UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Alinhamento Interativo de Ontologias: Uma Abordagem Baseada em Query-by-Committee

Vinicius Lopes

Orientadora
Fernanda Araujo Baião Amorim
Coorientadora
Kate Cerqueira Revoredo

Rio de Janeiro, RJ – Brasil
Abril de 2015

Alinhamento Interativo de Ontologias: Uma Abordagem Baseada em Query-by-Committee

Vinicius Lopes

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Fernanda Araujo Baião Amorim, D. Sc. – UNIRIO

Kate Cerqueira Revoredo, D. Sc. – UNIRIO

Bernadette Farias Lóscio, D. Sc. - UFPE

Sean Wolfgand Matsui Siqueira, D. Sc. - UNIRIO

Rio de Janeiro, RJ – Brasil

Abril de 2015

Lopes, Vinicius.
L864 Alinhamento interativo de ontologias: uma abordagem baseada em *Query-by-Committee* / Vinicius Lopes, 2015.
112 f. ; 30 cm

Orientadora: Fernanda Araujo Baião Amorim.
Coorientadora: Kate Cerqueira Revoredo.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2015.

1. Alinhamentos de ontologias. 2. Aprendizado do computador. 3. Active learning. 4. Interação homem-máquina.. I. Amorim, Fernanda Araujo Baião. II. Revoredo, Kate Cerqueira. III. Universidade Federal do Estado do Rio de Janeiro. Centro de Ciências Exatas e Tecnológicas. Curso de Mestrado em Informática. IV. Título.

CDD - 006.31

Agradecimentos

Parece que foi ontem que tudo começou. Ainda lembro-me da Fernanda nos perguntando sobre o motivo de estarmos ali, de termos escolhido cursar um mestrado. Algum tempo e muito estudo depois, eis aqui o produto desta grande jornada.

Agradeço em primeiro lugar a DEUS, pela força em todos os momentos e por me proporcionar esta oportunidade. Agradeço a minha mãe Jorgina, pelas palavras de incentivo e por compreender minhas ausências em determinados momentos. Agradeço a minha noiva e agora esposa Carla, uma companheira nesta jornada. Seu apoio nos momentos difíceis e a tranquilidade que me transmitia nos momentos de impaciência foram vitais para a conclusão deste trabalho. Agradeço as minhas orientadoras Fernanda e Kate, pelas inúmeras discussões sobre o trabalho, pela presença constante e pela motivação nos momentos em que as coisas não estavam dando certo (e foram muitos). For fim, agradeço a todos os professores da UNIRIO, com quem tive a oportunidade de conviver durante este período, aos amigos e familiares que tanto me apoiaram.

LOPES, Vinicius. **Alinhamento Interativo de Ontologias: Uma Abordagem Baseada em Query-by-Committee.** UNIRIO, 2015. 112 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Ontologias vêm recebendo grande destaque como ferramenta para a especificação de conceitos em diversos domínios. No entanto, o número crescente de ontologias heterogêneas, que descrevem o mesmo domínio, tem se apresentado como um desafio para a interoperabilidade de dados e aplicações. A área de Alinhamento de Ontologias apresenta uma possível solução para o problema da heterogeneidade, centrada na identificação de possíveis correspondências entre pares de entidades e na geração de um alinhamento entre as ontologias. Mesmo diante de todo avanço alcançado nos últimos anos, avaliações realizadas anualmente pela OAEI (*Ontology Alignment Evaluation Initiative*) demonstram que ainda existe um grande espaço de melhoria possível para as abordagens de alinhamento de ontologias. Diante do desafio da melhoria da qualidade dos alinhamentos, alguns estudos tem sido realizados na direção de abordagens interativas, que buscam tornar o usuário participante do processo de alinhamento. Dentre estas abordagens, estão aquelas que procuram explorar o conhecimento do usuário à respeito do domínio, solicitando a ele *feedback* sobre pares de entidades. Contudo, estudos vêm sendo realizados com o objetivo de aperfeiçoar este processo de solicitação de *feedback*, principalmente no que se refere a identificação de pares de entidades relevantes e propagação do efeito para outros pares de entidades, permitindo a redução do número de iterações necessárias e a maximização do efeito da participação do usuário. Neste trabalho, é apresentada uma abordagem interativa para alinhamento de ontologias baseada na estratégia de seleção de instâncias informativas, da área de estudo de *Active Learning*, conhecida como query-by-committee. O objetivo desta abordagem é explorar de forma efetiva o conhecimento do usuário à respeito de um determinado domínio de forma que esta participação possa contribuir para a melhoria da qualidade do alinhamento. A abordagem aqui apresentada é avaliada através de um conjunto de experimentos, utilizando as ontologias do *conference data set* da OAEI.

Palavras-chave: Alinhamento de Ontologias, Aprendizado de Máquina, Active Learning, Envolvimento do Usuário.

ABSTRACT

Ontologies have been receiving great prominence as a tool for specifying concepts in many different domains. However, the increasing number of heterogeneous ontologies, which describe the same domain, has emerged as a challenge for the data and applications interoperability. The Ontology Matching area presents a possible solution to the problem of heterogeneity, focused on the identification of possible correspondences between pairs of entities and generating an alignment between ontologies. Even before all progress achieved in recent years, evaluation performed annually by OAEI (Ontology Alignment Evaluation Initiative) show that there are still a large potential improvement space for Ontology Matching approaches. Faced with the challenge of improving the quality of alignments, studies have been performed in the direction of interactive approaches, which aim user involvement at the matching process. Among these approaches are those that looking to explore the user's knowledge about the domain, requesting feedback on pairs of entities. However, studies have been conducted with the aim of improving feedback request, especially as regards the identification of relevant entities pairs and propagation of the effect to other pairs of entities, allowing reducing the number of iterations required and maximizing user participation. This paper presents an interactive approach to Ontology Matching based on informative instances selection strategy, known as query-by-committee. The goal of this approach is to exploit effectively the user's knowledge to about a particular domain so that this participation can contribute to improving the alignment quality.

Keywords: Ontology Matching, Machine Learning, Active Learning, User Involvement.

Índice

Índice de Figuras

Índice de Tabelas

1 Introdução

Este capítulo apresenta uma visão geral desta pesquisa, incluindo um panorama da área de estudo de Alinhamento de Ontologias, o problema de pesquisa endereçado, a solução proposta, os objetivos que se pretende alcançar, bem como a metodologia científica aplicada.

1.1 Motivação e Caracterização do Problema

A Web Semântica é uma extensão da Web em que os dados disponibilizados são semanticamente definidos [1]. Esta característica permite que aplicações acessem dados de fontes diversificadas, processem estes dados e compartilhem seus resultados com outras aplicações.

Ontologias viabilizam a operação da Web Semântica, pois fornecem a conceitualização explícita e formal necessária aos dados, funcionando como um protocolo para a comunicação e o compartilhamento dos resultados entre aplicações.

O crescimento da quantidade de dados publicados na Web Semântica foi acompanhado de um incremento no número de ontologias definidas para especificar estes dados. Com o objetivo de estimular o reuso destas ontologias, foram criadas bibliotecas que armazenam coleções de ontologias. Estas bibliotecas permitem que usuários possam identificar, reusar e publicar ontologias que descrevam seu domínio de forma apropriada [2].

Em cenários de múltiplas ontologias representando conceitos de um mesmo domínio, é comum que exista sobreposição de conceitos entre ontologias. Em situações como esta, para estabelecer a interoperabilidade dos dados e o seu compartilhamento entre aplicações é necessário lidar com o desafio da heterogeneidade existente entre os conceitos que compõem as ontologias. A heterogeneidade pode se apresentar de

diversas formas, podendo ocorrer desde simples variações nos nomes dos conceitos (terminológica) até diferenças em seus significados, granularidade ou perspectiva (semântica).

Para lidar com o problema da heterogeneidade entre ontologias, foi criada a área de pesquisa de Alinhamento de Ontologias (*Ontology Matching*) [3]. A atividade de alinhar ontologias tem como objetivo principal identificar correspondências entre entidades das ontologias e gerar um alinhamento entre elas. Portanto, um alinhamento é formado por um conjunto de correspondências e uma correspondência é representada por um par de entidades e o tipo de relação existente entre elas, que pode ser de equivalência, generalização ou disjunção.

O processo de alinhamento de ontologias vem sendo utilizado para diversas aplicações, dentre elas a integração de dados [3]. Neste cenário, um conjunto de fontes de informação, nas quais os dados são potencialmente armazenados em diferentes formatos (*SQL, XML, RDF*), é acessado pelo usuário a partir de uma interface de consulta global. Cada uma das fontes de informação é representada por uma ontologia local *LO* e a interface de consulta por uma ontologia global *CO*. A comunicação entre cada ontologia local e a ontologia global é realizada através de um mediador (*mediator*) gerado a partir de um alinhamento entre a ontologia local e a ontologia global. A Figura 1.1 mostra o esquema de integração de dados descrito em Euzenat *et al.* [3].

Outras aplicações para a área de Alinhamento de Ontologias são destacadas ainda em [3], como: Composição de serviços web, compartilhamento de informação ponto a ponto (*P2P*) e engenharia de ontologias.

Com o avanço da pesquisa em alinhamento de ontologias, diversas abordagens para alinhamento foram propostas e para organizar a avaliação destas, foi criada a *Ontology Alignment Evaluation Initiative – OAEI*. A OAEI realiza campanhas anuais com o objetivo de avaliar a evolução dos sistemas de alinhamento de ontologias.

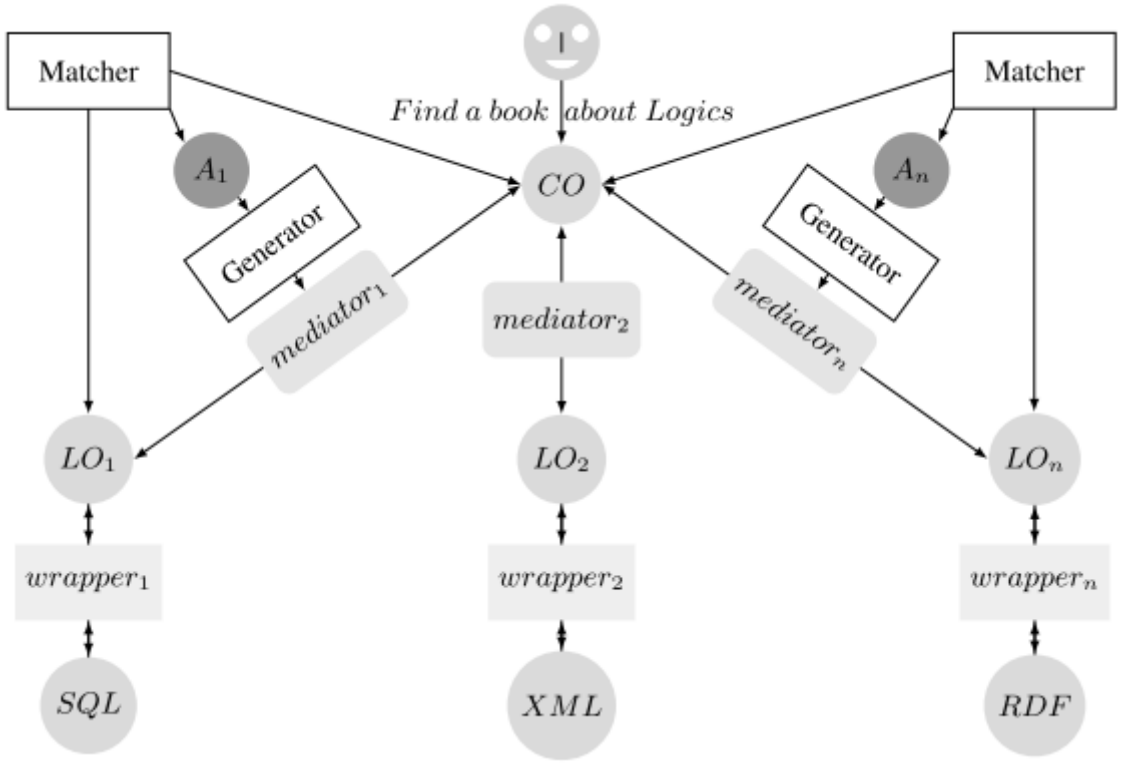

Figura 1.1- Cenário geral de integração de dados aplicando alinhamento de ontologias [4]

As campanhas realizadas são organizadas em trilhas e *data sets* que representam diferentes cenários, que buscam, por exemplo, avaliar a capacidade das soluções em alinhar ontologias expressivas, multilinguais e instâncias, automaticamente ou de forma interativa. A Figura 1.2 apresenta a evolução da medida-F (considerada a principal medida para avaliar a qualidade dos alinhamentos [5]) ao longo das campanhas anuais realizadas, considerando os diversos *data sets* disponibilizados e avaliados.

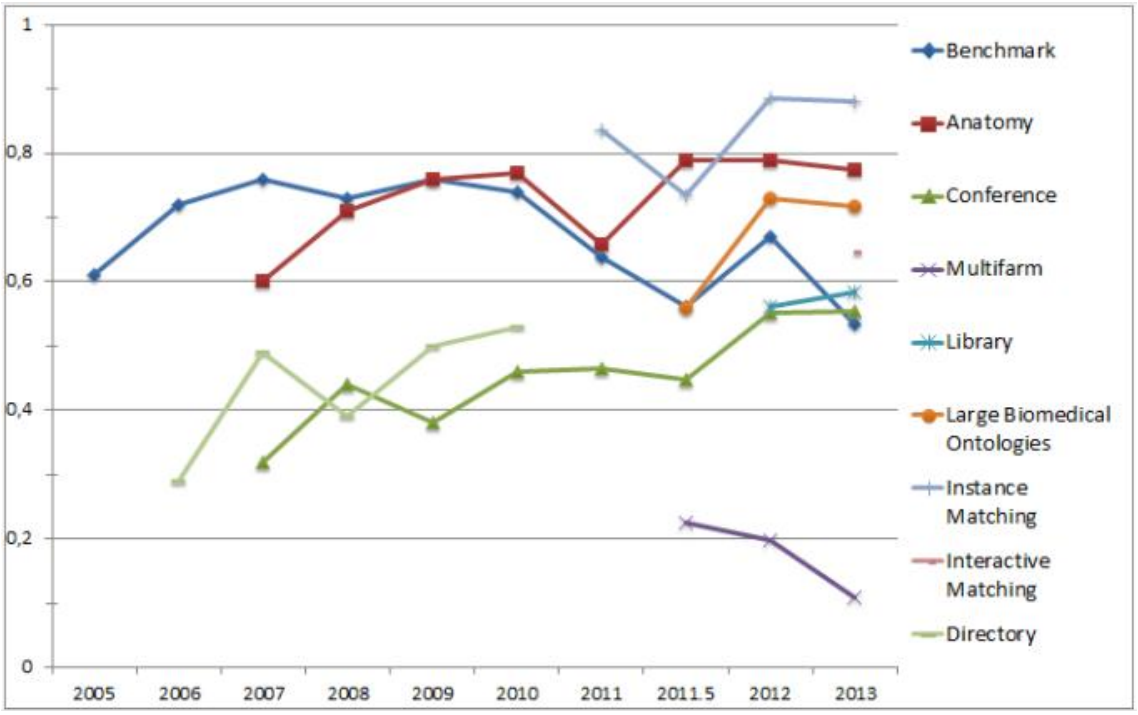

Figura 1.2 - Evolução da Medida-F média para os cenários da OAEI

Analisando o gráfico da Figura 1.2, é possível observar que, mesmo com a evolução das abordagens propostas, ainda existe um espaço para melhoria da qualidade dos alinhamentos. Neste caso, a principal questão a ser respondida é: Como aumentar a qualidade dos alinhamentos gerados?

A maioria das soluções propostas adota uma abordagem automática no processo de alinhamento de ontologias e segundo Ehrig [5], é muito difícil criar uma solução totalmente automática que apresente resultados satisfatórios. Desta forma, estudos recentes vêm sendo realizados com o objetivo de envolver o usuário no processo de forma a melhorar a qualidade dos alinhamentos, dando origem a abordagens interativas. De fato, o envolvimento do usuário no processo de alinhamento de ontologias é um dos sete desafios identificados por Euzenat *et al.* em seu levantamento do estado da arte desta área de pesquisa [6].

Este envolvimento deve ser pensado de forma que a interação com o usuário seja realizada de forma natural (em que se exija dele apenas conhecimento sobre o domínio, ao invés de conhecimento técnico sobre os algoritmos e sobre o processo de alinhamento), e dentre as possíveis estratégias neste sentido, está aquela que atribui ao usuário o papel de fornecer *feedback* sobre pares de entidades.

Um problema central em abordagens interativas, no entanto, é como reduzir o número de interações com o usuário, aumentando o efeito do seu esforço ao longo das iterações [7]. A necessidade de um número muito grande de *feedbacks* pode inviabilizar a solução, considerando que esta não seria tão vantajosa quando comparada a realização do alinhamento de forma manual.

A identificação de pares de entidades relevantes para *feedback* do usuário é o elemento central para a definição de uma abordagem interativa de alinhamento de ontologias. Portanto, este trabalho apresenta a seguinte questão a ser respondida: *Quais pares de entidades devem ser selecionados para feedback do usuário, de forma que sua participação contribua para a melhoria da qualidade do alinhamento?*

Para responder esta questão, o presente trabalho propõe a aplicação de técnicas de aprendizado de máquina, particularmente da subárea conhecida como *Active Learning*. Abordagens apresentadas em [8] e [9] têm empregado técnicas de aprendizado de máquina com sucesso no processo de alinhamento de ontologias.

Active learning é uma área de estudo que permite combinar estratégias de aprendizado de máquina com participação do usuário. A ideia geral é selecionar, dentre um conjunto de instâncias de entrada, as mais informativas para classificação por um oráculo. Uma vez classificadas, estas instâncias são utilizadas no treinamento de algoritmos que aprendem modelos de classificação [10]. Na abordagem proposta pelo presente trabalho, o oráculo é representado pelo usuário, as instâncias representam pares de entidades das ontologias a serem alinhadas e a classificação de uma instância indica se este par é (ou não é) uma correspondência.

A seleção de instâncias informativas representa um desafio para a área de *active learning* e dentre as estratégias propostas está a conhecida como *query-by-committee*. Na estratégia *query-by-committee,* um comitê de classificadores gera hipóteses de classificação para cada instância de entrada. Aquelas instâncias em que ocorre o maior desacordo entre os membros do comitê são consideradas as mais informativas.

A ideia de focar no desacordo entre os membros do comitê tem por objetivo reduzir o espaço de busca de instâncias para classificação pelo usuário. Esta característica parece ser bem apropriada no contexto da abordagem proposta para alinhamento de ontologias, que tem como objetivo aumentar o efeito da participação do usuário, solicitando *feedback* do usuário apenas sobre os pares de entidades relevantes.

Considerando as características da abordagem de seleção *query-by-committee* e a questão apresentada, este trabalho apresenta a seguinte hipótese de pesquisa: *Se a estratégia de seleção de instâncias informativas, conhecida como query-by-committee, for aplicada no processo de alinhamento de ontologias, então o efeito da participação do usuário será aumentado, contribuindo para a melhoria da qualidade dos alinhamentos obtidos.*

1.2 Objetivos

O objetivo geral deste trabalho é inserir o usuário (especialista no domínio) no contexto do processo de alinhamento de ontologias, explorando seu conhecimento a respeito do domínio e tornando a sua participação mais efetiva possível na busca por alinhamentos de melhor qualidade.

Para alcançar este objetivo, este trabalho propõe uma abordagem interativa para o alinhamento de ontologias. A elevação do efeito da participação do usuário é obtida selecionando pares de entidades relevantes aplicando a estratégia *query-by-committee.*

Além disso, este trabalho apresenta os seguintes objetivos específicos:

- Reduzir o número de iterações e *feedbacks* sobre pares de entidades solicitados ao usuário.
- Aumentar o efeito do *feedback* do usuário na melhoria da qualidade do alinhamento.

1.3 Metodologia de Pesquisa

Neste trabalho é empregado o método de pesquisa quantitativo. Este método é centrado na coleta de dados quantitativos com o objetivo de mensurar o estado de alguma variável de um determinado domínio no mundo real. Por esta característica, este método é utilizado para confirmar um modelo teórico previamente definido [11].

O método de pesquisa quantitativo aqui empregado segue o processo proposto em [11], que compreende as seguintes atividades:

1. **Geração da teoria e hipótese:** Inicialmente, foi realizada uma revisão da literatura a respeito dos desafios na área de Alinhamento de Ontologias, o que resultou na questão de pesquisa apresentada neste trabalho. Posteriormente,

foi feito um estudo direcionado à área de Aprendizado de Máquina com o objetivo de mapear possíveis soluções para a questão apresentada. O resultado deste estudo foi a hipótese de pesquisa apresentada anteriormente neste capítulo.

