0.84** | **0.92** | **0.85** | **0.88** |
| SVM | 0.40 | 0.44 | 0.42 | 0.43 | 0.43 | 0.43 | 0.39 | 0.45 | 0.42 |
| Máxima Entropia | 0.68 | 0.66 | 0.67 | 0.70 | 0.72 | 0.71 | 0.57 | 0.47 | 0.51 |
| TF-IDF | 0.60 | 0.59 | 0.60 | 0.52 | 0.63 | 0.57 | 0.47 | 0.59 | 0.52 |

Tabela ‑: Ganho por Classificador do Mapeamento Iterativo Ecolingua×Apes

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFICADOR** | *1**Iteração* | | | *2**Iteração* | | | *3**Iteração* | | |
| **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| k-NN | 0.60 | 0.47 | 0.53 | -3% | -10% | -8% | -7% | -7% | -8% |
| Naive Bayes | 0.80 | 0.63 | 0.70 | -5% | -17% | -13% | -18% | -21% | -20% |
| NB-Shrinkage | **0.82** | **0.76** | **0.79** | **+7%** | **+4%** | **+5%** | **+20%** | **+9%** | **+9%** |
| SVM | 0.40 | 0.44 | 0.42 | +3% | -1% | +1% | -1% | +1% | 0% |
| Máxima Entropia | 0.68 | 0.66 | 0.67 | +2% | +5% | +4% | -11% | -19% | -16% |
| TF-IDF | 0.60 | 0.59 | 0.60 | -8% | **+4%** | -3% | -13% | 0% | -8% |

Infelizmente, ganhos não são garantidos quando utilizamos outro classificador individual senão o NB-*Shrinkage*: o Máxima Entropia até obteve ganhos na segunda iteração, mas logo teve grandes perdas na terceira; o SVM manteve-se quase constante e os outros algoritmos obtiveram boas perdas, principalmente o *Naive Bayes*, que havia se saído tão bem na primeira iteração. Felizmente identificamos um algoritmo de classificação útil, o NB-*Shrinkage*, o que nos permite concluir que além da abordagem iterativa de mapeamento, o algoritmo de classificação utilizado também é importante.

Repetimos estes experimentos com algumas combinações (ensembles) do NB-*Shrinkage* com outros classificadores que, na primeira iteração, deram mostra que poderiam funcionar:

Tabela ‑: Avaliação por Ensemble do Mapeamento Iterativo Ecolingua×Apes

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFICADOR** | *1**Iteração* | | | *2**Iteração* | | | *3**Iteração* | | |
| **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| NB-Shrinkage | 0.82 | 0.76 | **0.79** | **0.89** | 0.80 | **0.84** | **0.92** | **0.85** | **0.88** |
| NB-Shrinkage+k-NN | 0.83 | 0.75 | **0.79** | 0.86 | 0.74 | 0.80 | 0.84 | 0.73 | 0.78 |
| NB-Shrinkage+NaiveBayes | **0.85** | 0.74 | **0.79** | 0.86 | 0.80 | 0.83 | 0.88 | 0.78 | 0.83 |
| NB-Shrinkage+SVM | 0.71 | **0.78** | 0.75 | 0.79 | **0.85** | 0.82 | 0.71 | 0.76 | 0.73 |
| NB-Shrinkage+MaxEnt | 0.75 | 0.76 | 0.76 | 0.79 | 0.78 | 0.78 | 0.72 | 0.70 | 0.71 |
| NB-Shrinkage+TF-IDF | 0.73 | 0.73 | 0.73 | 0.76 | 0.80 | 0.78 | 0.79 | 0.84 | 0.82 |
| k-NN+Naive Bayes | 0.82 | 0.68 | 0.74 | 0.83 | 0.60 | 0.69 | 0.81 | 0.46 | 0.58 |

Tabela ‑: Ganho por Ensemble do Mapeamento Iterativo Ecolingua×Apes

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFICADOR** | *1**Iteração* | | | *2**Iteração* | | | *3**Iteração* | | |
| **P** | **R** | **F** | **P** | **R** | **F** | **P** | **R** | **F** |
| NB-Shrinkage | 0.82 | 0.76 | **0.79** | +7% | +4% | +5% | **+20%** | +9% | **+9%** |
| NB-Shrinkage+k-NN | 0.83 | 0.75 | **0.79** | +3% | -1% | +1% | +1% | -2% | -1% |
| NB-Shrinkage+NaiveBayes | **0.85** | 0.74 | **0.79** | +1% | +6% | +4% | +3% | +4% | +4% |
| NB-Shrinkage+SVM | 0.71 | **0.78** | 0.75 | **+8%** | **+7%** | **+7%** | 0% | -2% | -2% |
| NB-Shrinkage+MaxEnt | 0.75 | 0.76 | 0.76 | +4% | +2% | +2% | -2% | -6% | -5% |
| NB-Shrinkage+TF-IDF | 0.73 | 0.73 | 0.73 | +3% | **+7%** | +5% | +6% | **+11%** | **+9%** |
| k-NN+Naive Bayes | 0.82 | 0.68 | 0.74 | +1% | -8% | -5% | -1% | -22% | -16% |

Nenhuma das combinações encontrou resultados melhores do que o NB-*Shrinkage* utilizado individualmente, porém as quedas de valor durante as iterações foram menores, exceto na última combinação, a qual não envolveu o NB-*Shrinkage*. Apesar dos ganhos na segunda iteração, as combinações com Máxima Entropia e SVM logo entraram em declínio; por outro lado as combinações com *Naive Bayes* e principalmente com TF-IDF também obtiveram ganhos consecultivos.