2. **Desenvolvimento de instrumentos para medição:** Para viabilizar a avaliação da proposta apresentada neste trabalho foi desenvolvido um protótipo de aplicação de alinhamento de ontologias que simula a interação com o usuário.
3. **Coleta de dados empíricos:** Com o objetivo de verificar as relações de causa e efeito entre as variáveis presentes na abordagem proposta foi utilizada uma abordagem experimental. Diferentes cenários foram definidos com variações no número de pares de entidades selecionados para feedback do usuário. Durante a execução destes cenários, a cada iteração realizada foram coletados dados de precisão, cobertura e medida-F, para cada par de ontologias que compõem o *data set* utilizado neste experimento.
4. **Análise de dados:** Os dados coletados foram analisados empregando a técnica descritiva. Seguindo a abordagem da OAEI, os dados foram agregados por *data set* e um valor médio foi determinado, considerando cada cenário e iteração realizada.
5. **Avaliação de resultados:** Os resultados obtidos pela abordagem proposta foram comparados com as soluções para alinhamento de ontologias avaliadas na trilha de *Interactive Matching* da OAEI.

1.4 Organização do Trabalho

Com o objetivo de orientar o leitor, esta dissertação está organizada em 7 capítulos incluindo esta introdução. A fundamentação teórica necessária à compreensão deste trabalho é apresentada no capítulo 2, que trata sobre os conceitos da área de Alinhamento de Ontologias, e no capítulo 3 que aborda a área de Aprendizado de Máquina e *Active Learning*. No capítulo 4 é apresentada a abordagem de alinhamento de ontologias proposta. Esta abordagem é avaliada no capítulo 5 através de um experimento utilizando o *conference data set* da OAEI. No capítulo 6, trabalhos da literatura que empregam a participação do usuário no processo de alinhamento e tem

relação com a abordagem proposta têm suas características apresentadas. Finalizando, no capítulo 7 são apresentadas as conclusões e os trabalhos futuros.

2 Alinhamento de Ontologias

Neste capítulo serão apresentados os principais conceitos relacionados à área de estudo conhecida como Alinhamento de Ontologias, incluindo as características de um conjunto de medidas de similaridade aplicadas neste trabalho, bem como medidas utilizadas para avaliação da qualidade dos alinhamentos. Finalizando, serão apresentadas as principais estratégias para envolvimento do usuário no processo de alinhamento de ontologias descritas na literatura.

2.1 Ontologias

Diversas definições foram dadas para Ontologia em diferentes áreas de conhecimento incluindo a Computação. Studer *et al.* [12] definem ontologia como uma *especificação explícita e formal de uma conceitualização compartilhada*.

O termo "conceitualização" refere-se ao fato de uma ontologia representar uma visão simplificada e abstrata de um universo de discurso [13], ou seja, um modelo abstrato formado pelos conceitos relevantes do universo de discurso analisado, e a forma como tais conceitos se relacionam entre si [12]. Estes conceitos devem ser especificados de maneira explicita, em que os tipos de conceitos e as restrições sobre sua conceituação são explicitamente definidos através de axiomas, e formal, ou seja, devem ser representados por uma linguagem que permita o processamento computacional [13] [12].

Apesar de na literatura existirem tipos de ontologias que variam com relação ao nível de abstração (ontologias de topo ou de fundamentação, ontologias de domínio, ontologias de tarefa e ontologias de aplicação) [13], no presente trabalho, é considerada apenas as ontologias de domínio.

A Figura 2.1 apresenta um recorte da ontologia *cmt*, que descreve um domínio de conferências acadêmicas, segundo a Microsoft[1].

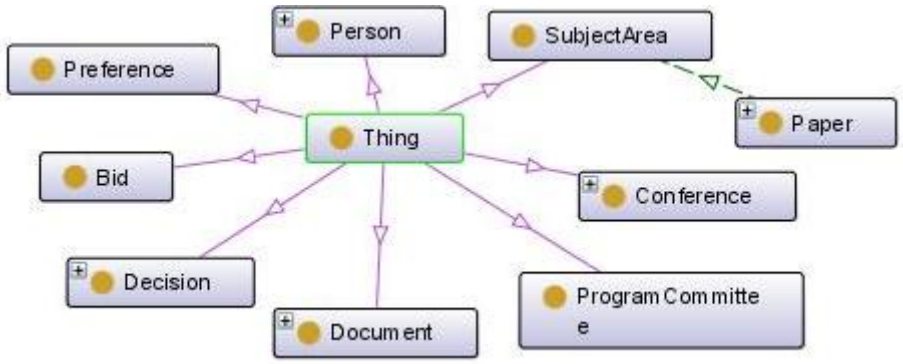

Figura 2.1 - Subconjunto de entidades da ontologia *cmt*

Ainda analisando a definição de Studer *et al.*[12], o termo "compartilhada" reflete o fato de que uma ontologia representa o conhecimento consensual de um grupo, e não de apenas um único indivíduo. Este termo está diretamente relacionado a características de interoperabilidade e comunicação estabelecidas entre indivíduos e entre máquinas [13]. Ehrig [5] destaca ainda que o termo "compartilhada" não significa necessariamente um consenso global, o que pode levar a definição de diferentes ontologias para um mesmo domínio. A Figura 2.2 apresenta um recorte da ontologia *conference*, que também descreve o domínio de conferências acadêmicas segundo a CRS[2]

[1] http://msrcmt.research.microsoft.com/cmt

[2] http://www.conferencereview.com/index.asp

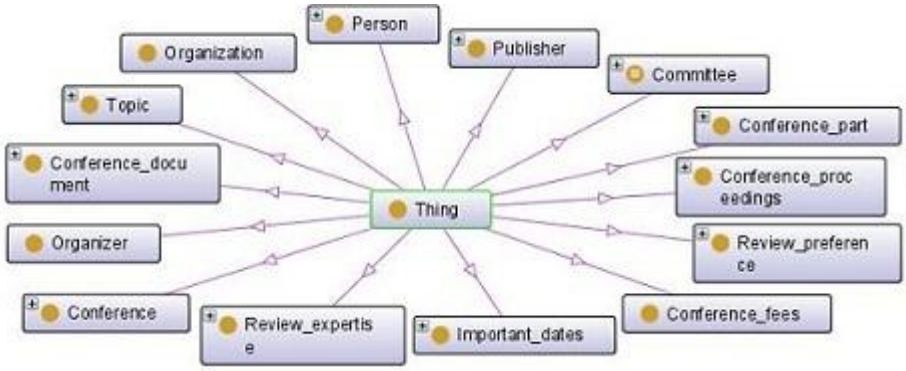

Figura 2.2 – Subconjunto de entidades da ontologia *conference*

2.2 Heterogeneidade Ontológica

A existência de diferentes ontologias para um mesmo domínio é cada vez mais frequente, uma vez que vários esforços têm sido empregados para o tratamento semântico de informações. Nestes cenários, é preciso lidar com diferentes níveis de heterogeneidade entre as ontologias de um mesmo domínio [3]:

- **Heterogeneidade sintática:** ocorre quando as ontologias não são definidas na mesma linguagem, por exemplo, OWL e F-logic.
- **Heterogeneidade terminológica:** ocorre quando os nomes de entidades equivalentes variam em ontologias diferentes, geralmente devido a linguagens naturais diferentes, ao uso de sinônimos, ou de terminologias proprietárias ou locais.
- **Heterogeneidade conceitual ou semântica:** ocorre quando há diferenças na modelagem de um mesmo domínio, ocasionadas pela utilização de diferentes axiomas para definir os conceitos ou devido ao uso de conceitos totalmente diferentes. Esta heterogeneidade ocorre por diferenças na cobertura (diferentes regiões do mundo), diferenças na granularidade (diferentes níveis e detalhes) e diferenças na perspectiva (diferentes perspectivas).
- **Heterogeneidade semiótica:** ocorre quando há variação na interpretação dada por pessoas para entidades com mesma interpretação semântica, dependo do contexto ou de como elas foram utilizadas ultimamente.

2.3 Processo de Alinhamento de Ontologias

A área de estudo de Alinhamento de Ontologias é uma das abordagens para o problema da heterogeneidade terminológica e semântica existente entre ontologias. A solução apresentada por esta área de estudo é baseada na identificação de correspondências entre entidades destas ontologias.

Dadas duas ontologias *O* e *O'* a serem alinhadas, um processo de alinhamento de ontologias busca encontrar correspondências entre entidades de *O* e *O'*. Uma correspondência é definida como uma 4-upla [4] [6]:

$$<id, e_1, e_2, r>$$

em que *id* é o identificador da correspondência, e_1 e e_2 são entidades de *O* e *O'*, respectivamente, e *r* é a semântica da relação entre e_1 e e_2, que pode ser: equivalência (=), disjunção (⊥) ou generalização (⊒) [6]. Um alinhamento entre *O* e *O*' é definido como o conjunto das correspondências encontradas. A Figura 2.3 apresenta graficamente o alinhamento entre as ontologias *conference* e *cmt*, em que cada aresta representa uma correspondência entre as entidades interligadas, e o *label* da aresta representa a semântica da correspondência.

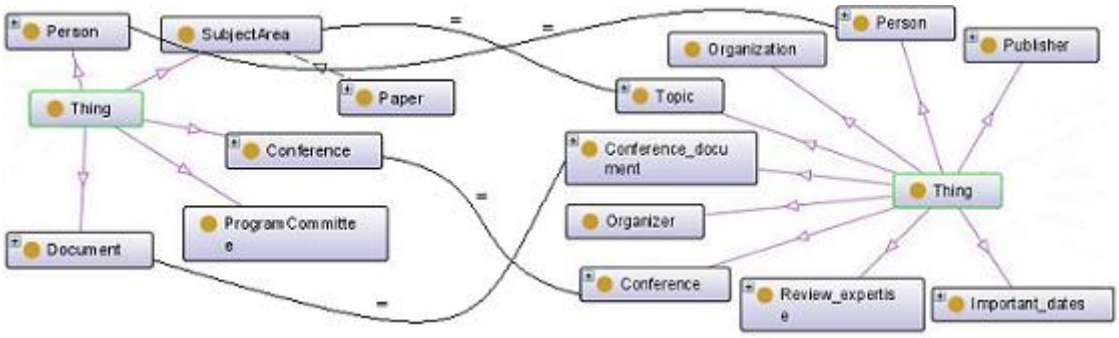

Figura 2.3 – Alinhamento entre as ontologias *cmt* e *conference*

O alinhamento da Figura 2.3 é, portanto, o conjunto das correspondências:

{< 1, *Person, Person,*= >, < 2, *SubjectArea, Topic,*= >, < 3, *Conference, Conference,*= >, < 4, *Document, Conference_Document,*= >}

Formalmente, Euzenat *et al.* [6] define o processo de alinhamento de ontologias como uma função *f* que, dado um par de ontologias *O* e *O'*, opcionalmente um alinhamento de entrada *A*, um conjunto de parâmetros *parameters* e um conjunto de recursos *resources*, retorna um alinhamento *A'* entre estas ontologias, conforme

apresentado na Figura 2.4. Considerando esta definição, parâmetros podem ser, por exemplo, pesos e *thresholds*, já os recursos externos podem ser, por exemplo, uma base léxica como a *Wordnet*[3].

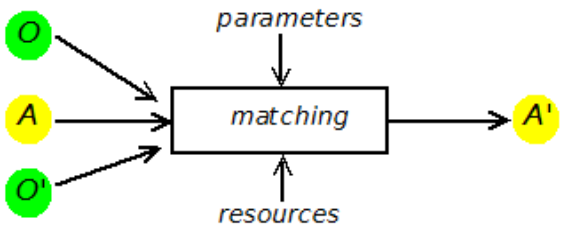

Figura 2.4 – Esquema do processo de alinhamento de ontologias [6]

2.4 Técnicas Básicas de Alinhamento de Ontologias

A identificação das correspondências durante o processo de alinhamento pode utilizar diversas técnicas básicas, que podem ser agrupadas segundo duas grandes perspectivas [4]:

- **Granularidade/interpretação da entrada (*Granularity/Input interpretation*):** Neste grupo, as técnicas básicas são organizadas pela granularidade, em que as entidades das ontologias são exploradas de forma isolada (*Element-level*) ou considerando as suas relações com outras entidades (*Structure-level*), e pela intepretação que fazem da entrada, considerando apenas a estrutura (*Syntactic*), utilizando apoio de recursos externos (*External*) ou algum formalismo semântico (*Semantics*).
- **Tipo de entrada (*Kind of input*):** Neste grupo, as técnicas básicas são organizadas considerando o tipo de entrada utilizado, que pode ser uma string (*Terminological*), estrutura (*Structural*), instâncias (*Extensional*) e modelos (*Semantic*). As técnicas terminológicas se subdividem ao considerar os termos como sequências de caracteres (*String-based*) ou como um objeto linguístico (*Linguistic*). O mesmo vale para as técnicas estruturais, que podem considerar apenas a estrutura interna como atributos e tipos (*Internal*) ou as relações entre entidades (*Relational*).

As técnicas básicas apresentadas na Figura 2.5 são descritas de forma resumida [4]:

[3] https://wordnet.princeton.edu/

- **Técnicas baseadas em string (*String-based*):** Estas técnicas exploram o nome e a descrição das entidades considerando estes como uma cadeia de caracteres. A ideia é que quanto mais similares estas strings forem maiores a possibilidade de representarem o mesmo conceito.
- **Técnicas baseadas em linguagem (*Language-based*):** Estas técnicas empregam processamento de linguagem natural nos nomes e descrições das entidades permitindo explorar características morfológicas das palavras.
- **Técnicas baseadas em restrições (*Constraint-based*):** Estas técnicas exploram restrições internas aplicadas na definição das entidades como tipos dos atributos, cardinalidade e chaves.
- **Recursos Linguísticos (*Linguistic resources*):** Empregam técnicas baseadas em linguagem, contudo utilizam recursos externos como tesauros, bases léxicas e dicionários, com o objetivo de explorar as relações linguísticas (sinônimos, por exemplo) entre as palavras.
- **Reuso de Alinhamento (*Alignment reuse*):** Estas técnicas apresentam outra forma de explorar recursos externos, no caso registros de alinhamentos de ontologias alinhadas anteriormente. A motivação é que ontologias a serem alinhadas são similares a outras ontologias já alinhadas.

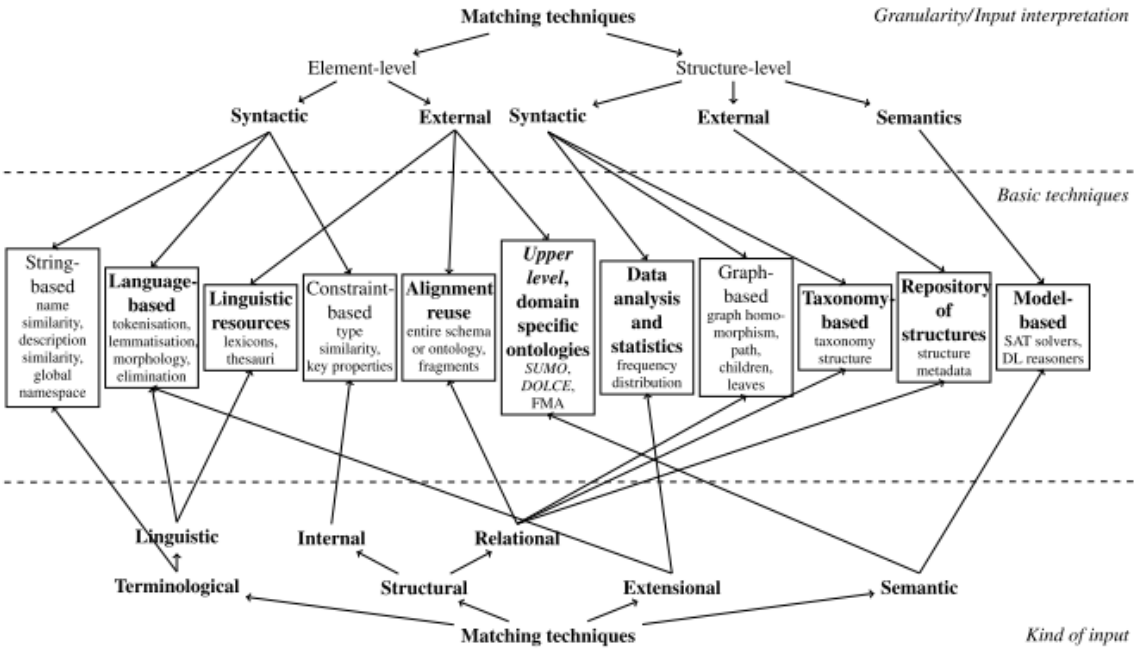

Figura 2.5 – Classificação das técnicas básicas de alinhamento de ontologias [4]

- **Ontologias de Fundamentação e Específicas de Domínio (*Upper level and domain specific ontologies*):** Ontologias de fundamentação podem ser utilizadas como fonte de conhecimento comum. Estas ontologias se caracterizam por serem sistemas baseados em lógica e podem ser explorados por técnicas de alinhamento baseadas em semântica. Por sua vez ontologias de domínio específico podem ser utilizadas como fonte de conhecimento prévio para o processo de alinhamento, neste caso focando em um domínio particular.
- **Técnicas Baseadas em Grafo (*Graph-based*):** Estas técnicas são baseadas em algoritmos de grafos, que consideram as ontologias como grafos rotulados. Geralmente a similaridade de um par de nós é determinada considerando a posição no grafo. A ideia é que se dois nós de duas ontologias são similares seus vizinhos também podem ser.
- **Técnicas Baseadas na Taxonomia (*Taxonomy-based*):** Estas técnicas também são baseadas em algoritmos de grafos, porém consideram somente as relações de especialização/generalização (*is-a*) entre entidades.
- **Repositório de Estruturas (*Repository of structures*):** Estes repositórios armazenam ontologias e seus fragmentos junto aos seus valores de similaridade. Diferentemente do reuso de alinhamento, o repositório de estrutura armazena valores de similaridade e não os alinhamentos. Ao alinhar uma nova estrutura,

primeiro é checada a similaridade das estruturas já existentes. O objetivo é identificar estruturas semelhantes.

- **Análise de Dados e Técnicas Estatísticas (*Data analysis and statistics*):** Estas técnicas exploram um conjunto de amostras de uma população com o objetivo de encontrar regularidades e discrepâncias entre elas.
- **Técnicas baseadas em Modelo (*Model-based*):** Estas técnicas analisam as entradas através de interpretações semânticas. A ideia é que duas entidades representam o mesmo conceito se compartilham a mesma interpretação. Dentre as técnicas empregas estão as baseadas em lógica de descrição [14].

As técnicas básicas de alinhamento apresentadas anteriormente identificam correspondências entre entidades das ontologias através do uso de medidas de similaridade propostas na literatura. Euzenat *et al.* [4] definem medida de similaridade $\sigma: O \times O' \rightarrow \mathbb{R}$ como uma função que dado um par de entidades retorna um número real que expressa a similaridade entre elas. Esta função deve satisfazer as seguintes propriedades:

$$\forall x \in O \wedge \forall y \in O', \sigma(x,y) \geq 0 \text{ (positividade)}$$

$$\forall x \in O \wedge \forall y, z \in O', \sigma(x,x) \geq \sigma(y,z) \text{ (maximalidade)}$$

$$\forall x \in O \wedge \forall y \in O', \sigma(x,y) = \sigma(y,x) \text{ (simetria)}$$

O número real retornado pela função pode variar dependendo do método empregado pela medida e dos recursos utilizados por ela no cálculo da similaridade. Neste trabalho, serão empregadas, no contexto do processo de alinhamento de ontologias, medidas de similaridade baseadas em *string* e medidas de similaridade semânticas. O processo de seleção das medidas foi baseado em dois critérios: implementações disponíveis e o resultado destas medidas em avaliações, como as realizadas em [15] e [16].

Medidas de similaridade baseadas em *string* focam na estrutura da *string*, como uma sequência de letras, ou caracteres. A comparação entre *strings* pode ser realizada analisando a sequência de letras corretas, a sequência de letras incorretas, um conjunto de letras ou mesmo um conjunto de palavras [4].

No contexto do processo de alinhamento de ontologias empregado neste trabalho, duas *strings s* e *s'* correspondem aos *labels* das entidades *e* e *e'* e a similaridade entre elas é determinada empregando as medidas descritas a seguir:

- **Jaccard:** Esta medida de similaridade considera *strings* como conjuntos de termos (palavras). Dados dois conjuntos de termos *A* e *B*, a similaridade jaccard é determinada conforme a expressão a seguir [17]:

$$\sigma_{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

em que $|A \cap B|$ representa o número de termos comuns aos conjuntos *A* e *B* e $|A \cup B|$ representa o número total de termos nos conjuntos *A* e *B*. Assumindo *s = "paper author"* e *s' = "article author"* tem-se os conjuntos A = {*"paper", "author"*} e B = {*"article", "author"*}. Aplicando a medida jaccard obtém-se: $|A \cap B|$ = 1 e $|A \cup B|$ = 3, portanto, a similaridade retornada por esta medida é igual a 0,33.

- **JaroWinkler:** A medida de similaridade jaro é baseada no número e na ordem dos caracteres comuns existentes entre duas *strings*, sendo calculada conforme a expressão seguinte [4]:

$$\sigma_{jaro}(s, s') = \frac{1}{3} \times \left(\frac{|com(s, s')|}{|s|} + \frac{|com(s', s)|}{|s'|} + \frac{|com(s, s')| - |transp(s, s')|}{|com(s, s')|} \right)$$

em que $|com(s, s')|$ corresponde ao número de caracteres comuns entre as *strings s* e *s'* e $|transp(s, s')|$ é o número de caracteres comuns transpostos nas *strings s* e *s'*. A medida de similaridade jarowinkler é uma variação da medida jaro, que busca melhorar os resultados de *strings* que possuem prefixos comuns. Esta medida é determinada conforme a expressão seguinte [17]:

$$\sigma_{jarowinkler}(s, s') = \sigma_{jaro}(s, s') + \frac{P}{10} \times (1 - \sigma_{jaro}(s, s'))$$

em que *P* é o tamanho do prefixo comum. Considerando, por exemplo, as *strings s* = "*author"* e *s'* = "*auhtor"*, e aplicando a medida jarowinkler, tem-se: $|com(s, s')|$ = 6, $|com(s', s)|$ = 6 e $|transp(s, s')|$ = 1. Portanto, o valor de similaridade aplicando esta medida é 0,93.