### Avaliação do Módulo de Mapeamento para Outros Pares de Ontologias

Experimentaremos outras ontologias no L-*Match* baseado apenas no NB-*Shrinkage*, pois concluímos que é o mais adequado. Iniciamos mapeando com duas iterações as ontologias sobre doenças, que são fáceis de mapear e se beneficiaram bastante do alinhamento vertical:

Tabela ‑: Avaliação da 1° Iteração de Mapeamento Disease 1 e 2 com NB-Shrinkage

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | | REVOCAÇÃO | | MEDIDA-F | |
| Diferença | 39 | 38 | 38 | 1.00 | 0.97 | | 0.99 | |
| Sobreposição | 0 | 0 | 0 | 1.00 | 1.00 | | 1.00 | |
| Menos Geral | 11 | 14 | 11 | 0.79 | 1.00 | | 0.88 | |
| Mais Geral | 7 | 6 | 6 | 1.00 | 0.86 | | 0.92 | |
| Equivalência | 6 | 5 | 5 | 1.00 | 0.83 | | 0.91 | |
| Total | 63 | 63 | 60 | 0.96 | 0.93 | | 0.94 | |

Tabela ‑: Avaliação da 2° Iteração de Mapeamento Disease 1 e 2 com NB-*Shrinkage*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | REVOCAÇÃO | MEDIDA-F |
| Diferença | 39 | 40 | 39 | 0.98 | 1.00 | 0.99 |
| Sobreposição | 0 | 1 | 0 | 0.00 | 1.00 | 0.00 |
| Menos Geral | 11 | 9 | 9 | 1.00 | 0.82 | 0.90 |
| Mais Geral | 7 | 7 | 7 | 1.00 | 1.00 | 1.00 |
| Equivalência | 6 | 6 | 6 | 1.00 | 1.00 | 1.00 |
| Total | 63 | 63 | 61 | 0.80 | 0.96 | 0.87 |

Figura ‑: Gráfico da 2° Iteração de Mapeamento Disease1×Disease2 com NB-*Shrinkage*

A primeira iteração teve resultados muito bons, mas outros ainda melhores foram alcançados novamente na segunda iteração: equivalência e mais geral alcançaram avaliação máxima e a medida-F de menos geral aumentou. As quedas na macro-avaliação são enganosas: um único falso-positivo zerou a precisão da sobreposição, influenciando exageradamente a macro-precisão, pois a média é uma função estatística tendenciosa a altos e baixos valores. De qualquer forma, a sobreposição é a relação menos importante.

O próximo par de ontologias a mapear corresponde à amostra retirada dos catálogos de cursos das universidades de Cornell e de Washington. Os resultados foram interessantes uma vez que todas as avaliações atingiram valor máximo logo na primeira iteração, veja:

Tabela ‑: Avaliação da 1° Iteração de Mapeamento Cornell×Washington com NB-*Shrinkage*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | REVOCAÇÃO | MEDIDA-F |
| Diferença | 506 | 506 | 506 | 1.00 | 1.00 | 1.00 |
| Sobreposição | 1 | 1 | 1 | 1.00 | 1.00 | 1.00 |
| Menos Geral | 49 | 49 | 49 | 1.00 | 1.00 | 1.00 |
| Mais Geral | 74 | 74 | 74 | 1.00 | 1.00 | 1.00 |
| Equivalência | 21 | 21 | 21 | 1.00 | 1.00 | 1.00 |
| Total | 651 | 651 | 651 | 1.00 | 1.00 | 1.00 |

Contudo, estes resultados não são tão notórios, pois estes catálogos são parecidos. De qualquer forma, este mesmo experimento foi realizado pelo S-*Match* que obteve o mesmo resultado máximo, portanto o L-*Match* não deixou a desejar. O GLUE também realizou diferentes experimentos com estes catálogos e, apesar de ter alcançado alguns resultados acima de 90%, não obteve valores máximos como o S-*Match* e o L-*Match*.

Por último, mapeamos as ontologias sobre a Rússia. Neste caso, a abordagem iterativa não ajudou, pois os resultados convergiram logo na primeira iteração:

Tabela ‑: Avaliação da 1° Iteração de Mapeamento Rússia 1 e 2 com NB-*Shrinkage*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | REVOCAÇÃO | MEDIDA-F |
| Diferença | 22908 | 22719 | 22246 | 0.98 | 0.97 | 0.98 |
| Sobreposição | 101 | 108 | 21 | 0.19 | 0.21 | 0.20 |
| Menos Geral | 952 | 827 | 532 | 0.64 | 0.56 | 0.60 |
| Mais Geral | 460 | 769 | 389 | 0.51 | 0.85 | 0.63 |
| Equivalência | 41 | 39 | 23 | 0.59 | 0.56 | 0.58 |
| Total | 24462 | 24462 | 23211 | 0.58 | 0.63 | 0.60 |

Figura ‑: Gráfico da 1° Iteração de Mapeamento Rússia 1 e 2 com NB-*Shrinkage*

Apesar de não serem notórios como os anteriores, estes resultados foram muito bons, dado que superam em cerca de 20% os experimentos realizados durante a I3CON em 2004[[15]](#footnote-16), que aplicou diferentes mapeadores sobre este mesmo par de ontologias, considerando os mesmos mapeamentos ideais, mas que malmente atingiram 40% na avaliação comparativa de media-f, contra 60% alcançado pelo L-*Match.* Além disso, estas ontologias sobre Rússia são conhecidas por serem realmente difíceis de mapear, pois apresentam muitas diferenças, principalmente estruturais.

Combinando NB-*Shrinkage* com *Naive Bayes* no L-*Match*, conseguimos resultados de mapeamento um pouco melhores para Rússia 1 e 2, mas a diferença foi pouca, o que não compensa já que praticamente o mesmo pode ser obtido executando apenas um classificador:

Tabela ‑: Avaliação da 1° Iteração de Mapeamento Rússia 1 e 2 com NB-*Shrinkage*+*NaiveBayes*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| RELAÇÃO | IDEAL | COMPUTADO | ACERTO | PRECISÃO | REVOCAÇÃO | MEDIDA-F |
| Diferença | 22908 | 22719 | 22246 | 0.98 | 0.97 | 0.97 |
| Sobreposição | 101 | 108 | 21 | 0.23 | 0.26 | 0.24 |
| Menos Geral | 952 | 827 | 532 | 0.74 | 0.56 | 0.64 |
| Mais Geral | 460 | 769 | 389 | 0.45 | 0.86 | 0.59 |
| Equivalência | 41 | 39 | 23 | 0.60 | 0.61 | 0.60 |
| Total | 24462 | 24462 | 23211 | 0.60 | 0.65 | 0.62 |

Capítulo 6

# CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho pesquisou a utilização de Aprendizado de Máquina para Integração de Informação em bases de dados heterogêneas, no caso, mapeamento semântico entre conceitos de ontologias distintas. Foram investigadas diferentes técnicas de classificação supervisionada adaptadas para criar uma nova abordagem de mapeamento semântico, cuja qualidade foi avaliada por meio de uma implementação preliminar chamada L-*Match*, a qual basicamente explora o conteúdo textual gerado a partir das propriedades das instâncias declaradas pelos autores das ontologias.

Dos bons resultados obtidos durante a avaliação dos experimentos, concluímos que a abordagem como um todo funciona muito bem, é bastante promissora e competitiva, que explorar instâncias no mapeamento é bom, que o NB-*Shrinkage* é o algoritmo de classificação que melhor funciona dentro da nossa abordagem, talvez pela intimidade entre *shrinkage* e hierarquias taxonômicas, e que a abordagem de mapeamento iterativo reutilizando mapeamentos computados pode incrementar muito a qualidade dos resultados. Contudo, infelizmente não foi possível fazer avaliação comparativa com o S-*Match*, que é o único mapeador semântico além do L-*Match*, pois seus criadores não o disponibilizam para terceiros.

## TRABALHOS FUTUROS

Observamos que o L-*Match* é apenas a peça inicial de uma solução mais completa, contudo é peça chave, dado o sucesso na avaliação dos experimentos. Certamente, está é a primeira etapa, pois muitas melhorias futuras são aplicáveis, com as quais esperamos ganhos de precisão e velocidade. A seguir, exemplificaremos trabalhos futuros espalhados pelos três módulos do L-*Match*.

Melhorias no Módulo de Extração são sempre bem vindas, pois é neste momento que os dados são extraídos de suas fontes, transformados (renderizados em texto, limpos, indexados, normalizados, etc.) e utilizados para alimentar os módulos seguintes, os quais necessitam receber informação útil, precisa e confiável: quanto menos ruído melhor. Trabalhos futuros vão desde coisas simples como expansão de siglas, passando pela experimentação de outras técnicas de desambiguação e de identificação de sinônimos sem utilizar *WordNet*, até a extensão da portabilidade do L-*Match* para trabalhar com diferentes formatos de ontologias. Seria interessante também gerar *tokens* combinando palavras que se modificam mutuamente (Chakravarthy, 1995): na Ecolíngua, por exemplo, *Temperature* e *OfTemperature* são conceitos cuja diferença advém de um ser substantivo e outro ser locução adjetiva, mas ao excluir a *stopword of* perdemos esta informação de diferença, levando a falsos positivos que comprometem a qualidade do texto.

Para o Módulo de Similaridade podemos adaptar o NB-*Shrinkage* para fazer classificação hierárquica e experimentar APIs de classificação mais atualizadas que a *Rainbow*, como Weka[[16]](#footnote-17) e LibSVM[[17]](#footnote-18). Podemos também aumentar o número de instâncias utilizando anotação automática ou, por outro lado, diminuir este número eliminando instâncias ruins contendo ruídos que atrapalham a classificação (*noisy exemplar pruning*), mantendo apenas instâncias com classificação confiável. Por fim, será interessante podar os valores de sobreposição entre conceitos pouco similares.

Nó Módulo de Mapeamento é ainda necessário observar melhor o comportamento da estratégia comparativa *TopDown* afim de otimizá-la, estender o L-*Match* para mapear propriedade (atributos), refinar o alinhamento vertical através de melhores técnicas para comparação de rótulos, refinar a abordagem iterativa controlando o conhecimento que é reutilizado (ou não) e, por fim, associar o L-*Match* com outras técnicas de mapeamento como, por exemplo, submeter os mapeamentos do L-*Match* ao raciocínio proposicional implementado no S-*Match*, que é o real diferencial deste mapeador. Podemos também tentar delimitar conceitos através de suas propriedades para então compará-los, o que é tão intuitivo quanto utilizar instâncias.

Capítulo 7

# REFERÊNCIAS

**AnHai, Doan, et al. 2004.** Ontology Matching: A Machine Learning Approach. *In Handbook on Ontologies.* s.l. : Springer-Verlag, 2004, pp. 385–404.

**Aumueller, David, et al. 2005.** Schema and Ontology Matching with COMA++. *In ACM SIGMOD International Conference on Management of Data.* June 2005, pp. 906–908.

**Baader, Franz, Horrocks, Ian and Sattler, Ulrike. 2002.** Description Logics as Ontology Languages for the Semantic Web. *In the International Workshop on Ontologies.* 2002.