- ***n*-gram:** Esta medida calcula o número de *n-grams* comuns, como uma sequência de *n* caracteres, entre duas *strings*. O valor de similaridade é dado pela expressão seguinte:

$$\sigma_{n-gram}(s, s') = \frac{|ngram(s,n) \cap ngram(s',n)|}{\min(|s|,|s'|) - n + 1}$$

em que *ngram(s,n)* é o conjunto de *substrings* de *s* com tamanho *n*. Considerando, por exemplo, *n* = 3 e as strings *s* = "*author"* e *s'* = "*autor"*, tem-se: *ngram(s,n)* = {*"aut"*, *"uth"*, *"tho"*, *"hor"*} e *ngram(s',n)* = {*"aut"*, *"uto"*, *"tor"*}. Portanto, aplicando a medida *n*-gram, obtém-se o valor de similaridade igual a 1.

As medidas de similaridade semânticas podem ser utilizadas para determinar a similaridade de conceitos que muitas vezes não são lexicamente similares. Para isto, estas medidas exploram relacionamentos linguísticos entre conceitos utilizando recursos externos, como a *WordNet* [15]. Devido a esta característica, estas medidas são classificadas como baseadas em recursos linguísticos, segundo a classificação apresentada na Figura 2.5.

A *WordNet* pode ser definida como uma base de dados léxica, organizada como uma taxonomia de conceitos. Na *WordNet,* cada conceito é representado por um *synset*, que corresponde a um conjunto de sentidos (*senses*). Um sentido (*sense*) corresponde a um dos possíveis significados de um conceito em um determinado contexto. Uma vez que estão organizados em uma estrutura taxonômica, os conceitos são relacionados a outros conceitos mais altos ou baixos na hierarquia através de diferentes tipos de relacionamentos, em que os mais comuns são os de generalização/especialização (*Hypernym/Hyponym*) e todo/parte de (*Meronym/Holonym*) [15]. Desta forma, dado um par de conceitos (*c*, *c'*) presente na taxonomia, pode-se definir o "conceito comum mais específico" (*c''*) deste par como sendo o conceito mais específico (mais baixo na hierarquia) que generaliza tanto *c* quanto *c'*.

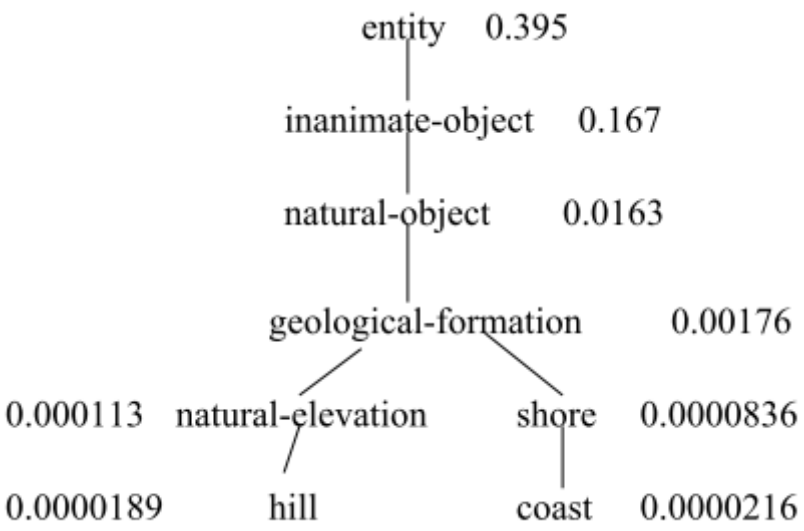

Figura 2.6 – Exemplo de relacionamento de generalização/especialização na *Wordnet*

A Figura 2.6 apresenta uma ilustração do relacionamento generalização/especialização para os conceitos *"hill"* (*c*) e *"coast"* (*c'*). Nesta figura, o conceito comum mais específico, considerando estes dois conceitos é *"geological-formation"* (*c''*). Nesta figura, os valores apresentados ao lado dos conceitos correspondem a probabilidade de ocorrência em um *corpus*.

Lin e Sandkuhl [18] classificam as medidas de similaridade semânticas em: **(i) baseadas nas arestas (*edge-based*)**, nas quais a similaridade entre dois conceitos é determinada pela distância (caminho) entre os conceitos e a posição do conceito na taxonomia; (ii) **baseadas na informação (*information-based*)**, nas quais a similaridade é determinada considerando a quantidade de informação existente entre os conceitos em função das suas probabilidades de ocorrência em um *corpus*; e (iii) **Híbridas,** nas quais a similaridade é determinada combinando as duas abordagens anteriores.

Dados dois conceitos *c* e *c'* presentes em uma taxonomia, e sendo *c''* o conceito comum mais específico de *c* e *c'*, o cálculo da similaridade entre *c* e *c'* pode ser realizado empregando alguma medida de similaridade semântica. As medidas de similaridade semânticas empregadas neste trabalho são as propostas em Wu e Palmer [19], Lin [20] e Jiang e Conrath [21], descritas a seguir:

- **Wu e Palmer** é uma medida de similaridade baseada nas arestas, que é determinada considerando o caminho entre *c''* e cada um dos conceitos de entrada e o caminho de *c''* até a raiz da taxonomia, conforme a expressão seguinte:

$$\sigma_{wupalmer}(c, c') = \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3}$$

em que N_1 é o número de nós no caminho de *c* até *c''*, N_2 é o número de nós no caminho de *c'* até *c''* e N_3 é o número de nós no caminho de *c''* até a raiz da taxonomia. Considerando a taxonomia da Figura 2.6, e os conceitos *"hill"* e *"coast"*, tem-se: $N_1 = 2$ e $N_2 = 2$ e $N_3 = 3$. Portanto, aplicando a medida Wu e Palmer, a similaridade entre os conceitos *"hill"* e *"coast"* é 0,6.

- **Lin** é uma medida de similaridade baseada na informação, que é determinada considerando a quantidade de informação necessária para indicar o quão comum dois conceitos são e que informações específicas são necessárias para descrevê-los completamente, conforme a expressão seguinte:

$$\sigma_{lin}(c, c') = \frac{2 \times \log p(c'')}{\log p(c) + \log p(c')}$$

em que *p(c)* é a probabilidades de ocorrência de *c* em um *corpus*. Considerando a taxonomia da Figura 2.6, e os conceitos *"hill"* e *"coast"*, tem-se: *p(c)* = 0,0000189 , *p(c')* = 0,0000216 e *p(c'')* = 0,00176. Portanto, aplicando a medida Lin, a similaridade entre os conceitos *"hill"* e *"coast"* é 0,58.

- **Jiang e Conrath** é uma medida de similaridade híbrida, derivada de ideias da abordagem baseada em aresta e que considera a quantidade de informação (*information content* - *ic*) como fator de decisão. A similaridade entre dois conceitos *c* e *c'* é determinada pelas expressões seguintes:

$$ic(x) = -\log p(x)$$
$$d(c, c') = ic(c) + ic(c') - 2 \times ic(ls(c, c'))$$
$$\sigma_{jiangconrath}(c, c') = \frac{1}{d(c, c')}$$

em que *d(c,c')* é a distância entre os conceitos *c* e *c'* e *ls(c,c')* corresponde ao conceito comum mais específico *c''*. Considerando a taxonomia da Figura 2.6, e os conceitos *"hill"* e *"coast"*, tem-se: *ic*(*c*) = 4,72 , *ic*(*c'*) = 4,66 e p(*c''*) = 2,75.

Portanto, aplicando a medida derivada da distância Jiang e Conrath, a similaridade entre os conceitos *"hill"* e *"coast"* é 0,26.

2.5 Medidas de Avaliação de alinhamentos

Com o desenvolvimento de novas propostas para o processo de alinhamento de ontologias, torna-se necessário estabelecer mecanismos de avaliação que permitam verificar a qualidade dos alinhamentos gerados por estas novas abordagens. O processo de avaliação geralmente consiste em realizar um cruzamento entre os alinhamentos gerados e um alinhamento de referência, formado pelas correspondências entre entidades realmente existentes ("gabarito").

As medidas de avaliação devem ser escolhidas de modo a não beneficiar abordagens específicas [5]. Dados um alinhamento de referência *R* e algum alinhamento *A*, as principais medidas empregadas na avaliação dos alinhamentos são descritas a seguir.

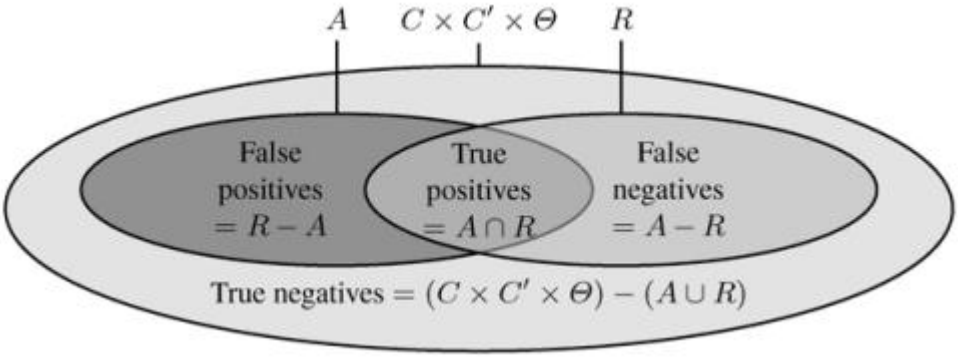

Figura 2.7 – Relação entre um alinhamento e sua referência [4]

- **Precisão:** mede a taxa de correspondências corretamente encontradas (*true positives*) sobre o número total de correspondências retornadas (A). A precisão é uma função *P*:

$$P(A,R) = \frac{|R \cap A|}{|A|}$$

- **Cobertura:** mede a taxa de correspondências corretamente encontradas (*true positives*) sobre o número total de correspondências esperadas (*R*). A cobertura é uma função *C*:

$$C(A,R) = \frac{|R \cap A|}{|R|}$$

- **Medida-F:** agrega as medidas de precisão e cobertura. Segundo Ehrig [5], esta é a medida mais importante para avaliar a qualidade de um alinhamento. Dado um número α entre 0 e 1, a medida-F é uma função F_α:

$$F_\alpha(A,R) = \frac{P(A,R) \times C(A,R)}{(1-\alpha) \times P(A,R) + \alpha \times C(A,R)}$$

  Geralmente o valor de α é igual a 0.5, neste caso a medida-F é uma média harmônica entre precisão e cobertura.

2.6 Alinhamento de Ontologias com Participação do Usuário

Processos de alinhamento de ontologias requerem a definição de uma estratégia que permita combinar diferentes técnicas básicas e agregar os seus resultados para geração de um alinhamento [4]. Na literatura, algumas soluções apresentam abordagens automáticas para o processo de alinhamento de ontologias, que envolve tanto combinação de técnicas, quanto agregação de resultados.

Ehrig [5], contudo, destaca que é muito difícil criar uma abordagem completamente automática e mesmo as melhores podem não apresentar resultados satisfatórios. Neste sentido, determinadas abordagens têm apresentado características customizáveis que acabam por determinar a forma de participação do usuário no processo de alinhamento de ontologias. Esta participação, segundo Euzenat *et al.* [4] pode ocorrer em três diferentes áreas:

- **Fornecendo entradas:** O usuário pode fornecer, além das ontologias a serem alinhadas, parâmetros ou mesmo um alinhamento inicial. Os parâmetros podem se apresentar, por exemplo, como pesos (para as técnicas básicas) ou *thresholds* (para identificar as correspondências que farão parte do alinhamento) e se bem configurados podem melhorar os resultados obtidos. Já um alinhamento inicial (em algumas abordagens chamado de *gold standard*) pode ser empegado para direcionar o comportamento do sistema de alinhamento, como por exemplo, em um processo de aprendizado.

- **Combinando técnicas de alinhamento:** Determinadas abordagens apresentam um conjunto de técnicas de alinhamento ao usuário e atribuem a ele o papel de selecionar quais destas técnicas serão utilizadas, realizar seu sequenciamento e combinação, e definir o método de agregação dos resultados a ser aplicado para geração do alinhamento. Este tipo de abordagem requer um conhecimento das técnicas de alinhamento e dos métodos de agregação por parte do usuário.
- **Fornecendo feedbacks relevantes:** Nesta área o usuário é solicitado a fornecer feedbacks sobre possíveis correspondências (correspondências candidatas) informativas. O principal desafio, conforme destacado por Shi *et al.* [7] e Cruz *et al.* [22], é identificar as mais informativas, reduzindo o número de interações necessárias com o usuário. Geralmente estas abordagens são definidas sob um processo iterativo, em que o resultado do *feedback* do usuário é utilizado na iteração seguinte para ajustar parâmetros, *thresholds* ou mesmo um classificador.

3 Aprendizado de Máquina

Neste capítulo será apresentada uma introdução à área de estudo de Aprendizado de Máquina, em que serão descritos os principais conceitos relacionados, bem como as abordagens de classificação empregadas neste trabalho. Ao final do capítulo será descrita a subárea conhecida como Active Learning.

3.1 Classificação

Um dos grandes desafios da Computação, desde a sua criação, está na busca por tornar os computadores capazes de aprender, de se aprimorar automaticamente com base na experiência [23]. Algoritmos que implementam determinadas tarefas de aprendizado têm sido desenvolvidos e aplicados em diversas áreas como: classificação de texto, processamento de linguagem natural, reconhecimento de voz, detecção de fraldes, diagnósticos médicos, sistemas de recomendação, reconhecimento de faces, entre outras [24].

Aprendizado de Máquina pode ser definido como a aquisição de descrições estruturais a partir de exemplos [25]. Para isso, realiza-se uma busca em um grande espaço de possíveis hipóteses com o objetivo de encontrar a solução mais aderente aos dados observados e ao conhecimento prévio trazido pelo *learner* [23].

O Aprendizado de Máquina fornece a base técnica para a mineração de dados, que pode ser definida como a extração de informação implícita, previamente desconhecida e potencialmente útil dos dados [25], com o objetivo de produzir um modelo preditivo [26]. Dentre os modelos preditivos que podem ser aprendidos, estão os classificadores [26].

Classificação é um processo cognitivo fundamental utilizado para organizar e aplicar o conhecimento a respeito do mundo. No dia-a-dia, a classificação é uma tarefa comumente realizada, pois em diversos momentos são atribuídas classes ou categorias significativas a instâncias de um determinado domínio [26].

Tan *et al.* [27] define classificação como a tarefa de aprender uma função *f* que mapeia um conjunto de atributos *x* em uma das classes predefinidas *y*. Informalmente, a função *f* é conhecida como modelo de classificação, ou classificador. Um modelo de classificação pode ser empregado para prever a classe de instâncias desconhecidas, dado seu conjunto de atributos. A Figura 3.1 apresenta o modelo de classificação (*classification model*) como uma caixa preta, em que dado uma instância *x* e seus atributos (*attribute set*), é retornada a classe (*class label*) *y*.

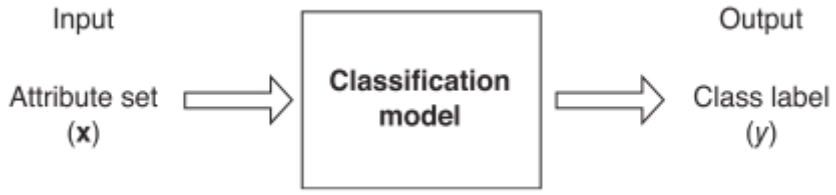

Figura 3.1 – Modelo de Classificação [27]

Classificação é, portanto, uma abordagem sistemática que constrói modelos de classificação a partir de um conjunto de dados de entrada. Para isto, são empregados algoritmos de aprendizado que encontram o modelo mais apropriado considerando o conjunto de atributos e a classe dos dados de entrada. Este modelo deve ser aderente aos dados de entrada, bem como capaz de prever a classe de instâncias desconhecidas [27]. A Figura 3.2 apresenta a abordagem geral para classificação.

Na Figura 3.2, o conjunto de treinamento (*training set*) é formado por instâncias em que o atributo de classe é conhecido. Este conjunto é utilizado pelo algoritmo de aprendizado (*learning algorithm*) para indução de um modelo de classificação (*learn model*) que posteriormente é aplicado em um conjunto de instâncias para dedução do atributo de classe antes desconhecido [27].

Nas subseções seguintes serão descritas as técnicas de classificação utilizadas neste trabalho, a saber: Naive Bayes [28], Random Forest [29] e Multilayer Perceptron [30]. Tais técnicas foram escolhidas por apesentarem bons resultados em trabalhos anteriores

na área de Alinhamento de Ontologias [31] [32] e por permitirem perspectivas diferentes na classificação de instâncias.

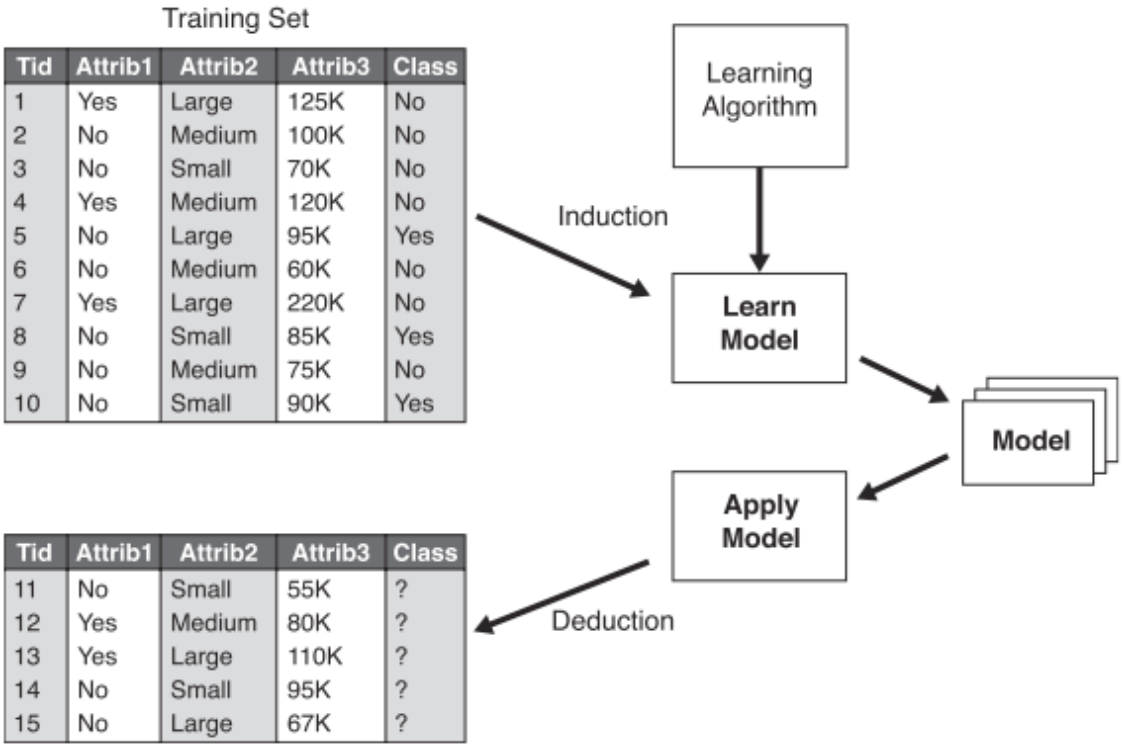

Figura 3.2 – Abordagem geral para classificação [27]

3.1.1 Naive Bayes

Classificadores bayesianos fornecem uma visão probabilística para o problema da classificação [23]. O objetivo principal desta abordagem é determinar a probabilidade condicional de cada classe em função dos atributos. Para isto, estes classificadores utilizam o teorema de Bayes [33]:

$$\Pr(C \mid A_1,...,A_k) = \alpha \Pr(C) \Pr(A_1,...,A_k|C)$$

em que Pr(C) é a probabilidade prévia da classe C (estimada a partir das instâncias de treinamento), Pr (A_1,...,A_k|C) é a distribuição de probabilidades dos valores dos atributos A_1,...,A_k dada a classe C, e α é um fator de normalização que garante que a probabilidade condicional de todas as possíveis classes somadas alcance o valor 1 [33].

Um dos classificadores bayesianos mais simples é conhecido como Naive Bayes [23]. Este classificador tem como principal característica a independência entre cada atributo de entrada e o atributo de classe, ou seja: para uma classe *C*, a probabilidade de cada atributo é estimada de forma independente dos demais, como na expressão:

$$\Pr(C \mid A_1,...,A_k) = \alpha \Pr(C) \Pr(A_1|C)... \Pr(A_k|C)$$

As probabilidades são estimadas a partir das instâncias de treinamento e a probabilidade posterior é calculada para cada classe. Neste caso, a predição é feita para a classe com a maior probabilidade posterior obtida [33].