**Bouquet, Paolo, et al. 2003.** A SAT-Based Algorithm for Context Matching. *In Proceedings of CONTEXT2003.* 2003.

**Bouquet, Paolo, et al. 2003.** C-OWL: Contextualizing Ontologies. *In Second International Semantic Web Conference (ISWC-2003).* 2003, Vol. 2870, pp. 164-179.

**Bouquet, Paolo, Serafini, Luciano and Zanobini, Stefano. 2003.** Semantic Coordination: a New Approach and an Application. *In the 2nd International Semantic Web Conference (ISWC'03).* october 2003.

**Braga, Alessandra Lopes, et al. 2006.** Comparação entre as Classificações Híbrida eSupervisionada no Mapeamento do Uso do Solo Usando Imagens de Alta Resolução. *In Congresso Brasileiro de Cadastro Técnico Multifinalitário.* october 15-19, 2006.

**Brilhante, Virgínia. 2004.** An Ontology for Quantities in Ecology. *In 17th Brazilian Symposium on Artificial Intelligence.* 2004, pp. 144-153.

**Bron, Coen and Kerbosch, Joep. 1973.** Algorithm 457: Finding All Cliques of an Undirected Graph. *Communications of the ACM.* 1973, Vol. 16.

**Buitelaar, Paul and Sacaleanu, Bogdan. 2001.** Ranking and Selecting Synsets by Domain Relevance. *In Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL 2001 Workshop.* June 2001.

**Buitelaar, Paul, Cimiano, Philipp and Magnini, Bernardo. 2005.** *Ontology Learning from Text: Methods, Evaluation and Applications.* s.l. : IOS Press, 2005. Vol. 123 Frontiers in Artificial Intelligence.

**Burges, Christopher J.C. 1998.** A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery.* 1998, pp. 121 - 167.

**Calado, Pável, et al. 2003.** Combining Link-Based and Content-Based Methods for Web Document Classification. *In Conference on Information and Knowledge Management (CIKM03).* 2003, pp. 394-401.

**Chakrabarti, Soumen. 2002.** *Mining the Web: Discovering Knowledge from Hypertext Data.* Bombay, India : Morgan-Kaufmann Publishers, 2002. ISBN 1-55860-754-4.

**Chandrasekaran, B., Josephson, John R. and Benjamins, V. Richard. 1999.** What Are Ontologies, and Why Do We Need Them? *In IEEE Intelligent Systems.* January 1999, Vol. 14.

**Charniak, Eugene. 1991.** Bayesians Networks without Tears. *IA Magazine.* 1991, Vol. 12, pp. 50-63.

**Cimiano, Philipp, et al. 2004.** Learning Taxonomic Relations from Heterogeneous Evidence. *in Proceedings of the Ontology Learning and Population Workshop, 16th European Conference on Artificial Intelligence (ECAI04).* 2004.

**Do, Hong-Hai and Rahm, Erhard. 2002.** COMA - A System for Flexible Combination of Schema Matching Approaches. *In Proceedings of VLDB.* 2002, pp. 610-621.

**Doan, AnHai, et al. 2002.** Learning to Map Between Ontologies on the Semantic Web. *In 11th International World Wide Web Conference.* May 2002.

**Du, Ding-Zhu and Pardalos, Panos M. 2007.** *Handbook of Combinatorial Optimization.* s.l. : Springer-Verlag, 2007. ISBN:0792359240.

**Euzenat, Jérôme and Valtchev, Petko. 2004.** An Algorithm and an Implementation of Semantic Matching. *In Proceedings of the First European Semantic Web Symposium.* 2004.

**—. 2003.** An Integrative Proximity Measure for Ontology Alignment. *In Semantic Integration Workshop, Second International Semantic Web Conference (ISWC-03).* 2003.

**—. 2004.** Similarity-based Ontology Alignment in OWL-Lite. *In European Conference on Artificial Intelligence (ECAI-04).* 2004, pp. 333–337.

**Euzenat, Jerome, et al. 2004.** Ontology Alignment with OLA. *In Proceedings of the International Semantic Web Conference Workshop on Evaluation of Ontology-Based Tools (EON'04).* 2004, pp. 59–68.

**Fellbaum, Christiane. 1998.** WordNet - An Electronic Lexical Database. *The MIT Press.* 1998.

**Fürst, Frédéric and Trichet, Francky. 2005.** Axiom-Based Ontology Matching. *In The 3rd International Conference on Knowledge Capture.* October 2005, pp. 195-196.

**—. 2005.** Axiom-based ontology matching: a method and a experiment. *Relatório Técnico N° 05-02, Laboratório de Informática de Nantes-Atrantique (LINA).* Março 2005.

**Ghidini, Chiara and Giunchiglia, Fausto. 2001.** Local Models Semantics, or Contextual Reasoning = Locality + Compatibility. *Artificial Intelligence.* April 2001, Vol. 127, 2, pp. 221-259.

**Giunchiglia, Fausto, Shvaiko, Pavel and Yatskevich, Mikalai. 2005.** S-Match: an Algorithm and an Implementation of Semantic Matching. *In Semantic Interoperability and Integration.* 2005.

**Gruber, Thomas R. 1995.** Toward Principles for the Design of Ontologies used for Knowledge Sharing. *In International Journal of Human-Computer Studies.* 1995, Vol. 43, pp. 907-928.

**Guarino, Nicola and Giaretta, Pierdaniele. 1995.** Ontologies and Knowledge Bases Towards a Terminological Clarification. *In Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases .* 1995, pp. 25-32.

**Haase, Peter and Motik, Boris. 2005.** A Mapping System for the Integration of OWL-DL Ontologies. *Proceedings of the First International Workshop on Interoperability of Heterogeneous information Systems.* 2005, pp. 9-16.

**Huang, Wen-Lin, et al. 2008.** ProLoc-GO: Utilizing Informative Gene Ontology Terms for Sequence-Based Prediction of Protein Subcellular Localization. *BMC Bioinformatics.* 2008.

**I., Bomze, et al. 1999.** The Maximum Clique Problem. *Handbook of Combinatorial Optimization.* 1999.

**Ichise, Ryutaro, Takeda, Hiedeaki and Honiden, Shinichi. 2003.** Integrating Multiple Internet Directories by Instance-based Learning. *in Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI'03).* 2003, pp. 22–28.

**—. 2001.** Rule Induction for Concept Hierarchy Alignment. *In Proceedings of the Workshop on Ontology Learning at IJCAI.* 2001.

**Jensen, Finn V. 2001.** *Bayesian Networks and Decision Graphs.* s.l. : Springer-Verlag, 2001. ISBN 0387952594.

**Joachims, Thorsten. 1997.** A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. *Proceedings of the 14th International Conference on Machine Learning.* 1997.

**Juan, Alfons and Ney, Hermann. 2002.** Reversing and Smoothing the Multinomial Naive Bayes Text Classifer. *In Proceedings of the 2nd International Workshop on Pattern Recognition in Information Systems (PRIS'02).* 2002, pp. 200-212.

**Kalfoglou, Yannis and Schorlemmer, Marco. 2003.** If-Map: an Ontology Mapping Method. *Journal on Data Semantics.* October 2003, pp. 98–127.

**—. 2003.** Ontology Mapping: the State of the Art. *In The Knowledge Engineering Review.* January 2003, Vol. 18, pp. 1 - 31.

**Kent, Robert E. 2000.** The Information Flow Foundation for Conceptual Knowledge Organization. *In the 6th International Conference of the International Society for Knowledge Organization (ISKO).* July 2000, pp. 10-13.

**—. 2000.** The Information Flow Foundation for Conceptual Knowledge Organization. *In Proceedings of the 6th International Conference of the International Society for Knowledge Organization (ISKO'00).* 2000.

**Klinger, Stefan and Austin, Jim. 2005.** Chemical Similarity Searching Using a Neural Graph Matcher. *In Proceedings of European Symposium on Artificial Neural Networks (ESANN'05).* April 2005.

**Korhonen, Anna and Briscoe, Ted. 2004.** Extended Lexical-Semantic Classification of English Verbs. *In Workshop on Computational Lexical Semantics.* May 2004, pp. 38–45.

**Kumar, Mukesh and Miller, Douglas A. 2006.** A Non- Parametric Classification Strategy for Remotely Sensed Images using both Spectral and Textural Information. *In Proceedings of the 24th IASTED International Multi-Conference Signal Processing, Pattern Recognition and Applications.* February 15-17, 2006, pp. 81-89.

**Lacher, Martin S. and Groh, Georg. 2001.** Facilitating the Exchange of Explicit Knowledge Through Ontology Mappings. *In Proceedings of the First International FLAIRS Conference.* May 2001.

**Lamma, Evelina and Mello, Paola. 1999.** *AI\*IA 99: Advances in Artificial Intelligence: 6th Congress of the Italian Association for Artificial Intelligence.* Bologna, Italy : Springer-Verlag, 1999. ISBN:3540673504.

**Madhavan, Jayant, Bernstein, Philip A. and Rahm, Erhard. 2001.** Generic Schema Matching with Cupid. *In Proceedings of VLDB.* 2001, pp. 49-58.

**Magnini, Bernardo, Serani, Luciano and Speranza, Manuela. 2003.** Making Explicit the Semantics Hidden in Schema Models. *In Proceedings of the Workshop on HLT form Semantic Web and Web Services at ISWC'03.* 2003.

**Magnini, Bernardo, Speranza, Manuela and Girardi, Christian. 2004.** A Semantic-Based Approach to Interoperability of Classification Hierarchies: Evaluation of Linguistic Techniques. *In Proceedings of COLING-2004.* August 23 - 27, 2004.

**McCallum, Andrew, et al. 1998.** Improving Text Classification by Shrinkage in a Hierarchy of Classes. *In Proceedings of the 15th International Conference on Machine Learning (ICML'98).* 1998, pp. 359-367.

**McGuinness, Deborah L., et al. 2000.** An Environment for Merging and Testing Large Ontologies. *In Proceedings of the 7th International Conference on Principles of Knowledge Representation and Reasoning.* 2000, pp. 483–493.

**Melnik, Sergey, Rahm, Erhard and Bernstein, Philip A. 2003.** Rondo: A Programming Platform for Generic Model Management. *In Proceedings of SIGMOD.* 2003, pp. 193–204.

**Miller, George A. 1995.** Wordnet: A Lexical Database for English. *In Communications of the ACM.* 1995, pp. 39-41.

**Minsky, Marvin. 1975.** A Framework for Representing Knowledge. *Reprinted in The Psychology of Computer Vision.* 1975, pp. 221–280.

**Nigamy, Kamal, Lafferty, John and McCallum, Andrew. 1999.** Using Maximum Entropy for Text Classification. *In IJCAI-99 Workshop on Machine Learning for Information.* 1999, pp. 61-67.

**Noy, Natalia and Musen, Mark. 2001.** Anchor-PROMPT: Using Non-Local Context for Semantic Matching. *Workshop on Ontologies and Information Sharing at the Seventeenth International JointConference on Artificial Intelligence (IJCAI'2001).* 2001.