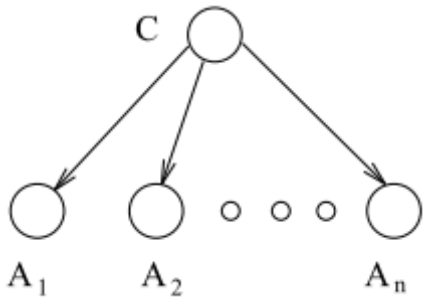

Figura 3.3 – Estrutura de uma rede naive bayes [28]

Um classificador Naive Bayes pode ser representado através de uma rede bayesiana, conforme Figura 3.3. Uma rede bayesiana é um grafo direcionado acíclico que permite representar de forma eficiente a distribuição de probabilidades, em que cada vértice corresponde a um atributo e cada aresta representa a correlação direta entre estes atributos [28].

A rede bayesiana permite definir a independência entre os atributos (representados pelas folhas da rede) em relação ao atributo de classe (representado pela raiz). A independência é explorada com o objetivo de reduzir o número de parâmetros necessários para caracterizar a distribuição de probabilidades permitindo calcular de forma eficiente a probabilidade posterior [28].

3.1.2 Random Forest

Diferentemente da visão probabilística apresentada por Naïve Bayes, alguns classificadores como árvores de decisão apresentam uma abordagem baseada em regras.

A indução de árvores de decisão é um método para obter uma função de classificação baseada em valores discretos, em que a função aprendida (classificador) é representada por uma árvore de decisão [23].

Uma árvore de decisão é uma estrutura de decisão que tem por objetivo determinar a classe de uma dada instância. Cada nó da árvore de decisão representa um atributo, e cada aresta representa uma condição que particiona o espaço de instâncias de acordo com o resultado do teste. Cada subconjunto da partição corresponde a um subproblema de classificação que é resolvido a partir da subárvore [34]. A Figura 3.4 apresenta um exemplo de árvore de decisão.

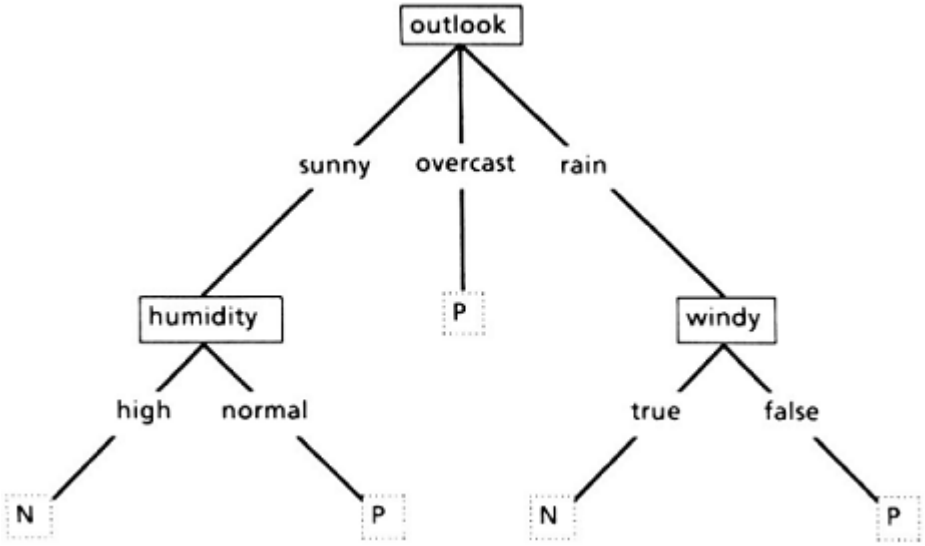

Figura 3.4 – Árvore de decisão [35]

Estruturalmente, uma árvore de decisão é composta de nós folha ou de resposta (que representam uma classe) e nós não folha ou de decisão, que contém um atributo de teste com ramificações para outra árvore de decisão para cada possível valor do atributo [34].

Diversos algoritmos foram desenvolvidos para aprendizado de máquinas através de árvores de decisão, dentre eles o algoritmo ID3, com implementações como a J4.8 [34]. Ambos os algoritmos são variações de um algoritmo principal que realiza uma busca top-down no espaço de árvores de decisão possíveis [23].

Árvores de decisão podem ser empregadas na criação de um *ensemble*. Um *ensemble* tem como ideia central a criação e combinação de múltiplos classificadores para um mesmo domínio com o objetivo de melhorar a qualidade da classificação. O principal desafio neste caso é criar modelos de classificação que sejam ao mesmo tempo

de boa qualidade e diversificados. Entre as estratégias para combinação estão a utilização de diferentes instâncias de treinamento, diferentes algoritmos, diferentes parâmetros de configuração ou mesmo algoritmos randômicos [26].

Random Forest pode ser definido como um *ensemble* formado pela combinação de classificadores baseados em árvores de decisão, de forma que cada árvore é definida com base nos valores de um vetor de amostras aleatórias independentes e com a mesma distribuição para todas as árvores da floresta [29].

Breiman [29] define Random Forest como um classificador composto por uma coleção de classificadores representados como árvores de decisão estruturadas $\{h(x,\Theta_k), k=1,...\}$, em que $\{\Theta_k\}$ são vetores aleatórios distribuídos identicamente e independentes e cada árvore representa um único voto para a classe mais popular na entrada x.

Em um modelo de classificação com Random Forest, as instâncias de treinamento são amostras geradas a partir das substituições de instâncias do conjunto de treinamento inicial. Esta técnica é conhecida como *bootstrapping*. Para cada conjunto de treinamento é gerada uma árvore em que os nós são atributos selecionados de forma aleatória. A classificação de novas instâncias é realizada agregando as classificações atribuídas por cada uma das árvores na forma de votos [36].

3.1.3 Multilayer Perceptron

Outra abordagem de aprendizado de máquina muito utilizada e que emprega técnicas diferentes das apresentadas anteriormente para classificação de instâncias é conhecida como Multilayer Perceptrons (MLP). MLP é definida como uma rede neural *feed-forward* formada por um conjunto de unidades chamadas neurônios conectados através de links com pesos definidos.

Os neurônios são distribuídos em um conjunto de camadas (*layers*) formadas por uma camada de entrada (*input layer*), uma ou mais camadas ocultas (*hidden layer*) e uma camada de saída (*output layer*) [37]. A Figura 3.5 apresenta a estrutura do Multilayer Perceptron.

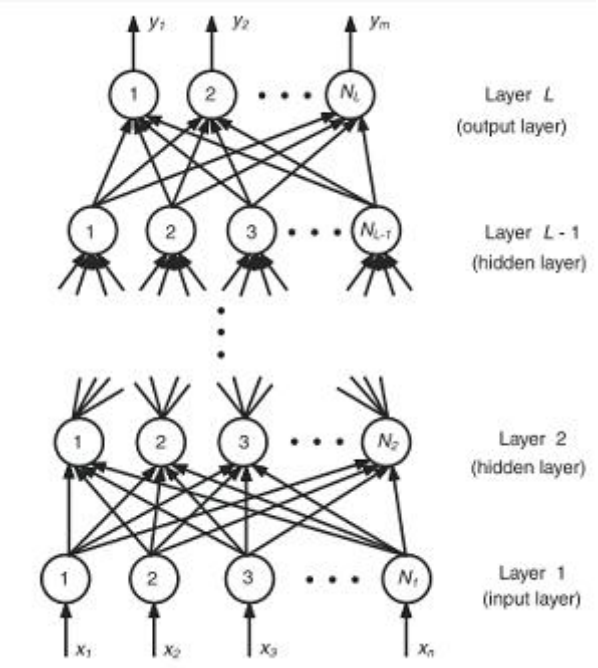

Figura 3.5 – Estrutura do Multilayer Perceptron [30]

Os neurônios que compõem a rede processam a informação recebendo entradas *z* de outros neurônios. Cada entrada é multiplicada pelo peso *w* correspondente e o resultado obtido é adicionado para calcular o peso total. Este peso é passado para uma função de ativação σ produzindo a saída do neurônio. Geralmente a função *sigmoid* é aplicada como função de ativação. A Figura 3.6 apresenta a estrutura de processamento do neurônio [30].

O principal objetivo do modelo neural é determinar pesos ótimos fazendo um mapeamento entre ativações de entrada para saída considerando um conjunto de exemplos [37]. O algoritmo *backpropagation* aprende os pesos dada uma rede composta de um conjunto fixo de unidades e suas interconexões. Este algoritmo aplica a técnica *gradient descent* para reduzir o erro entre os valores de saída da rede e os valores esperados para esta saída [23].

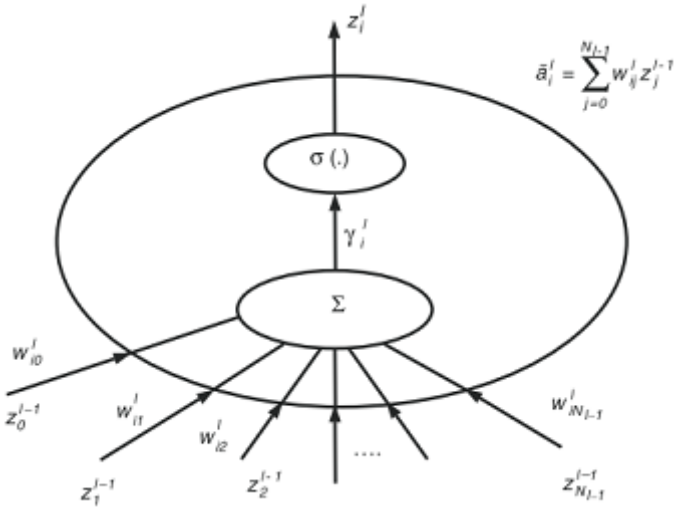

Figura 3.6 – Estrutura de processamento do neurônio [30]

Durante o treinamento o desempenho da rede é avaliado calculando a diferença entre as saídas geradas pela rede atual e as saídas desejadas. Esta diferença é calculada aplicando uma função *E* que determina o erro [30]. A *gradient* de *E* indica a direção em que ocorre o incremento mais acentuado de *E*, sendo determinada calculando a derivada de *E* em relação a cada componente do vetor *w* [23].

A negativa da *gradient* indica a direção de decremento do valor de *E*, portanto, a *gradient descent* é calculada como a negativa da *gradient* multiplicada pela taxa de aprendizado $\in$. A taxa de aprendizado permite moderar o grau de mudança nos pesos a cada iteração.

A atualização do peso ocorre de forma iterativa considerando cada instância de treinamento. A cada iteração a alteração no peso é determinada calculando a *gradient descent* e adicionando a alteração no peso, determinada na iteração anterior, multiplicada pela constante *momentum* μ. A constante momentum permite definir a influência da iteração anterior no cálculo corrente [23] [25]. A expressão seguinte apresenta o cálculo para determinação da alteração do peso a cada iteração [37].

$$\Delta w_{ij}(t) = -\in \frac{\partial E}{\partial w_{ij}}(t) + \mu \Delta w_{ij}(t-1)$$

3.2 Clustering

A classificação é uma importante atividade na área de mineração de dados, pois permite atribuir uma categoria pré-definida a instâncias de um determinado domínio. Contudo,

em determinados cenários, estas categorias pré-definidas não existem e para sua identificação, torna-se necessário compreender estas instâncias considerando os dados que as descrevem. Esta atividade pode ser realizada empregando a técnica de aprendizado de máquina conhecida como *clustering* [27].

A técnica de *clustering* consiste na divisão de um conjunto de instâncias de um determinado domínio em grupos considerando a sua similaridade. A similaridade é determinada através dos atributos que descrevem estas instâncias. O objetivo é que instâncias em um grupo sejam similares entre si e diferentes das instâncias em outros grupos [26] [27].

Dentre os diferentes tipos de *clustering* apresentados na literatura, está a família conhecida como *k-center*. Algoritmos nesta família possuem o número de clusters predefinido e representado pela variável *k* e um vetor de valores (atributos de uma instância) que representam os clusters genericamente, chamados *clusters centers*. O processo de formação dos *clusters* é executado atribuindo as instâncias iterativamente ao cluster com *cluster center* mais similar, seguido do deslocamento do *cluster center* para refletir o conteúdo do cluster com a nova instância [26].

3.2.1 Farthest First

A definição dos *clusters centers* é considerada um problema NP-difícil. O algoritmo Farthest First Traversal [38] foi proposto como uma aproximação para o problema do *k-center*. A ideia central deste algoritmo é selecionar um ponto (instância) inicialmente, e então escolher o ponto mais distante deste, seguido pela escolha do ponto mais distante dos dois primeiros e assim sucessivamente até que sejam obtidos os *k* pontos. Os pontos selecionados são definidos como *clusters centers*, e os demais pontos são atribuídos ao cluster com *cluster center* mais similar [38]. A Figura 3.7 apresenta um exemplo da aplicação deste algoritmo, considerando dez pontos.

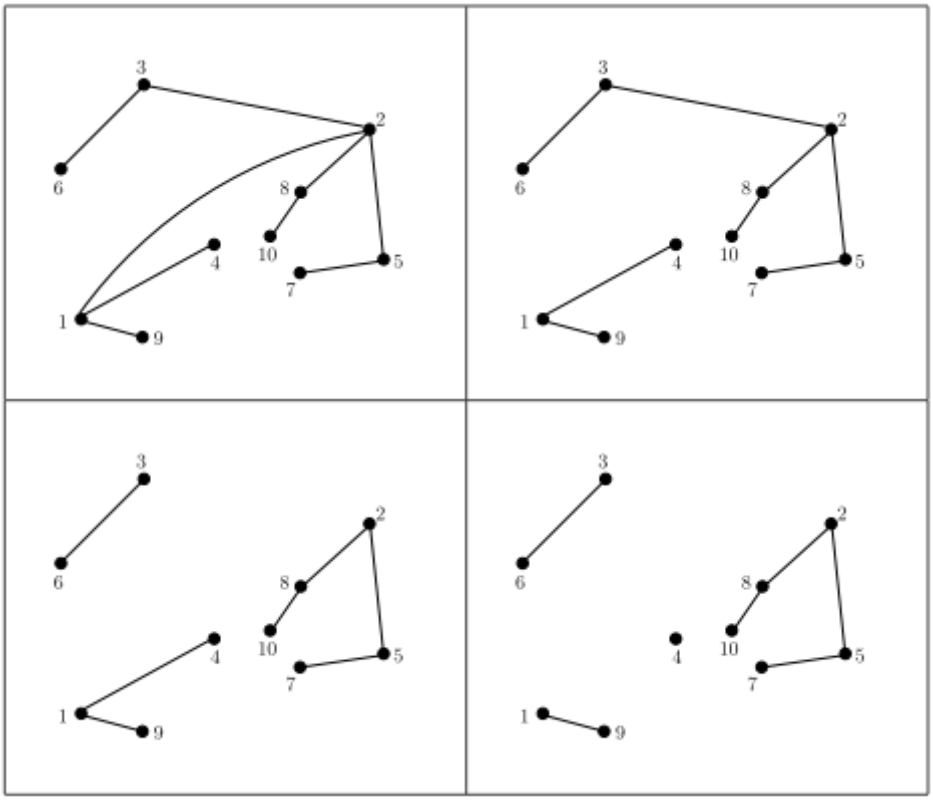

Figura 3.7 – Agrupamento em 1, 2, 3 e 4 clusters com Farthest First [38]

Analisando a Figura 3.7, é possível observar que os pontos 1 e 2 foram selecionados como *clusters centers*, no agrupamento em dois clusters. Nesta divisão, o ponto 2 é o ponto mais distante do ponto 1, sendo então o ponto selecionado. No caso da divisão em três *clusters*, o ponto 3 foi escolhido, por ser o ponto mais distante dos dois primeiros. O mesmo vale para a divisão em quatro *clusters*, com a seleção do ponto 4, por ser o ponto mais distante dos três primeiros.

3.3 Active Learning

Conforme apresentado na Figura 3.2, para construir um modelo de classificação é necessário um grande conjunto de treinamento formado por instâncias que possuem seu atributo de classe definido. Contudo, em diversos cenários em que o aprendizado de máquina é empregado, instâncias classificadas não estão disponíveis e a sua classificação é custosa e consome tempo [10].

Cenários como este, em que existe uma grande quantidade de instâncias não classificadas e poucas classificadas, são ideais para se empregar a subárea do

aprendizado de máquina conhecida como *active learning*. *Active learning* tem por objetivo alcançar alta acurácia na classificação fazendo uso de poucas instâncias classificadas, e consequentemente, minimizando o custo de sua obtenção [10].

No *Active learning*, o *learner* possui um determinado nível de controle que permite a ele selecionar as instâncias que serão classificadas e adicionadas ao conjunto de treinamento. A classificação das instâncias ocorre na forma de consultas a um oráculo, como por exemplo, um usuário [10] [39].

Algumas variações de *active learning* vêm sendo estudadas e uma das que recebe mais atenção na literatura, principalmente na área de mineração de dados, é conhecida como *pool-based sampling* [40]. *Pool-based sampling* ocorre em cenários em que há um pequeno conjunto de instâncias classificadas e um grande *pool* de instâncias não classificadas. Instâncias consideradas mais informativas são selecionadas do *pool* de instâncias não classificadas para consulta ao oráculo. Uma vez classificadas, estas instâncias são então incluídas no conjunto de instâncias classificadas e posteriormente utilizadas para treinamento do classificador. A Figura 3.8 apresenta uma ilustração de *pool-based sampling* [10].

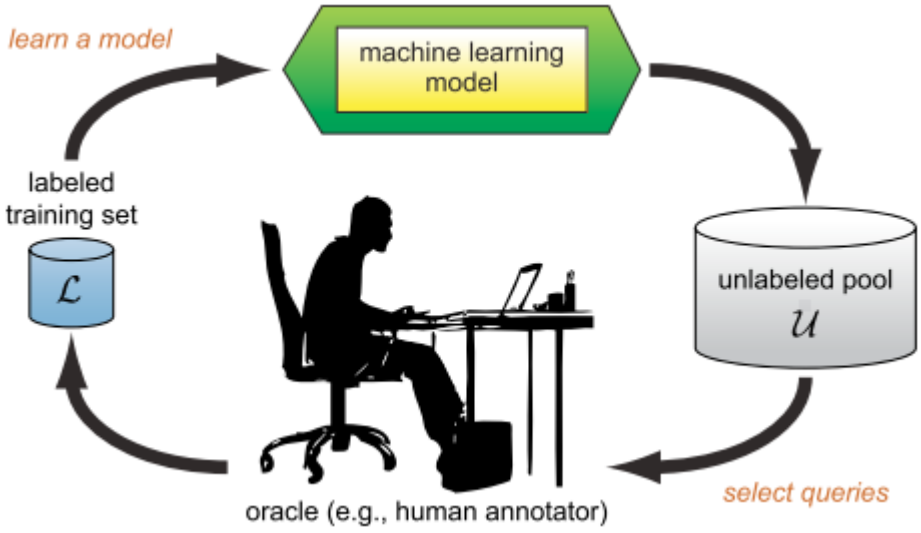

Figura 3.8 – Pool-based Sampling [10]

O ponto chave da abordagem de *active learning* é a identificação e seleção da instância que será classificada pelo oráculo. Para identificar esta instância dentro de um grande conjunto de instâncias não classificadas, é necessário determinar as instâncias mais informativas. Algumas estratégias foram propostas na literatura, e dentre elas a estratégia conhecida como *query-by-committee* [39].

A abordagem *query-by-committee* seleciona iterativamente instâncias não classificadas utilizando um comitê de classificadores, que representa diferentes hipóteses de classificação [39], gerados a partir do conjunto de treinamento corrente [40]. Cada membro do comitê vota em uma das possíveis classes, e as instâncias consideradas mais informativas são aquelas em que ocorre o maior desacordo entre os membros [10].

Para implementar um algoritmo de seleção baseado na abordagem *query-by-committee,* é necessário construir um comitê de classificadores que representem diferentes hipóteses consistentes com o conjunto de treinamento e utilizar alguma medida que permita mensurar o desacordo entre os membros do comitê [10].

Dentre as abordagens para mensurar o desacordo, está a apresentada por Dagan e Engelson [41] chamada *vote entropy*. *Vote entropy* foi originalmente aplicada na área de *part-of-speech*, com o objetivo de determinar o desacordo entre os classificadores quanto à atribuição de uma *tag* a uma palavra de uma sentença. A expressão seguinte apresenta o cálculo da medida *vote entropy.*

$$VE_{(w)} = -\sum_{t} \frac{V(t,w)}{k} \log \frac{V(t,w)}{k}$$

em que *V(t,w)* é o número de membros do comitê que votam na *tag t* para a palavra *w*, e *k* corresponde ao número de membros do comitê [41].

4 Alinhamento Interativo baseado em Query-by-Committee

Neste capítulo será descrita a abordagem proposta para alinhamento de ontologias. Para ilustrar a solução, será apresentado um exemplo utilizando um subconjunto de entidades das ontologias conference e cmt, apresentadas na introdução. Finalizando este capítulo, será apresentada a arquitetura do protótipo de sistema de alinhamento criado para implementar a proposta.

4.1 Abordagem Proposta

A abordagem proposta neste trabalho tem por objetivo inserir o usuário (especialista no domínio) no contexto do processo de alinhamento de ontologias, de forma efetiva, ou seja, fazendo com que sua participação resulte em alinhamentos de melhor qualidade.