**Pelillo, Marcello, Siddiqi, Kaleem and Zucker, Steven W. 1998.** Matching Hierarchical Structures Using Association Graphs. *In the European Conference on Computer Vision (ECCV'98).* 1998, pp. 3-16.

**Porter, Martin. 1997.** An Algorithm for Suffix Stripping. *Reprinted in Readings in Information Retrieval.* San Francisco : Morgan Kaufmann Publishers Inc., 1997, pp. 313-316.

**Qu, Yuzhong, Hu, Wei and Cheng, Gong. 2006.** Constructing Virtual Documents for Ontology Matching. *In Proceedings of the 15th International Conference on World Wide Web (WWW '06).* May 23 - 26, 2006.

**Rahm, Erhard and Bernstein, Philip A. 2001.** A Survey of Approaches to Automatic Schema Matching. *In VLDB Journal: Very Large Data Bases.* December 2001, Vol. 10, pp. 334–350.

**Resnik, Philip. 1990.** Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research.* 1990, pp. 95–130.

**Ribeiro, Regiane Maria Paes, Soares, Vicente Paulo and Vieira, Carlos Antônio Oliveira. 2005.** Avaliação de Métodos de Classificação de Imagens IKONOS para o Mapeamento da Cobertura Terrestre. *In Anais XII Simpósio Brasileiro de Sencoriamento Remoto.* april 16-21, 2005, pp. 4277-4283.

**Rish, Irina. 2001.** An Empirical Study of the Naive Bayes Classifier. *International Joint Conferences on Artificial Intelligence, Workshop on Empirical Methods in Artificial Intelligence.* 2001.

**Shakhnarovich, Gregory, Darrell, Trevor and Indyk, Piotr, [ed.]. 2006.** *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice.* s.l. : The MIT Press, 2006. ISBN 0-262-19547-X.

**Shannon, Claude. 1948.** A Mathematical Theory of Communication. *In Bell System Technical Journal.* July and October 1948, Vol. 27, pp. 379-423 and 623-656.

**Shiba, Marcelo Hiroshi, et al. 2005.** Classificação de Imagens de Sensoriamento Remoto pela Aprendizagem por Árvore de Decisão: uma Avaliação de Desempenho. *In Anais XII Simpósio Brasileiro de Sensoriamento Remoto.* April 16-21, 2005, pp. 4319-4326.

**Shvaiko, Pavel and Euzenat, Jerome. 2005.** Tutorial on Schema and Ontology Matching. *In the 2nd European Semantic Web Conference.* 2005.

**Snow, Rion, Jurafsky, Daniel and Ng, Andrew Y. 2006.** Semantic Taxonomy Induction from Heterogenous Evidence. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL.* 2006, pp. 801 - 808 .

**Sorokine, Alexandre, Bittner, Thomas and Renschler, Chris. 2005.** Ontological Investigation of Ecosystem Hierarchies and Formal Theory for Multiscale Ecosystem Classifications. *Geoinformatica.* 2005, Vol. 10, pp. 313-335.

**Sowa, John F. 2000.** *Knowledge Representation - Logical, Philosophical and Computational Foundations, 1st Edition.* s.l. : Brooks Cole Publishing Co., Pacific Grove, CA, 2000. 0534949657.

**Stuckenschmidt, Heiner, et al. 2004.** Using C-OWL for the Alignment and Merging of Medical Ontologies. *In Proceedings of KR-MED workshop at KR.* 2004.

**Stumme, Gerd and Maedche, Alexander. 2001.** FCA-MERGE: Bottom-Up Merging of Ontologies. *17th International Joint Conference on Artificial Intelligence (IJCAI'2001).* 2001, pp. 225–234.