Dentre as formas de envolvimento do usuário no processo de alinhamento apresentadas na literatura, foi escolhida aquela que o atribui o papel de fornecer *feedbacks* sobre pares de entidades relevantes. Esta forma se apresenta como a mais natural, por necessitar apenas do conhecimento do domínio, em comparação com outras formas que exijam conhecimento do processo de alinhamento, seus algoritmos e parâmetros.

A identificação e seleção dos pares de entidades para *feedback* do usuário é baseada na estratégia de seleção de instâncias informativas, da área de *Active Learning*, conhecida como *query-by-committee*. Para viabilizar a ligação entre a estratégia *query-by-committee* e o problema de alinhamento de ontologias com participação do usuário, foi proposto o processo de alinhamento de ontologias apresentado na Figura 4.1. Este processo é estruturado em duas fases:

- **Selecionar correspondências candidatas:** Nesta fase, são selecionados os pares de entidades das ontologias de entrada, considerando diferentes medidas de similaridade que exploram diferentes aspectos das ontologias, gerando correspondências candidatas.
- **Classificar correspondências candidatas:** Nesta fase, o alinhamento de ontologias é definido como um problema de classificação, o que permite determinar se correspondências candidatas identificadas na fase de seleção são de fato correspondentes ou não. *Feedbacks* sobre correspondências candidatas são solicitados ao usuário durante esta fase.

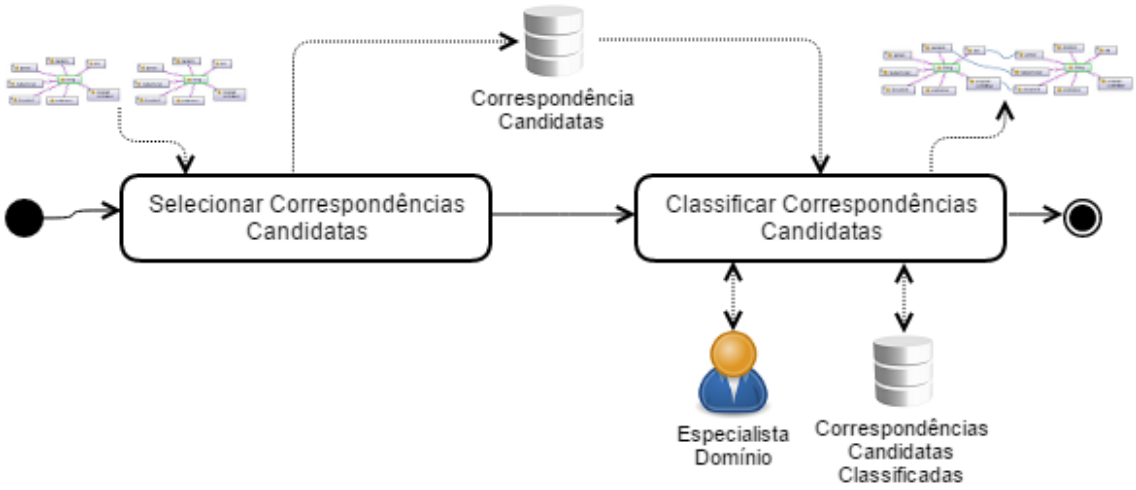

Figura 4.1 - Fases da abordagem proposta

Durante o processo de alinhamento, apresentado na Figura 4.1, os pares de entidades obtidos a partir das ontologias a serem alinhadas transitam por diferentes estados, que variam à medida que as atividades que compõem o processo proposto são executadas. A Figura 4.2 apresenta o ciclo de vida destes pares de entidades:

- **Correspondência candidata:** Um par de entidades torna-se uma correspondência candidata, quando, para uma dada medida de similaridade, existe algum indício de que este par de entidades seja uma possível correspondência. A transição de um par de entidades para o estado de correspondência candidata ocorre na fase de seleção de correspondências candidatas.
- **Correspondência candidata classificada:** Uma correspondência candidata torna-se uma correspondência candidata classificada, quando classificada automaticamente (com base na premissa da similaridade máxima) ou quando

classificada pelo usuário (através de um *feedback*), durante as iterações realizadas.

- **Correspondência:** Tanto correspondências candidatas classificadas quanto correspondências candidatas podem se tornar correspondências. A transição para este estado ocorre durante a geração do alinhamento. Uma correspondência candidata classificada torna-se uma correspondência, se esta tiver sido classificada como correspondente. No caso de uma correspondência candidata, a transição para o estado de correspondência, ocorre se ao fim do processo, a maioria dos membros do comitê indicar que esta correspondência candidata é uma correspondência.

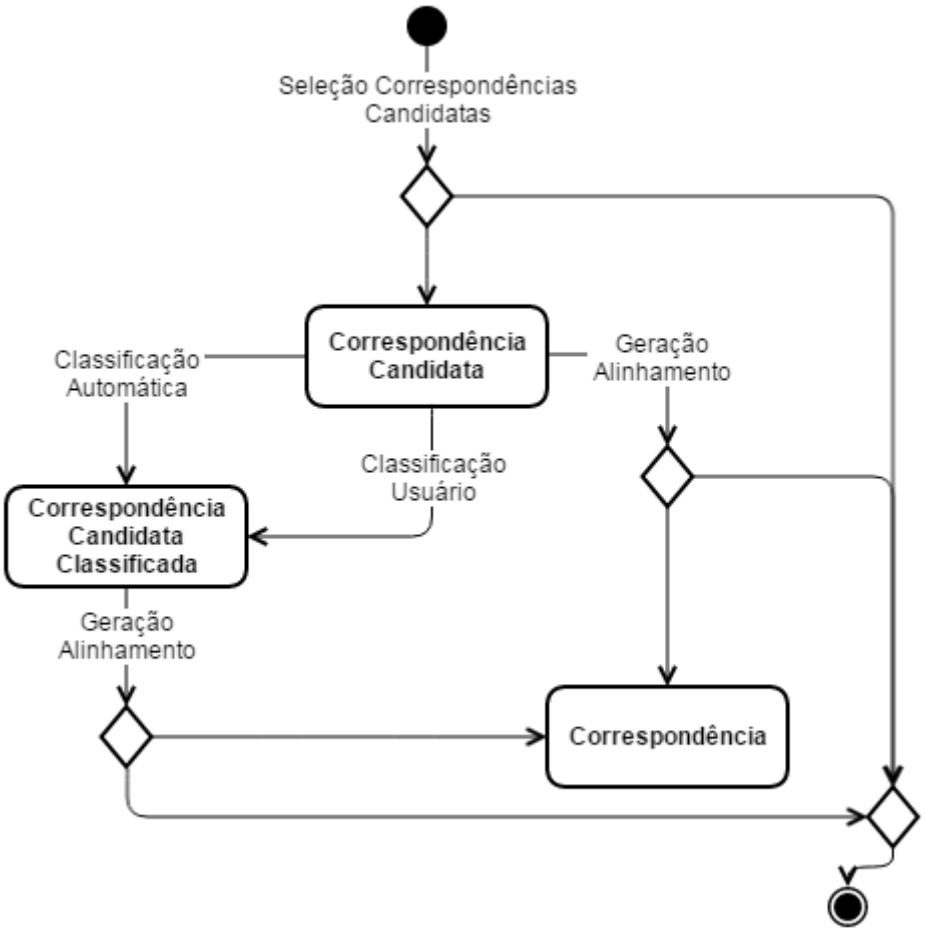

Figura 4.2 - Ciclo de vida de um par de entidades

Nas seções seguintes será discutida cada uma das atividades que compõem a solução proposta.

4.1.1 Selecionar Correspondências Candidatas

Pares de entidades de duas ontologias podem ser analisados sob diferentes perspectivas, explorando aspectos como nome, descrição, propriedades, relações que estabelecem com outras entidades e instâncias que as compõem. Cada uma destas perspectivas pode revelar um determinado nível de similaridade entre estas entidades. Estas perspectivas podem ser representadas por medidas que empregam diferentes métodos e recursos na análise da similaridade, conforme descrito no Capítulo 2.

A identificação de possíveis correspondências candidatas dá-se pela aplicação de medidas de similaridade. Dadas duas ontologias *O* e *O'* e suas respectivas entidades *e* e *e'*, pode-se representar as diversas combinações possíveis de entidades destas ontologias como um conjunto de *n*-uplas:

$$< id, e, e', m_1, m_2,..., m_j>$$

em que $m_1,...,m_j$ correspondem aos valores de similaridades obtidos a partir das *j* medidas aplicadas.

À medida que o número de entidades das ontologias a serem alinhadas cresce, o conjunto de possíveis pares de entidades cresce também. Somado a isto, pode-se observar que grande parte dos pares de entidades formados tem baixa similaridade e consequentemente não são correspondentes. Com o objetivo de reduzir o espaço de busca de pares de entidades para possíveis solicitações de *feedback* ao usuário, foi definida a atividade de seleção de correspondências candidatas. Esta atividade é centrada na aplicação do Algoritmo I apresentado na Figura 4.3.

O Algoritmo I seleciona um subconjunto de *n*-uplas considerando a perspectiva de cada medida de similaridade isoladamente. Ao final, o conjunto de correspondências candidatas resultante da atividade de seleção de correspondências candidatas é a união dos subconjuntos selecionados para cada medida. As correspondências candidatas podem ser entendidas como os pares de entidades que são candidatos a correspondência, segundo pelo menos uma das perspectivas de similaridade empregadas no processo. As *n*-uplas que não são selecionadas por nenhuma das medidas são descartadas, não sendo utilizadas na etapa seguinte.

Algoritmo I – Seleção de *n-uplas* por medida de similaridade

Entrada: $S_{(n\text{-}uplas)}$: Conjunto de *n*-uplas formadas por entidades de *O* e *O'*
m_x: Medida de similaridade a ser avaliada
Saída: $S_{(n\text{-}uplas\ selecionadas)}$: Conjunto de *n*-uplas selecionadas para a medida m_x
para cada ***e*** de ***O***
e' ← obter entidade ***e'*** de maior similaridade para m_x dado: $S_{(n\text{-}uplas)}$, ***e***
e'' ← obter entidade ***e*** de maior similaridade para ***m*** dado: $S_{(n\text{-}uplas)}$, ***e'***
se (***e*** = ***e''***) então
Adicionar a $S_{(n\text{-}uplas\ selecionadas)}$ a *n-upla* formada por ***e***, ***e'***
fim se
fim para

Figura 4.3 - Algoritmo de Seleção de n-uplas por medida de similaridade

O Algoritmo I realiza uma análise bidirecional sobre o conjunto de *n*-uplas, para uma medida de similaridade m_x ($1 <= x <= j$) fornecida como entrada. Considerando a medida m_x, esta análise ocorre da seguinte maneira: Para cada entidade *e* da ontologia *O*, é obtida a entidade *e'* da ontologia *O'* com quem *e* forma o par de maior valor de similaridade para a medida m_x, no conjunto de *n*-uplas. Dada a entidade *e'* retornada, é obtida a entidade *e''* da ontologia *O*, com que *e'* forma o par de maior valor de similaridade para a medida m_x, no conjunto de *n*-uplas. Se *e* e *e''* representarem a mesma entidade na ontologia *O*, então a *n*-upla que contém *e* e *e'* é selecionada como correspondência candidata na perspectiva de m_x.

A Tabela 4.1 apresenta um subconjunto das *n*-uplas formadas a partir das ontologias *cmt* e *conference*, e os valores obtidos aplicando-se três medidas de similaridade (m_1, m_2 e m_3) para cada par de entidades. Os pares de entidades deste subconjunto serão utilizados no decorrer deste capítulo para exemplificar a abordagem proposta e foram selecionados de um conjunto formado por 4.145 pares de entidades por questões de espaço.

Tabela 4.1 – Subconjunto de n-uplas das ontologias cmt e conference

id	e	e'	m_1	m_2	m_3
01	Author	Person	0,45	0,89	0,11
02	Author	Paper	0,09	0,38	0,08

03	Author	Abstract	0,00	0,00	0,31
04	Author	Regular author	0,50	0,50	0,36
05	Author	Topic	0,00	0,25	0,00
06	Author	Program Committee	0,00	0,22	0,10
07	Author	Chair	0,19	0,48	0,06
08	Chairman	Person	0,34	0,84	0,17
09	Chairman	Paper	0,08	0,35	0,12
10	Chairman	Abstract	0,00	0,00	0,13
11	Chairman	Regular author	0,12	0,38	0,12
12	Chairman	Topic	0,00	0,24	0,08
13	Chairman	Program Committee	0,00	0,21	0,12
14	Chairman	Chair	1,00	1,00	0,62
15	Co author	Person	0,22	0,61	0,07
16	Co author	Paper	0,05	0,50	0,15
17	Co author	Abstract	0,00	0,00	0,00
18	Co author	Regular author	0,33	0,62	0,43
19	Co author	Topic	0,00	0,25	0,13
20	Co author	Program Committee	0,00	0,22	0,18
21	Co author	Chair	0,09	0,38	0,28
22	Paper	Person	0,14	0,43	0,31
23	Paper	Paper	1,00	1,00	1,00
24	Paper	Abstract	0,00	0,00	0,06
25	Paper	Regular author	0,05	0,19	0,10
26	Paper	Topic	0,00	0,33	0,20
27	Paper	Program Committee	0,00	0,28	0,15
28	Paper	Chair	0,08	0,35	0,07
29	Paper Abstract	Person	0,07	0,33	0,29
30	Paper Abstract	Paper	0,50	0,63	0,36
31	Paper Abstract	Abstract	0,00	0,00	0,51
32	Paper Abstract	Regular author	0,03	0,20	0,12
33	Paper Abstract	Topic	0,06	0,37	0,07
34	Paper Abstract	Program Committee	0,18	0,23	0,19
35	Paper Abstract	Chair	0,04	0,28	0,07
36	Person	Person	1,00	1,00	1,00
37	Person	Paper	0,14	0,43	0,31
38	Person	Abstract	0,00	0,00	0,13
39	Person	Regular author	0,22	0,44	0,13
40	Person	Topic	0,00	0,29	0,00

41	Person	Program Committee	0,00	0,24	0,17
42	Person	Chair	0,28	0,53	0,00
43	Subject Area	Person	0,16	0,41	0,08
44	Subject Area	Paper	0,05	0,40	0,14
45	Subject Area	Abstract	0,00	0,00	0,22
46	Subject Area	Regular author	0,07	0,24	0,14
47	Subject Area	Topic	0,50	0,40	0,08
48	Subject Area	Program Committee	0,06	0,42	0,10
49	Subject Area	Chair	0,09	0,34	0,14

Aplicando o Algoritmo I e considerando a medida m_1, para a entidade *Author*, na ontologia *O*, é obtida a entidade *Regular author*, na Ontologia *O'* (*id = 04*), pois *Regular author* é a entidade com quem *Author* forma o par de maior valor de similaridade para a medida m_1, no caso 0,50. O mesmo vale para o inverso, ou seja, dada a entidade *Regular author*, na ontologia *O'*, é obtida a entidade *Author*, na ontologia *O*, pois *Author* é a entidade com quem *Regular author* forma o par de maior valor de similaridade para a medida m_1. Neste caso, podemos dizer que há simetria entre as entidades (*Author* e *Regular author*), portanto, a *n*-upla <04, *Author*, *Regular author*, 0,50, 0,50, 0,36> é selecionada como correspondência candidata segundo a medida m_1.

As Tabelas 4.2, 4.3 e 4.4 apresentam os subconjuntos de *n*-uplas selecionados da Tabela 4.1 pelo Algoritmo I, para as medidas de similaridade m_1, m_2 e m_3, respectivamente.

Tabela 4.2 – *n*-uplas selecionadas a partir de m_1

id	**e**	**e'**	**m_1**	**m_2**	**m_3**
04	Author	Regular author	0,50	0,50	0,36
14	Chairman	Chair	1,00	1,00	0,62
23	Paper	Paper	1,00	1,00	1,00
36	Person	Person	1,00	1,00	1,00
47	Subject Area	Topic	0,50	0,40	0,08

Tabela 4.3 - *n*-uplas selecionadas a partir de m_2

id	**e**	**e'**	**m_1**	**m_2**	**m_3**
14	Chairman	Chair	1,00	1,00	0,62

18	Co author	Regular author	0,33	0,62	0,43
23	Paper	Paper	1,00	1,00	1,00
36	Person	Person	1,00	1,00	1,00
48	Subject Area	Program Committee	0,06	0,42	0,10

Tabela 4.4 – *n*-uplas selecionadas a partir de m_3

id	**e**	**e'**	**m_1**	**m_2**	**m_3**
14	Chairman	Chair	1,00	1,00	0,62
18	Co author	Regular author	0,33	0,62	0,43
23	Paper	Paper	1,00	1,00	1,00
31	Paper Abstract	Abstract	0,00	0,00	0,51
36	Person	Person	1,00	1,00	1,00

A Tabela 4.5 apresenta o conjunto resultante de correspondências candidatas, considerando as três medidas de similaridade utilizadas no exemplo.

Tabela 4.5 – Correspondências candidatas

id	**e**	**e'**	**m_1**	**m_2**	**m_3**
04	Author	Regular author	0,50	0,50	0,36
14	Chairman	Chair	1,00	1,00	0,62
18	Co author	Regular author	0,33	0,62	0,43
23	Paper	Paper	1,00	1,00	1,00
31	Paper Abstract	Abstract	0,00	0,00	0,51
36	Person	Person	1,00	1,00	1,00
47	Subject Area	Topic	0,50	0,40	0,08
48	Subject Area	Program Committee	0,06	0,42	0,10

4.1.2 Classificar Correspondências Candidatas

Esta etapa do método proposto tem como objetivo determinar se as correspondências candidatas, identificadas na fase anterior, são de fato correspondentes ou não.

Ichise [9] descreve o processo de alinhamento de ontologias como um problema de classificação [23], em que cada par de entidades possui um atributo de classe, que indica se este par é correspondente ou não. A classe do par de entidades é determinada por um classificador, utilizando para isto os valores de similaridade obtidos a partir das medidas

aplicadas no processo. Para gerar este classificador, é necessário que exista um conjunto inicial de pares de entidades com o atributo de classe corretamente determinado e com valores de similaridade obtidos empregando as mesmas medidas de similaridade dos pares a serem classificados. É neste contexto do alinhamento de ontologias como um problema de classificação, que esta fase está definida.

O objetivo principal deste trabalho é envolver o usuário de forma a explorar o seu conhecimento a respeito do domínio representado pelas ontologias a serem alinhadas. Contudo, esta participação deve ser pensada de forma que o seu efeito reflita de maneira positiva na qualidade do alinhamento gerado. Este objetivo pode ser alcançado, no contexto da área de estudo de aprendizado de máquina, aplicando a abordagem conhecida como *active learning* [10].

No contexto do *active learning*, o aumento do efeito do *feedback* do usuário está diretamente relacionado ao nível de informação de uma instância, neste trabalho, representada por uma correspondência candidata. Quanto mais informativa uma instância é, maior é o aprendizado gerado pela sua classificação e, consequentemente, o efeito provocado por ela. Dentre as principais estratégias para seleção de instâncias informativas está a conhecida como *query-by-committee*.

Query-by-committee é baseada na ideia de um comitê de classificadores que geram hipóteses de classificação sobre as instâncias. As instâncias em que ocorre o maior desacordo entre os classificadores são consideradas as mais informativas, sendo, portanto selecionadas para classificação do usuário.

A estratégia de seleção de instâncias informativas apresentada por *query-by-committee* é a base da abordagem proposta. Estruturalmente, esta fase está organizada em um ciclo iterativo, em que a cada iteração são solicitados *feedbacks* ao usuário sobre correspondências candidatas e hipóteses de classificação são geradas. A Figura 4.4 apresenta o fluxo de atividades nesta fase.

A primeira iteração possui uma atividade de inicialização do repositório de correspondências candidatas classificadas, em que amostras do conjunto de correspondências candidatas são selecionadas para classificação do usuário. Já na última iteração é executada a atividade de geração do alinhamento. Cada uma das atividades que compõem esta fase é apresentada nas seções seguintes.

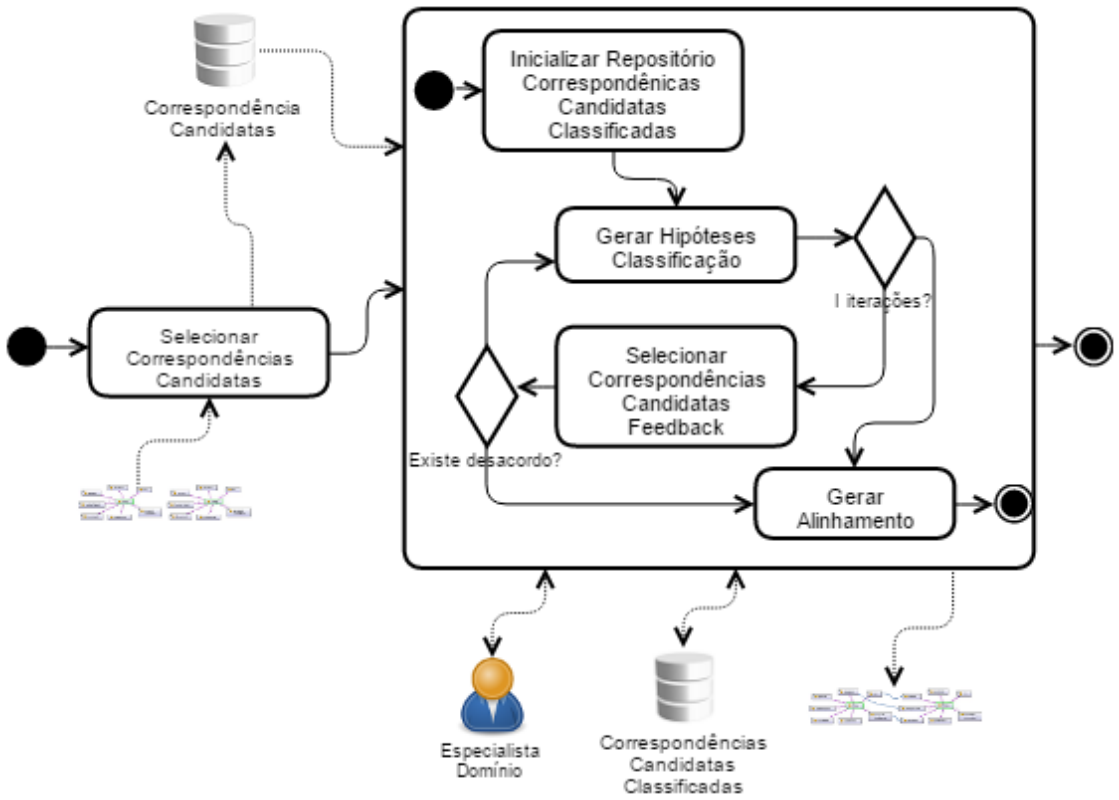

Figura 4.4 – Fluxo de atividade da fase de classificação

4.1.2.1 Inicializar Repositório de Correspondências Candidatas Classificadas

Para treinar o comitê de classificadores é necessário um conjunto inicial de instâncias classificadas, na abordagem proposta representado pelo repositório de correspondências candidatas classificadas.

Uma correspondência candidata classificada é representada por uma *n*-upla com a mesma estrutura da *n*-upla de uma correspondência candidata, acrescentando-se o atributo *classe*:

$$< id, e, e', m_1, m_2,..., m_j, classe>$$

em que o atributo *classe* indica se o par de entidades (*e*, *e'*) é ('SIM') ou não é ('NÃO') uma correspondência.

Para inicializar o repositório de correspondências candidatas classificadas, a abordagem proposta emprega duas estratégias de classificação:

Classificação automática segundo a premissa da similaridade máxima. Algumas correspondências candidatas são automaticamente classificadas, levando em conta a

seguinte premissa: *Quando um par de entidades é analisado sob diferentes perspectivas através de medidas de similaridade e todas elas retornam o valor de similaridade máximo para este par, então ele é considerado correspondente*.

Assim sendo, todas as correspondências candidatas que se adequam a premissa acima são classificadas automaticamente e passam a integrar o repositório de correspondências candidatas classificadas, com seu atributo de classe recebendo o valor 'SIM' e deixando, portanto, de fazer parte do repositório de correspondências candidatas.

A Tabela 4.6 apresenta o repositório de correspondências candidatas classificadas, atualizado com as correspondências candidatas da Tabela 4.5 que foram automaticamente classificadas. A tabela 4.7 apresenta o repositório de correspondências candidatas atualizado, ou seja, removendo-se as *n*-uplas já classificadas automaticamente.

Tabela 4.6 – Correspondências candidatas classificadas automaticamente

id	**e**	**e'**	**m_1**	**m_2**	**m_3**	**Classe**
23	Paper	Paper	1,00	1,00	1,00	SIM
36	Person	Person	1,00	1,00	1,00	SIM

Tabela 4.7 – Repositório de correspondências candidatas atualizado, após a classificação automática

id	**e**	**e'**	**m_1**	**m_2**	**m_3**
04	Author	Regular author	0,50	0,50	0,36
14	Chairman	Chair	1,00	1,00	0,62
18	Co author	Regular author	0,33	0,62	0,43
31	Paper Abstract	Abstract	0,00	0,00	0,51
47	Subject Area	Topic	0,50	0,40	0,08
48	Subject Area	Program Committee	0,06	0,42	0,10

Classificação segundo o feedback do usuário. Uma vez classificadas as correspondências candidatas que atendem a premissa da similaridade máxima, a abordagem prossegue selecionando amostras do repositório de correspondências candidatas, com o objetivo de inicializar o repositório de correspondências candidatas classificadas com um número mínimo de instâncias que permitam treinar os

classificadores. As amostras selecionadas devem ser diversificadas e capazes de representar o universo de faixas de valores de similaridade.

A seleção de amostras emprega o algoritmo Farthest First [38], em que as correspondências candidatas são consideradas pontos em um espaço *j*-dimensional (onde *j* é o número de medidas de similaridade) descritos pelos valores das medidas de similaridade. O algoritmo Farthest First seleciona *k* pontos, em que o primeiro é escolhido aleatoriamente, o segundo é o ponto mais distante do primeiro, o terceiro é o ponto mais distante do primeiro e do segundo, e assim sucessivamente até que se alcancem os *k* pontos.

Por exemplo, considerando as correspondências candidatas da Tabela 4.7 e *k* = 3, o algoritmo Farthest First obtém as correspondências candidatas destacadas na Figura 4.5, em que o *label* de cada nó representa o *id* da correspondência candidata.

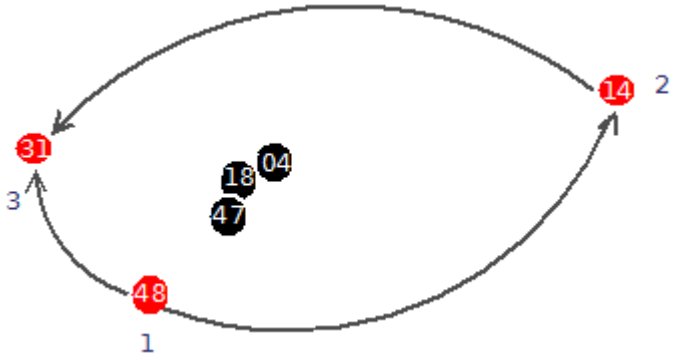

Figura 4.5 – Amostra de três correspondências candidatas da Tabela 4.7 selecionadas pelo algoritmo Farthest First

Considerando que inicialmente foi selecionada a correspondência candidata <48, *Subject Area*, *Program Committee*, 0,06, 0,42, 0,10>, o algoritmo busca o próximo ponto mais distante deste, no caso a correspondência candidata <14, *Chairman*, *Chair*, 1,00, 1,00, 0,62>. Posteriormente, é selecionado o ponto mais distante destes dois pontos, aqui representado pela correspondência candidata <31, *Paper abstract*, *Abstract*, 0,00, 0,00, 0,51>

Selecionados os *k* pontos (correspondências candidatas), é solicitado ao usuário que realize a classificação destes. As correspondências candidatas classificadas pelo usuário são então armazenadas no repositório de correspondências candidatas classificadas.

No exemplo, a Tabela 4.8 apresenta o repositório de correspondências candidatas classificadas atualizado após o *feedback* do usuário, enquanto a Tabela 4.9 apresenta o

repositório de correspondências candidatas atualizado, ou seja, removendo-se as *n*-uplas já classificadas pelo usuário.

Tabela 4.8 – Correspondências candidatas classificadas incluindo as amostras classificadas pelo usuário

id	e	e'	m_1	m_2	m_3	classe
14	Chairman	Chair	1,00	1,00	0,62	SIM
23	Paper	Paper	1,00	1,00	1,00	SIM
31	Paper Abstract	Abstract	0,00	0,00	0,51	SIM
36	Person	Person	1,00	1,00	1,00	SIM
48	Subject Area	Program Committee	0,06	0,42	0,10	NÃO

Tabela 4.9 – Repositório de correspondências candidatas atualizado, após a classificação pelo usuário.

id	e	e'	m_1	m_2	m_3
04	Author	Regular author	0,50	0,50	0,36
18	Co author	Regular author	0,33	0,62	0,43
47	Subject Area	Topic	0,50	0,40	0,08

4.1.2.2 Gerar Hipóteses de Classificação

Esta etapa é responsável por gerar hipóteses de classificação para as correspondências candidatas, através de um comitê de classificadores. Este comitê é o componente central da solução apresentada, pois é a partir dele que correspondências candidatas mais informativas são identificadas e o efeito do *feedback* do usuário é propagado. O repositório de correspondências candidatas classificadas pela etapa anterior representa o conjunto de instâncias de treinamento para os classificadores.

Com o objetivo de explorar ao máximo o conjunto de treinamento e também estimular o desacordo, que na abordagem *query-by-committe* é a chave para identificar instâncias informativas, a abordagem proposta define um comitê heterogêneo, formado por classificadores que empregam diferentes estratégias no processo de classificação de instâncias. A seleção destes classificadores foi realizada com base em trabalhos na área de alinhamento de ontologias, em que estes classificadores foram aplicados apresentando bons resultados [32] [31]. São eles:

- **Naive Bayes:** Este classificador fornece uma visão probabilística para a classificação considerando cada atributo de forma independente [33].
- **Random Forest:** Combinação de classificadores baseados em árvores de decisão, que fornecem uma visão baseada em regras [29].
- **Multilayer Perceptron:** Classificador baseado em redes neurais, em que o processo de aprendizado está associado a determinação dos pesos de cada neurônio [30].

Na abordagem proposta, a cada iteração, os classificadores que compõem o comitê são treinados utilizando o repositório de correspondências candidatas classificadas, e cada um deles gera hipóteses de classificação para cada uma das correspondências candidatas (ainda não classificadas). Nesta fase da abordagem, a *n*-upla de uma correspondência candidata é estendida para inclusão de atributos que representem as hipóteses de classificação do comitê, assumindo o formato:

$$< id, e, e', m_1, m_2,..., m_j, classe_{NB}, classe_{RF}, classe_{MP} >$$

em que os atributos $classe_{NB}$, $classe_{RF}$, $classe_{MP}$ representam a hipótese de classificação ('SIM' ou 'NÃO') dos classificadores Naive Bayes, Random Forest e Multilayer Perceptron, respectivamente.

A Tabela 4.10 apresenta as correspondências candidatas da Tabela 4.9 com as hipóteses de classificação geradas pelos membros do comitê.

Tabela 4.10 – Correspondências candidatas e hipóteses de classificação geradas pelo comitê

id	**e**	**e'**	**m_1**	**m_2**	**m_3**	**$classe_{NB}$**	**$classe_{RF}$**	**$classe_{MP}$**
04	Author	Regular author	0,50	0,50	0,36	SIM	SIM	SIM
18	Co author	Regular author	0,33	0,62	0,43	NÃO	NÃO	SIM
47	Subject Area	Topic	0,50	0,40	0,08	SIM	NÃO	SIM

4.1.2.3 Selecionar Correspondências Candidatas para Feedback

Medir o nível de informação de uma instância é um elemento chave quando se aplica *active learning*. Na estratégia *query-by-committe*, isto significa medir o desacordo entre os membros do comitê, pois quanto maior o desacordo entre os membros do comitê sobre a classificação de uma instância, mais informação esta instância carrega para o aprendizado.

Na abordagem proposta, os membros do comitê geram diferentes hipóteses de classificação para cada correspondência candidata, cada um seguindo uma técnica de classificação. Para mensurar o desacordo entre eles, e consequentemente o nível de informação de uma correspondência candidata, foi empregada a medida *vote entropy* proposta por Dagan e Engelson [41].

A ideia geral da medida *vote entropy* é que cada membro do comitê "vota", para cada uma das correspondências candidatas, em uma das classes ("SIM" ou "NÃO"). A *Vote entropy* para uma correspondência candidata *cc* ($Ve_{(cc)}$) é calculada através da expressão seguinte:

$$Ve_{(cc)} = -\sum_{l} \frac{V(l,cc)}{k} \log \frac{V(l,cc)}{k}$$

em que *V(l,cc)* retorna o número de votos para a classe *l* obtidos para a correspondência candidata *cc*, e *k* corresponde ao número de membros do comitê.

Contudo, Settles [39] destaca que considerar apenas o desacordo na determinação do nível de informação de uma instância pode ser insuficiente e levar a escolha de *outliers*. Para lidar com este problema, a abordagem proposta considera a vizinhança da correspondência candidata ao selecioná-la para *feedback*, calculando a distância da correspondência candidata selecionada em relação a todas as outras. Quanto menor for esta distância, maior o efeito do *feedback* do usuário.

Para calcular a distância de uma correspondência candidata em relação a todas as outras contidas no repositório de correspondências candidatas, é utilizada a distância euclidiana média $Dem_{(cc)}$, determinada pela expressão abaixo:

$$Dem_{(cc)} = \frac{1}{U} \sum_{i=1}^{U} \sqrt{(m_{1_{(cc)}} - m_{1_{(cc\,i)}})^2 + \cdots + (m_{j_{(cc)}} - m_{j_{(cc\,i)}})^2}$$

em que $m_1,\ldots,m_j$ correspondem aos valores de similaridade obtidos a partir das *j* medidas aplicadas no processo de alinhamento, *cc* é uma correspondência candidata e *U* é o número total de correspondências candidatas no repositório de mesmo nome. A Tabela 4.11 apresenta as medidas *Vote entropy* (*Ve*) e distância euclidiana média (*Dem*) das correspondências candidatas contidas na Tabela 4.10.

Tabela 4.11 – Vote entropy e distância euclidiana média das correspondências candidatas

id	m_1	m_2	m_3	$classe_{NB}$	$classe_{RF}$	$classe_{MP}$	Ve	Dem
04	0,50	0,50	0,36	SIM	SIM	SIM	0	0,26

18	0,33	0,62	0,43	NÃO	NÃO	SIM	0,28	0,33
47	0,50	0,40	0,08	SIM	NÃO	SIM	0,28	0,37

A cada iteração realizada nesta etapa, são selecionadas *w* correspondências candidatas para *feedback* do usuário, segundo os seguintes critérios, em ordem decrescente de prioridade:

1. Correspondências candidatas com maior valor para a medida *vote entropy*;
2. Correspondências candidatas com menor valor para a distância euclidiana média;

As *w* correspondências candidatas selecionadas são classificadas pelo usuário e então armazenadas no repositório de correspondências candidatas classificadas. No exemplo apresentado, com *w* = 1, a correspondência candidata <18, *Co author*, *Regular author*, 0,33, 0,62, 0,43> é selecionada para *feedback* do usuário. A Tabela 4.12 apresenta o repositório de correspondências candidatas classificadas atualizado com o novo par de entidades classificado.

Tabela 4.12 – Correspondências candidatas classificadas incluindo a mais informativa

id	**e**	**e'**	**m_1**	**m_2**	**m_3**	**classe**
14	Chairman	Chair	1,00	1,00	0,62	SIM
23	Paper	Paper	1,00	1,00	1,00	SIM
31	Paper Abstract	Abstract	0,00	0,00	0,51	SIM
36	Person	Person	1,00	1,00	1,00	SIM
48	Subject Area	Program Committee	0,06	0,42	0,10	NÃO
18	Co author	Regular author	0,33	0,62	0,43	NÃO

Tabela 4.13 – Repositório de correspondências candidatas atualizado

id	**e**	**e'**	**m_1**	**m_2**	**m_3**	**$classe_{NB}$**	**$classe_{RF}$**	**$classe_{MP}$**
04	Author	Regular author	0,50	0,50	0,36	SIM	SIM	SIM
47	Subject Area	Topic	0,50	0,40	0,08	SIM	NÃO	SIM

4.1.2.4 Gerar Alinhamento

As atividades, gerar hipóteses de classificação e selecionar correspondências candidatas para *feedback*, ocorrem em um ciclo que se repete por um determinado número de

iterações. O número de iterações executadas é variável podendo ser determinado de duas formas:

- **Manualmente:** O usuário indica o número de iterações a serem executadas. Este número nem sempre será alcançado, pois depende da existência de desacordo entre os membros do comitê;
- **Automaticamente:** O número de iterações é determinado dinamicamente, com as atividades sendo executadas enquanto houver desacordo entre os membros do comitê.

Após alcançar o número de iterações, o alinhamento entre as ontologias é gerado. As correspondências que compõem o alinhamento são formadas por pares de entidades do repositório de correspondências candidatas classificadas, com valor de classe = "SIM", e do repositório de correspondências candidatas, que obtiveram o maior número de votos para a hipótese de classificação "SIM".

No exemplo, supondo que o número de iterações tenha sido alcançado, o alinhamento gerado é composto pelas correspondências da Tabela 4.14.

Tabela 4.14 – Alinhamento entre o subconjunto de entidades das ontologias *cmt* e *conference*

id	**e**	**e'**	**Relação**
04	Author	Regular author	=
14	Chairman	Chair	=
23	Paper	Paper	=
31	Paper Abstract	Abstract	=
36	Person	Person	=
47	Subject Area	Topic	=

O algoritmo II apresentado na Figura 4.6 representa o fluxo de atividades da fase "Classificar Correspondências Candidatas", em que o número de iterações é definido manualmente pelo usuário.

Algoritmo II – Classificação de Correspondências Candidatas

Entrada: $S_{(cc)}$: Conjunto de correspondências candidatas

i: Número de iterações

Saída: $S_{(id, e1, e2, r)}$: Alinhamento entre as Ontologias

inicializar correspondências candidatas classificadas ($S_{(cc)}$)

gerar hipóteses de classificação ($S_{(cc)}$)

enquanto $i <> 0$

    selecionar correspondências candidatas para feedback do usuário ($S_{(cc)}$)

    gerar hipóteses de classificação ($S_{(cc)}$)

    decrementar (i)

fim enquanto

$S_{(id, e1, e2, r)} \leftarrow$ gerar alinhamento

Figura 4.6 - Algoritmo para a fase de classificação

4.2 Arquitetura JARVIS

Para avaliar a abordagem proposta, foi desenvolvido um protótipo de sistema de alinhamento de ontologias que implementa cada uma das atividades definidas na seção 4.1.

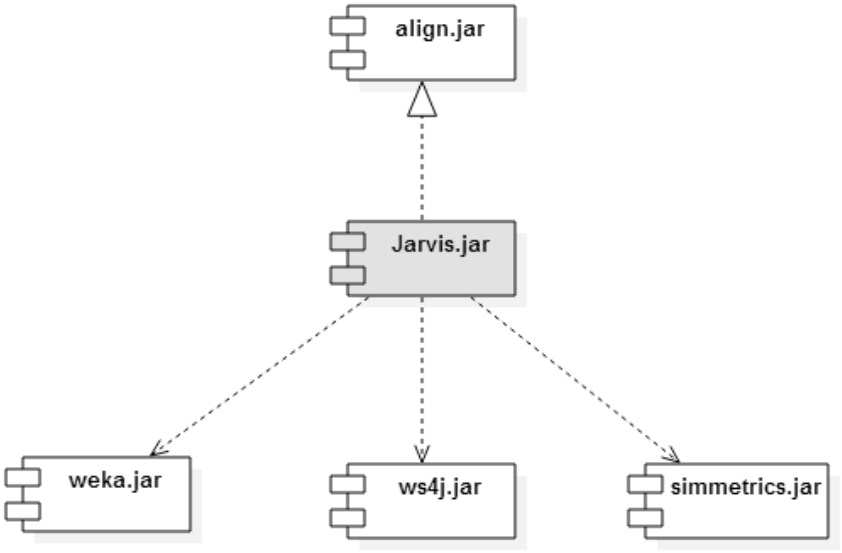

Figura 4.7 - Diagrama de componentes da arquitetura JARVIS

Este protótipo foi desenvolvido utilizando como base um conjunto de componentes, que fornecem as capacidades necessárias para o desenvolvimento da solução. Estes componentes são apresentados na Figura 4.7 e descritos a seguir:

- A Align API [42] fornece um arcabouço para o desenvolvimento de sistemas de alinhamento de ontologias oferecendo um conjunto de abstrações que permitem expressar, acessar e compartilhar alinhamentos. Esta API fornece uma estrutura mínima de processamento que possui como componentes mais importantes a *AlignmentProcess*, que é uma interface para implementação dos sistemas de alinhamento e *Evaluator*, que é a interface para avaliação dos alinhamentos gerados.
- A WEKA API (*Waikato Environment for Knowledge Analysis*) [43] permite acessar de maneira simples o estado da arte das técnicas de aprendizado de máquina, provendo a implementação de diversos algoritmos incluindo os de classificação e clustering.
- A SimMetrics API fornece um conjunto de implementações das principais medidas de similaridade baseadas em string disponíveis na literatura, incluindo as medidas de similaridade Jaccard, JaroWinkler e nGram utilizadas neste trabalho. Estas medidas foram selecionadas por apresentarem os melhores resultados em avaliações realizadas no trabalho [16].

- A WS4J API (*WordNet Similarity for Java*) fornece a implementação em java de um conjunto de medidas de similaridade semânticas, utilizando a *WordNet* como recurso linguístico, incluindo as medidas de similaridade WuPalmer, Lin e JiangConrath utilizadas neste trabalho. Estas medidas foram selecionadas por apresentarem os melhores resultados em avaliações realizadas no trabalho [15].