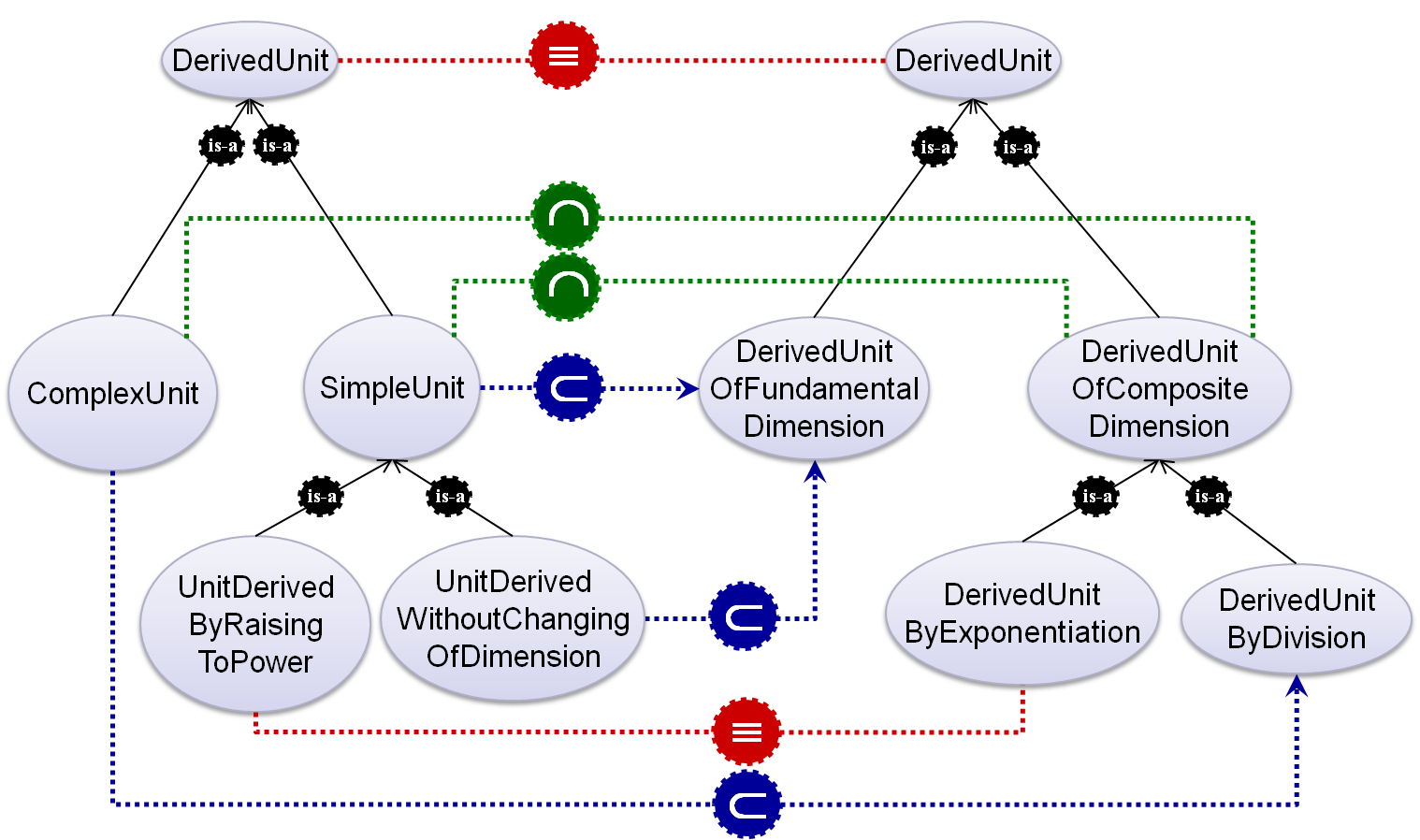
**Udrea, Octavian and Getoor, Lise. 2007.** Combining Statistical and Logical Inference for Ontology Alignment. *In Workshop on Semantic Web for Collaborative Knowledge Acquisition at the 20th International Joint Conference on Artificial Intelligence.* 2007.

**Udrea, Octavian, Getoor, Lise and Miller, Renée J. 2007.** Leveraging Data and Structure in Ontology Integration. *In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data.* 2007, pp. 449 - 460.

**Vale, Rodrigo F., et al. 2001.** Recuperação de Informação em Coleções Médicas Utilizando Categorização Automática de Documentos. *In Proceedings of the XVI Simpósio Brasileiro de Banco de Dados.* October 2001, pp. 243-258.

**Venant, Fabienne. 2006.** A Geometric Approach to Meaning Computation: Automatic Disambiguation of French Adjectives. *International Joint Conference.* October 2006.

//// RASCUNHO



 É importante que a semântica não seja ferida por mapeamentos errôneos, o que em princípio poderia ser garantido se especialistas humanos criassem os mapeamentos. Entretanto, manualmente este seria um trabalho braçal, árduo e propenso a erros de lógica, portanto soluções automáticas e precisas são inevitáveis.

An exciting and potentially far-reaching development in computer science is the invention and application of methods of machine learning. These enable a computer program to automatically analyse a large body of data and decide what information is most relevant. This crystallised information can then be used to automatically make predictions or to help people make decisions faster and more accurately.

The semantics introduced in De\_nition 4 can be viewed as an instance of the

compatibility relation between contexts as de\_ned in Local Models Semantics [8,

5]. Indeed, suppose we take a set of documents D as the domain of interpreta-

tion of the local models of two contexts c1 and c2, and each concept as a unary

predicate. If we see the documents associated to a concept as the interpretation

of a predicate in a local model, then the relation we discover between concepts of

di\_erent contexts can be viewed as a compatibility constraint between the local

models of the two concepts. For example, if the algorithm returns an equivalence

between the concepts k1 and k2 in the contexts c1 and c2, then it can be inter-

preted as the following constraint: if a local model of c1 associates a document d

to k1, then any compatible model of c2 must associate d to k2 (and vice versa);

analogously for the other relations.

De\_nition 4. A mapping function M from Hs to Ht is extensionally correct

with respect to two hierarchical classi\_cations \_s and \_t of the same set of doc-

uments D in Hs and Ht, respectively, if the following conditions hold for any

ks 2 Ks and kt 2 Kt:

*the directionality of information flow* (Bouquet, et al., 2003)

(Lacher, et al., 2001) também utilizou the *Bow* toolkit para mapear conceitos de ontologias.

(Resnik, 1990) shows how concept hierarchies

can also be used for the resolution of syntactic and semantic

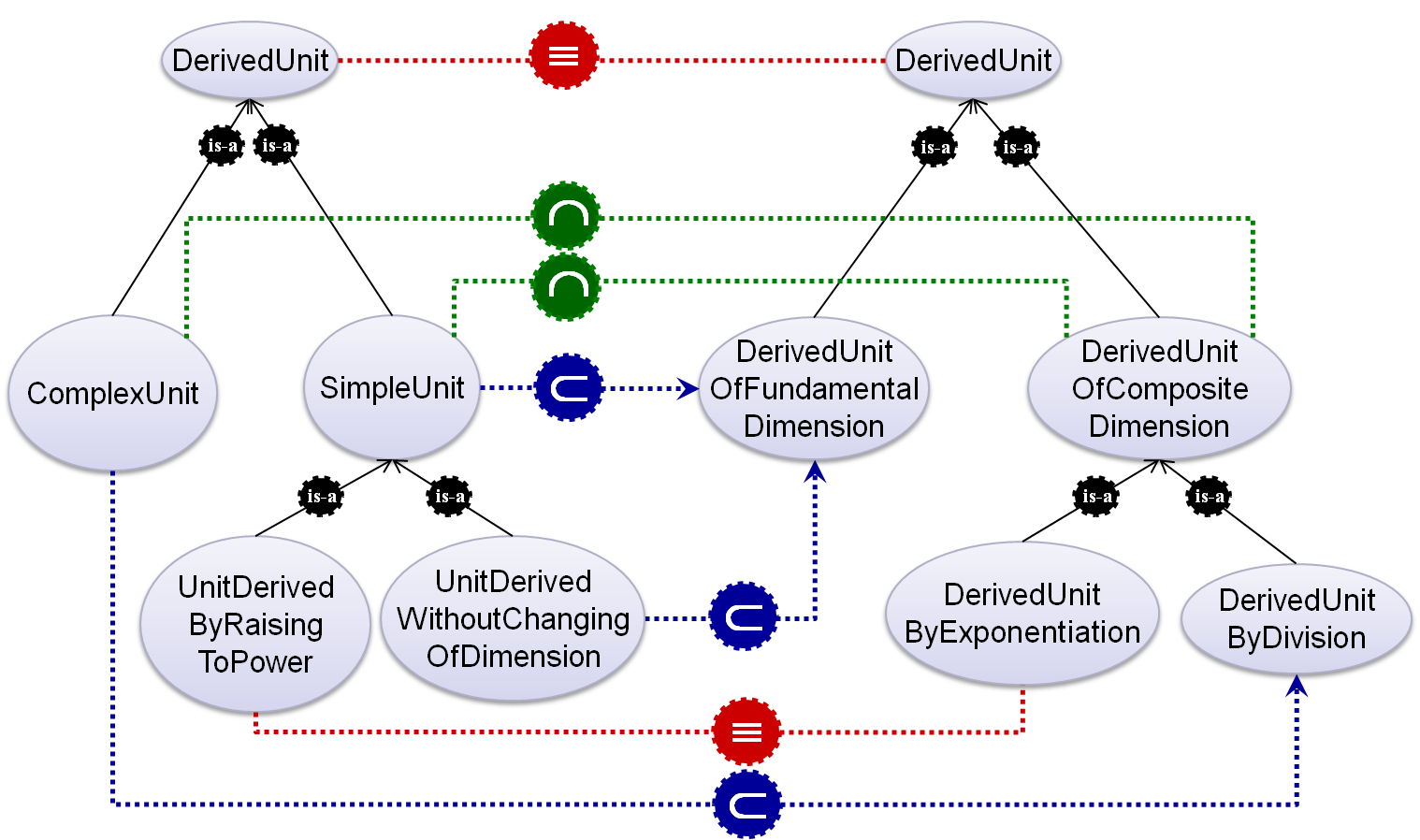
ambiguities.

Veja um exemplo simples de mapeamento semântico dado por (Giunchiglia, et al., 2005):

Basicamente, ontologias são formadas por conceitos que tipicamente estão organizados numa hierarquia taxonômica e para cada conceito são atribuídas algumas instâncias

teoricamente que as instâncias dos conceitos ontológicos são uma boa evidência para relacionar conceitos que receberam interpretações dadas por comunidades diferentes

## 



Ontologias são compostas primariamente por conceitos freqüentemente numa hierarquia taxonômica e que por vezes recebem algumas instâncias: para nossa abordagem, que se completa nesta seção, tanto a quantidade de instanciações quanto de conceitos instanciados é importante, uma vez que a eficácia da classificação e do método que apresentaremos a seguir é influenciada, dentre outras coisas, pela quantidade de instâncias. Felizmente, ao contrário do que se temia, a quantidade reduzida de instâncias nas ontologias não foi problema para o mapeamento devido ao bom desempenho dos classificadores.

Por exemplo, nas ontologias Ecolingua e Apes, como o conceito *Ecolingua.EcologicalData.Quantity* aparece num contexto específico (dados ecológicos), é teoricamente menos geral que *Apes.SeamCore.Quantity* (dados de mensuração), contudo ambos os conceitos têm a mesma utilidade nas ontologias e, portanto, podem também ser considerados equivalentes sem problema.

1. http://www.w3.org/ [↑](#footnote-ref-2)
2. http://protege.stanford.edu/ [↑](#footnote-ref-3)
3. http://www.cyc.com/ [↑](#footnote-ref-4)
4. http://www.w3.org/TR/owl-features [↑](#footnote-ref-5)
5. http://tartarus.org/~martin/PorterStemmer/ [↑](#footnote-ref-6)
6. http://dico.isc.cnrs.fr/en/index.html [↑](#footnote-ref-7)
7. http://www.ontologymatching.org/evaluation.html [↑](#footnote-ref-8)
8. http://jena.sourceforge.net/index.html [↑](#footnote-ref-9)
9. http://www.cs.cmu.edu/~mccallum/bow/rainbow/ [↑](#footnote-ref-10)
10. http://wordnet.princeton.edu/ [↑](#footnote-ref-11)
11. http://www.seamless-ip.org/ [↑](#footnote-ref-12)
12. http://www.cuinfo.cornell.edu/Academic/Courses/ [↑](#footnote-ref-13)
13. http://www.washington.edu/students/crscat/ [↑](#footnote-ref-14)
14. http://www.atl.external.lmco.com/projects/ontology/i3con.html [↑](#footnote-ref-15)
15. http://www.atl.external.lmco.com/projects/ontology/papers/I3CON-Results.pdf [↑](#footnote-ref-16)
16. http://www.cs.waikato.ac.nz/ml/weka/ [↑](#footnote-ref-17)
17. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ [↑](#footnote-ref-18)