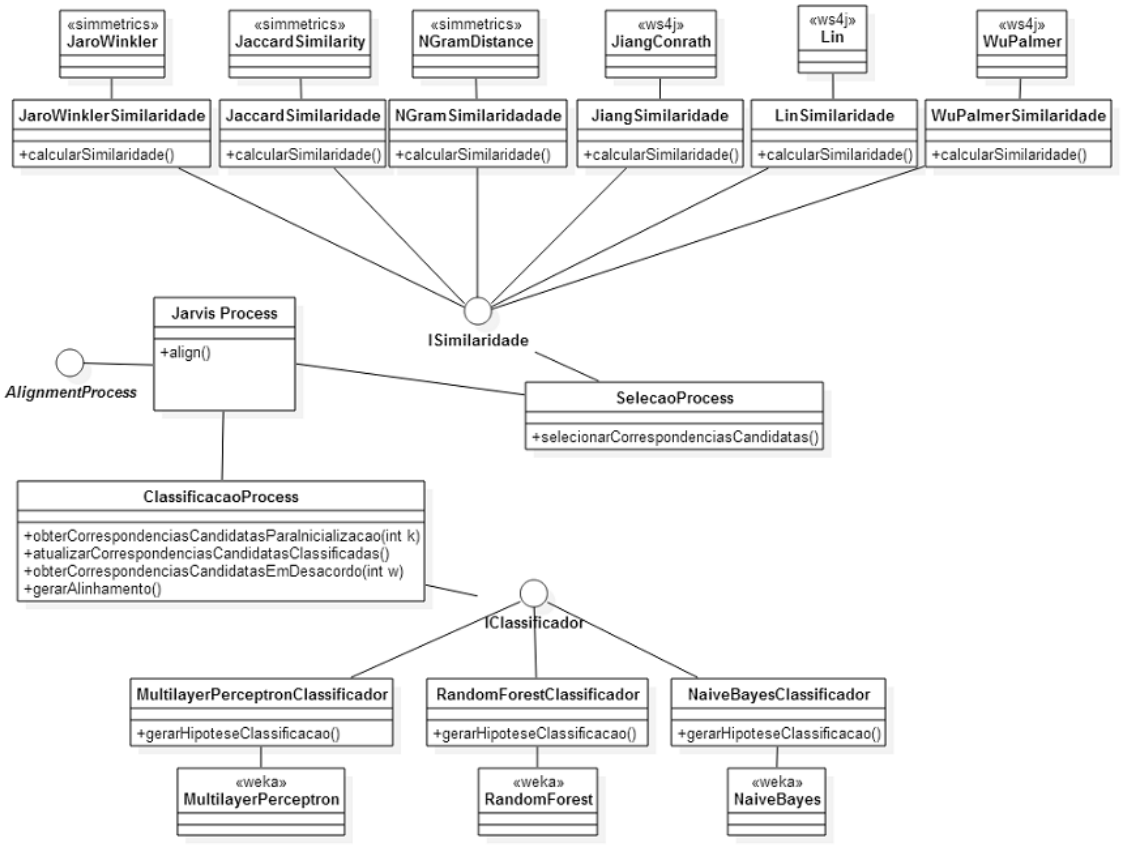

Figura 4.8 – Diagrama de classes do protótipo JARVIS

A Figura 4.8 apresenta o diagrama de classes do protótipo implementado. Esta visão mostra a integração da aplicação com os componentes apresentados anteriormente.

No capítulo seguinte a abordagem para alinhamento de ontologias apresentada será avaliada através de um conjunto de experimentos, que simulam diversos cenários de interação com o usuário.

5 Experimentos e Análise dos Resultados

Neste capítulo será apresentado o experimento realizado para avaliação da abordagem proposta, descrita no capítulo anterior, incluindo os cenários planejados, ontologias a serem alinhadas, e dados coletados durante a execução de cada um dos cenários avaliados. Ao final deste capítulo, será realizada a análise dos dados coletados, bem como a avaliação dos resultados considerando os sistemas de alinhamento avaliados na trilha de interactive matching.

5.1 Planejamento do Experimento

Na abordagem proposta, correspondências candidatas são classificadas pelo comitê e o desacordo é mensurado aplicando a medida *vote entropy*. A cada iteração realizada, o usuário é solicitado a classificar as *w* correspondências candidatas que possuem o maior valor para esta medida.

A variável *w* representa o número de instâncias informativas que serão selecionadas para classificação do usuário a cada iteração realizada, portanto, a análise do comportamento desta variável em diferentes cenários é um dos pontos principais a ser avaliado neste experimento, pois afeta diretamente os resultados obtidos pela abordagem proposta.

Considerando o objetivo de reduzir o esforço do usuário ao longo das iterações, aumentando o efeito da participação, é desejável que a variável *w* assuma valores baixos, de forma que a cada iteração realizada com o usuário sejam solicitados *feedbacks* sobre um número reduzido de pares de entidades.

A variável *k* representa o número de instâncias selecionadas, durante a primeira iteração, através do algoritmo Farthest First para classificação do usuário. Estas

instâncias farão parte do conjunto de treinamento utilizado para inicialização do comitê de classificadores. O valor de *k* influencia a capacidade dos membros do comitê em classificar as instâncias, aumentando ou reduzindo o desacordo entre eles. Este comportamento pode afetar diretamente os resultados alcançados pela abordagem proposta.

O experimento deve incluir cenários para avaliar o comportamento da abordagem proposta frente às variações de *k*. É desejável que esta variável assuma valores baixos pois como o objetivo é o aumento do efeito do *feedback*, então não se justifica o custo do usuário classificar uma grande quantidade de pares de entidades inicialmente.

Para realização do experimento, além das características a serem consideradas na especificação dos cenários de avaliação da abordagem proposta, outros elementos devem ser definidos, como as ontologias a serem alinhadas durante a execução dos cenários e os dados a serem coletados a cada iteração executada, que permitiram avaliar e comparar a abordagem proposta com o estado da arte dos sistemas de alinhamento.

As seções seguintes descrevem cada um dos elementos envolvidos no processo experimental, bem como os cenários planejados.

5.1.1 Ontologias

O primeiro insumo a ser definido para realização do experimento são as ontologias a serem utilizadas no processo de alinhamento. A *Ontology Alignment Evaluation Iniciative* – OAEI é uma iniciativa internacional coordenada que organiza a avaliação de um número crescente de sistemas de alinhamento de ontologias [44].

Anualmente a OAEI realiza campanhas com o objetivo de comparar sistemas e algoritmos e concluir sobre as melhores estratégias e soluções para alinhamento de ontologias. As campanhas realizadas são organizadas em trilhas, que representam diferentes modalidades de avaliação, e um conjunto de *data sets*. No ano de 2014 a campanha foi organizada em sete trilhas de avaliação [44].

Dentre as modalidades de avaliação definidas pela OAEI encontra-se a trilha de *interactive matching*. Esta trilha oferece a possibilidade de comparar diferentes soluções para alinhamento de ontologias que requerem a participação do usuário. O objetivo é verificar se a interação com o usuário pode contribuir para a melhoria da qualidade dos alinhamentos, identificar quais abordagens são mais promissoras e

contabilizar o número de interações necessárias para realização do processo. As abordagens participantes desta trilha são avaliadas no *Conference data set*. O *Conference data set* originalmente integra a trilha de ontologias expressivas, composta de *data sets* formados por ontologias do mundo real, modeladas em OWL. Para avaliação deste trabalho será utilizado o *Conference data set*, por se tratar do *data set* oficial para avaliação de abordagens interativas. Este *data set* tem suas características descritas na seção seguinte.

5.1.1.1 Conference

O *Conference data set* é composto de dezesseis ontologias que descrevem o domínio da realização de conferências. Estas ontologias são adequadas para avaliação do processo de alinhamento devido ao seu caráter heterogêneo [44].

Das dezesseis ontologias que compõem este *data set*, são disponibilizados vinte e um alinhamentos de referência, que correspondem ao alinhamento completo da combinação de sete destas ontologias: *Cmt*, *Conference*, *ConfOf*, *Edas*, *Ekaw*, *Iasted* e *Sigkdd*. Suas características são apresentadas na Tabela 5.1.

Para a avaliação são consideradas duas variações do alinhamento de referência, ra1 e ra2. O ra1 é o alinhamento de referência original e disponível para uso. O ra2 é derivado do ra1, em que possíveis correspondências conflitantes foram inspecionadas e resolvidas por avaliadores, com o objetivo de se obter alinhamentos mais coerentes. O ra2 não está disponível para uso, sendo exclusivo para avaliações realizadas pela OAEI.

Tabela 5.1 – Ontologias do Conference data det

Ontologia	Número de classes	Número de data properties	Número de object properties
Cmt	36	10	49
Conference	60	18	46
ConfOf	38	23	13
Edas	104	20	30
Ekaw	74	0	33
Iasted	140	3	38
Sigkdd	49	11	17

Na avaliação realizada em 2014 para a trilha de *interactive matching*, foi empregado o *Conference data set* utilizando o alinhamento de referência ra1 [44]. Para a avaliação deste trabalho será utilizado o mesmo *data set* e também o mesmo alinhamento de referência empregado pela OAEI, permitindo que a abordagem proposta seja comparada com o estado da arte dos sistemas de alinhamento de ontologias avaliados pela trilha de *interactive matching*.

5.1.2 Questões e Cenários

Nesta seção são definidas as questões a serem respondidas por este experimento de avaliação, e o conjunto de cenários de execução avaliados.

Conforme apresentado no início deste capítulo, a solução proposta pode variar seu comportamento e consequentemente os resultados obtidos dependo do valor assumido por duas variáveis que apresentam funções distintas:

Variável k: Na atividade de inicialização do repositório de correspondências candidatas classificadas, é solicitado ao usuário a classificação das *k* correspondências candidatas selecionadas pelo algoritmo Farthest First. Estas *k* correspondências, uma vez classificadas, irão compor o repositório de correspondências candidatas classificadas, que é a base para treinamento do comitê de classificadores.
Variável w: Na atividade de seleção de correspondências para *feedback*, é solicitado ao usuário a classificação das *w* correspondências candidatas de maior *vote entropy* e menor distância euclidiana média. Estas *w* correspondências, uma vez classificadas, serão armazenadas no repositório de correspondências candidatas classificadas e participarão da geração do comitê de classificadores na iteração seguinte.

Considerando o papel desempenhado pelas variáveis *k* e *w* na abordagem proposta, foram definidas duas questões a serem respondidas durante a realização do experimento de avaliação:

Questão 1: *Qual o efeito da variação no valor assumido pela variável k na qualidade dos alinhamentos gerados (precisão, cobertura e medida-F)?*

Questão 2: *Qual o efeito da variação no valor assumido pela variável w na qualidade dos alinhamentos gerados (precisão, cobertura e medida-F)?*

Estas questões visam a verificação do comportamento da abordagem proposta frente às variações nos valores de *k* e *w*. Para responder estas questões foram especificados nove cenários de execução, conforme apresentados na Tabela 5.2. Esta tabela apresenta a relação entre as questões e os cenários de execução, bem como o número máximo de iterações em cada cenário, os valores assumidos pelas variáveis *k* e *w* e o número máximo de feedbacks que poderão ser solicitados ao usuário durante a execução.

Ao definir o número de iterações, optou-se pela opção manual, em que o usuário indica o número máximo de iterações a serem executadas. Contudo, este número nem sempre será alcançado, pois dependerá da existência de desacordo entre os membros do comitê. Na primeira iteração de cada cenário o usuário será solicitado a classificar até *k* + *w* pares de entidades, já nas iterações seguintes ele será solicitado a classificar até *w* pares de entidades.

Tabela 5.2 – Cenários de execução

Questões	**Cenários**	**Iterações**	***k***	***w***	**Número máximo Feedbacks**
Questão 1	Cenário 1	5	4	1	9
	Cenário 2	5	5	1	10
	Cenário 3	5	6	1	11
	Cenário 4	5	7	1	12
	Cenário 5	5	8	1	13
Questão 2	Cenário 1	5	4	1	9
	Cenário 6	5	4	2	14
	Cenário 7	5	4	3	19
	Cenário 5	5	8	1	13
	Cenário 8	5	8	2	18
	Cenário 9	5	8	3	23

O objetivo principal da abordagem proposta é a melhoria da qualidade dos alinhamentos gerados, envolvendo o usuário neste processo e, ao mesmo tempo, reduzindo o esforço, ao fornecer *feedbacks*, e aumentando o efeito destes *feedbacks* ao longo das iterações. Para verificar o comportamento da abordagem frente a este objetivo, as variáveis *k* e *w* foram configuradas com valores baixos baseados no número de pares de entidades correspondentes, de acordo com o alinhamento de referência.

5.1.3 Coleta de Dados

A escolha das medidas a serem empregadas na avaliação dos alinhamentos gerados pela abordagem proposta foi realizada considerando o estado da arte da área de alinhamento de ontologias e as avaliações realizadas pela OAEI em suas campanhas anuais.

As principais medidas empregadas com o objetivo de avaliar a qualidade dos alinhamentos focam na verificação da conformidade do alinhamento gerado por alguma abordagem em relação a um padrão (alinhamento de referência). As medidas mais utilizadas com este objetivo são: precisão, cobertura e medida-F.

O processo de coleta de dados para avaliação experimental da solução proposta seguirá as seguintes definições:

- Serão executados nove cenários de avaliação;
- Em cada cenário serão alinhadas todas as vinte e uma combinações de ontologias do *Conference data set*;
- A cada iteração realizada, o alinhamento será gerado e avaliado utilizando o respectivo alinhamento de referência. Nesta avaliação serão coletados os valores das medidas de precisão (*P*), cobertura (*C*) e medida-F (*F*).

5.2 Análise de Dados

Nesta seção serão apresentados os resultados obtidos pela abordagem proposta em cada cenário de execução. Os resultados dos cenários serão agrupados por questão definida e serão apresentados de forma consolidada, por iteração.

Questão 1: *Qual o efeito da variação no valor assumido pela variável k na qualidade dos alinhamentos gerados (precisão, cobertura e medida-F)?*

Para responder a esta questão foram executados os cenários 1, 2, 3, 4 e 5, em que são atribuídos diferentes valores para a variável *k* com o valor da variável *w* fixa. Este conjunto de cenários tem por objetivo verificar o comportamento da abordagem proposta frente à variação no número de pares de entidades selecionados através do algoritmo Farthest First, e que uma vez classificados pelo usuário, serão utilizados para a inicialização do comitê de classificadores.

Os resultados das execuções dos cenários 1, 2, 3, 4 e 5, para cada um dos vinte e um pares de ontologias experimentados, são apresentados nas Tabelas 0.1, 0.2, 0.3, 0.4 e 0.5 respectivamente e estão disponíveis para consulta no apêndice deste trabalho.

O gráfico da Figura 5.1 apresenta a variação da medida-F nos cenários 1, 2, 3, 4 e 5. Os resultados apresentados foram consolidados por cenário e iteração, e determinados calculando o valor médio da medida-F obtida a partir de cada um dos vinte e um pares de ontologias avaliados.

Analisando o gráfico da Figura 5.1 é possível verificar a evolução da medida-F a cada iteração. Considerando a primeira iteração dos cenários avaliados, é possível observar um aumento no valor da medida-F ao passo que a variável *k* é incrementada. A exceção é o cenário 2, em que ocorre redução neste valor quando comparado com o cenário 1.

Este tipo de comportamento (redução no valor da medida-F com o incremento da variável *k*) é ocasionado por um *feedback* que não segue o padrão até então existente nas instâncias de treinamento. Isto ocorre, por exemplo, quando um par de entidades possui valores de similaridade altos e o usuário o classifica como não correspondente, ou quando possui valores de similaridade baixos e o usuário o classifica como correspondente. Cada membro do comitê tem um comportamento diferente (comitê heterogêneo) para este tipo de *feedback*, o que geralmente afeta o resultado obtido.

Comparando a primeira iteração dos cenários 1 e 5 é possível perceber um incremento de 0,02 no valor da medida-F, com a variável *k* assumindo os valores 4 e 8 respectivamente. Este mesmo incremento é observado na quinta iteração destes mesmos cenários.

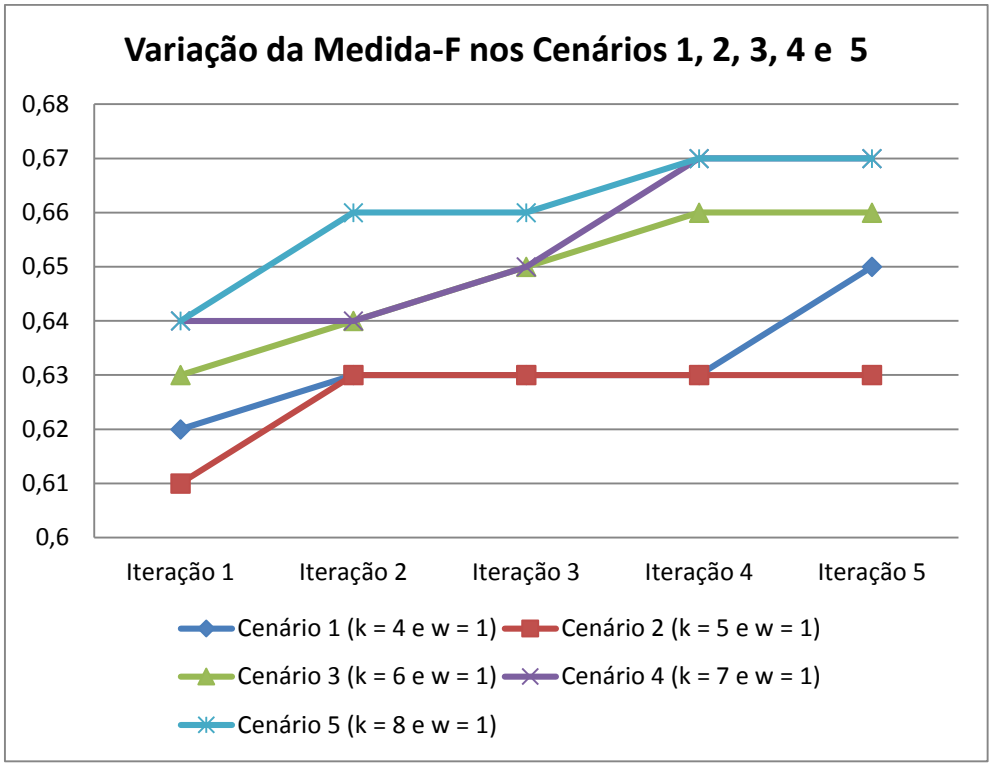

Figura 5.1 – Variação da Medida-F média para os cenários 1, 2, 3, 4 e 5

Os resultados obtidos pela abordagem proposta em termos de precisão e cobertura para os cenários 1, 2, 3, 4 e 5 são apresentados nos gráficos da Figura 5.2 e 5.3 respectivamente. Os resultados apresentados foram consolidados por cenário e iteração, e determinados calculando o valor médio das medidas de precisão e cobertura obtidas a partir de cada um dos vinte e um pares de ontologias avaliados.

Considerando os resultados em termos de precisão, é possível observar a evolução desta medida ao longo das iterações. Analisando a primeira iteração dos cenários avaliados é possível observar um aumento no valor da precisão ao passo que a variável *k* é incrementada. A exceção é o cenário 2, em que ocorre redução neste valor quando comparado com o cenário 1.

Analisando a quarta iteração do cenário 1, é possível observar uma redução no valor da medida-F quando comparada com a terceira iteração. Este comportamento, conforme explicado anteriormente, é ocasionado por um *feedback* que não segue o padrão até então existente nas instâncias de treinamento. Contudo, é importante destacar a capacidade de recuperação da abordagem proposta, pois na quinta iteração, deste mesmo cenário, com apenas um *feedback*, o resultado obtido foi superior ao da terceira

iteração, o que demonstra a capacidade de propagação do efeito do *feedback* para outros pares de entidades.

Comparando a primeira iteração dos cenários 1 e 5 é possível perceber um incremento de 0,02 no valor da precisão, com a variável *k* assumindo os valores 4 e 8 respectivamente. Este mesmo incremento é observado na quinta iteração destes mesmos cenários. Analisando os cenários 1 e 5 individualmente, pode-se perceber um incremento de 0,06 nos valores de precisão, considerando a primeira e quinta iterações.

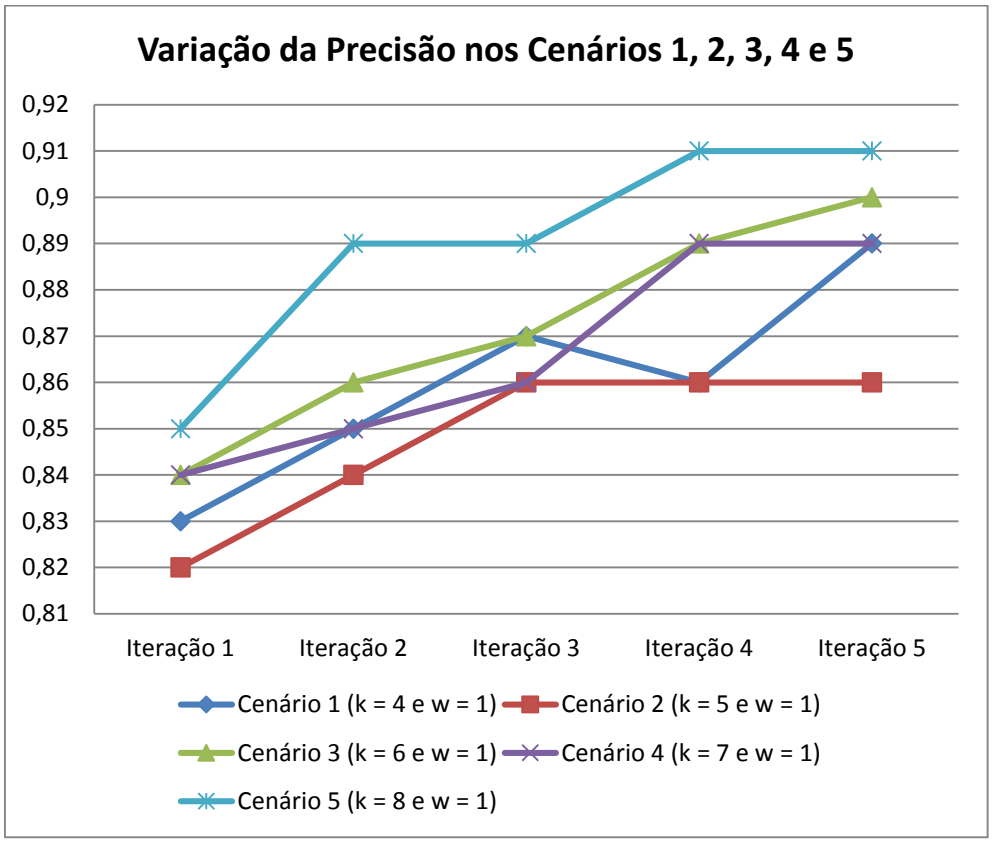

Figura 5.2 – Variação da Precisão média para os cenários 1, 2, 3, 4 e 5

Ao analisar os resultados da abordagem proposta considerando a cobertura é possível observar certa constância nos valores obtidos em cada cenário. Analisando a primeira iteração dos cenários avaliados é possível observar um aumento no valor da cobertura ao passo que a variável *k* é incrementada, com exceção para o cenário 2.

Comparando a primeira iteração dos cenários 1 e 5 é possível perceber um incremento de 0,02 no valor da cobertura, com a variável *k* assumindo os valores 4 e 8 respectivamente. Este mesmo incremento é observado na quinta iteração destes mesmos cenários.

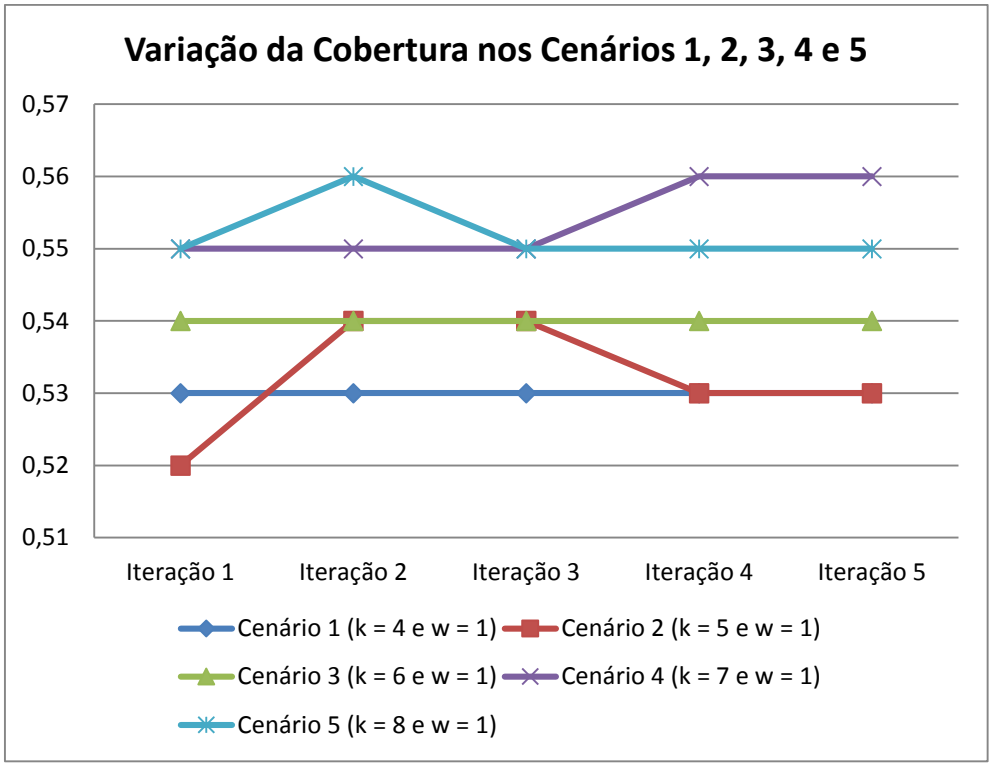

Figura 5.3 – Variação da Cobertura média para os cenários 1, 2, 3, 4 e 5

Conforme a análise realizada, é possível concluir que o incremento da variável *k* teve efeito positivo nos valores de precisão, cobertura e medida-F obtidos pela abordagem proposta. Contudo, a variação destes valores não alcançou 0,1 em uma escala de 0 a 1. Isto permite concluir que em cenários de aplicação da abordagem proposta esta variável possa ser configurada com valores baixos, resultando em um número menor de solicitações de *feedback* ao usuário. A Figura 5.4 apresenta o gráfico da Figura 5.1 com os valores em termos de medida-F apresentados em uma escala de 0 a 1.

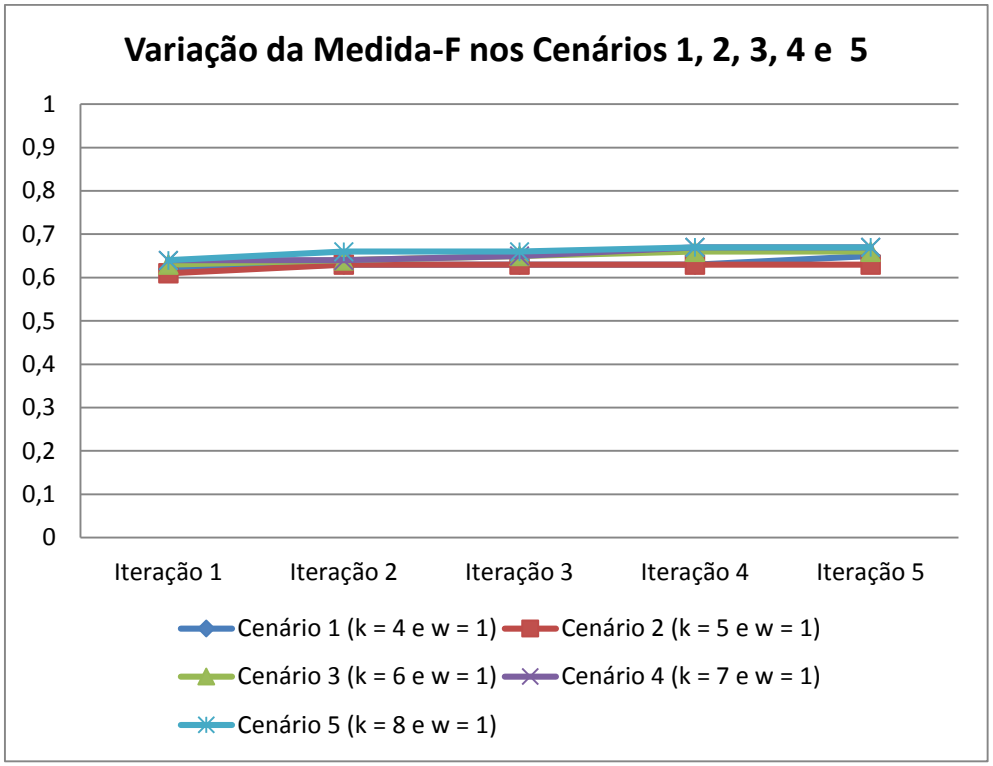

Figura 5.4 – Medida-F nos cenários 1, 2, 3, 4 e 5 em escala de 0 a 1

Questão 2: *Qual o efeito da variação no valor assumido pela variável w na qualidade dos alinhamentos gerados (precisão, cobertura e medida-F)?*

Finalizada a análise do comportamento da variável *k*, o próximo passo foi verificar o efeito da variação da variável *w* nos valores de precisão, cobertura e medida-F. Para realizar esta verificação, a variável *w* dos cenários 1 e 5 foi incrementada, passando a solicitar *feedbacks* sobre até dois pares de entidades nos cenários 6 e 8 e até três pares de entidades nos cenários 7 e 9 a cada iteração. Os resultados das execuções dos cenários 6, 7, 8 e 9 são apresentados nas Tabelas 0.6, 0.7, 0.8 e 0.9 respectivamente e estão disponíveis para consulta no apêndice deste trabalho.

O gráfico da Figura 5.5 apresenta a variação da medida-F nos cenários 1, 5, 6, 7, 8 e 9. Os resultados apresentados foram consolidados por cenário e iteração, e determinados calculando o valor médio da medida-F obtida a partir de cada um dos vinte e um pares de ontologias avaliados.

Analisando a primeira iteração dos cenários 1, 6 e 7, é possível observar um incremento de 0,02 no valor da medida-F, quando comparados os cenários 1 e 6. Contudo, este comportamento não é observado entre os cenários 6 e 7, com o valor

desta medida se mantendo constante. Na quinta iteração, o valor da medida-F se mantém constante, quando comparados os cenários 1 e 6, e apresenta incremento de 0,01, quando comparados os cenários 6 e 7.

Realizando análise análoga com os cenários 5, 8 e 9, é possível observar um incremento de 0,02 no valor da medida-F, quando comparados os cenários 5 e 8, e de 0,01, quando comparados os cenários 8 e 9. Na quinta iteração, o valor da medida-F se mantém constante, quando comparados os cenários 5 e 8, e apresenta incremento de 0,01, quando comparados os cenários 8 e 9.

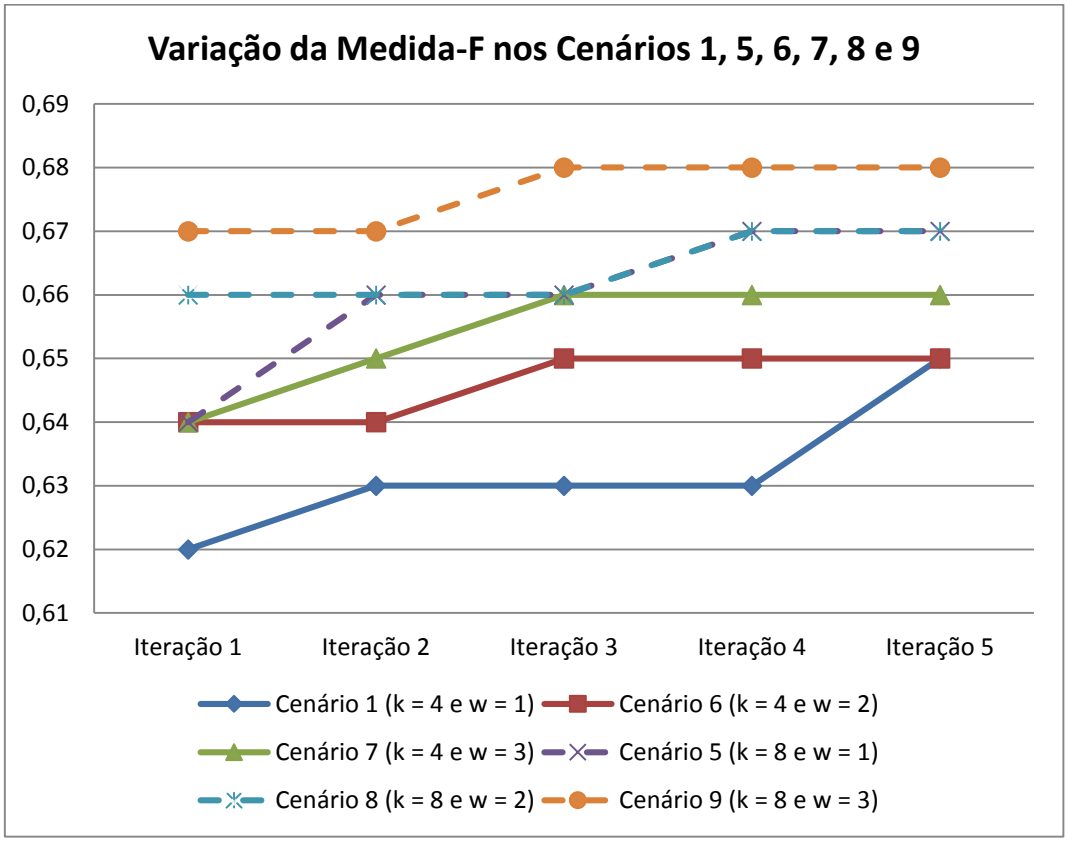

Figura 5.5 – Variação da Medida-F média nos cenários 1, 5, 6, 7, 8 e 9

Os resultados obtidos pela abordagem proposta em termos de precisão e cobertura para os cenários são apresentados nos gráficos da Figura 5.6 e 5.7 respectivamente. Os resultados apresentados foram consolidados por cenário e iteração, e determinados calculando o valor médio das medidas de precisão e cobertura obtidas a partir de cada um dos vinte e um pares de ontologias avaliados.

Analisando a primeira iteração dos cenários 1, 6 e 7, é possível observar um incremento de 0,05 no valor da precisão, quando comparados os cenários 1 e 6, e de 0,01, quando comparados os cenários 6 e 7. Na quinta iteração, o valor da precisão se

mantém constante, quando comparados os cenários 1 e 6, e apresenta incremento de 0,01, quando comparados os cenários 6 e 7. Analisando os cenários 1 e 7, é possível observar que na primeira iteração ocorre um incremento de 0,06, com a variável *w* assumindo os valores 1 e 3 respectivamente. Contudo, na quinta iteração esta diferença passa a ser de apenas 0,01, demonstrando o bom desempenho da abordagem ao considerar um número mínimo de *feedbacks* ao longo das iterações.

Realizando análise análoga entre os cenários 5, 8 e 9, é possível observar um incremento de 0,02 no valor da precisão, quando comparados os cenários 5 e 8, e de 0,01, quando comparados os cenários 8 e 9. Na quinta iteração, o valor da precisão se mantém constante, considerando os três cenários analisados, o que novamente comprova o desempenho da abordagem ao considerar um número mínimo de *feedbacks* ao longo das iterações.

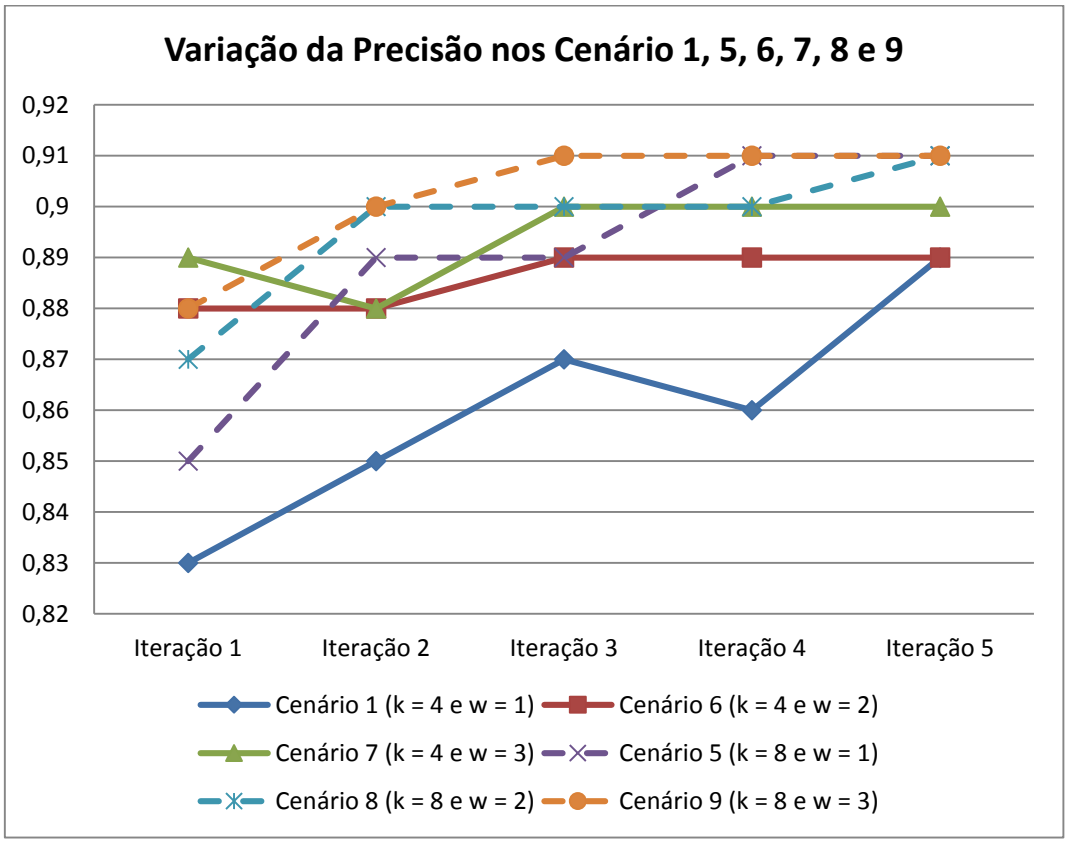

Figura 5.6 – Variação da Precisão média nos cenários 1, 5, 6, 7, 8 e 9

Analisando a primeira iteração dos cenários 1, 6 e 7, em termos de cobertura, é possível observar um decremento de 0,01 no valor desta medida, quando comparados os cenários 1 e 6. Nos cenários 6 e 7 este valor se mantém constante. Na quinta iteração, o

valor da cobertura apresenta incremento de 0,01, quando comparados os cenários 1 e 6 e os cenários 6 e 7.

Realizando análise análoga entre os cenários 5, 8 e 9, é possível observar que o valor da cobertura se mantém constante, quando comparados os cenários 5 e 8, e apresenta incremento de 0,01, quando comparados os cenários 8 e 9. Na quinta iteração, o valor da cobertura apresenta incremento de 0,01, entre os cenários 5 e 8 e entre os cenários 8 e 9.

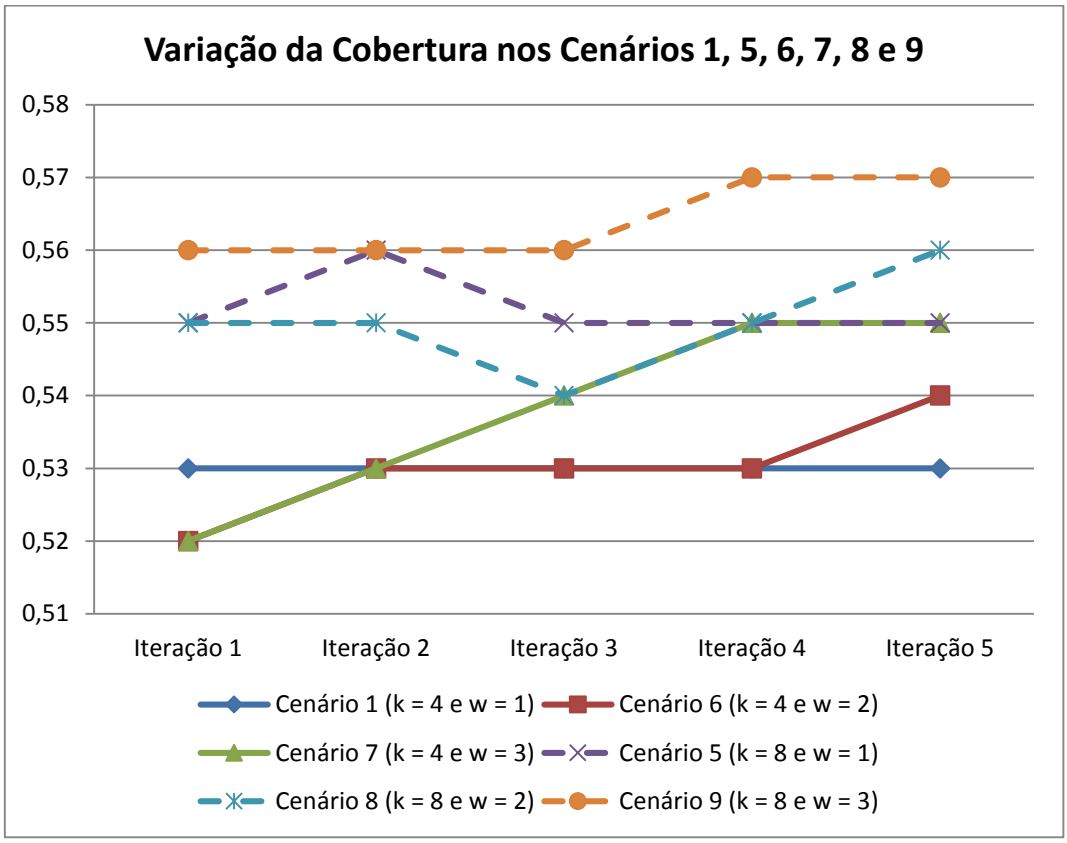

Figura 5.7 – Variação da Cobertura média nos cenários 1, 5, 6, 7, 8 e 9

De acordo com a análise realizada sobre os resultados dos cenários 1, 6 e 7 e dos cenários 5, 8 e 9, em termos dos valores de precisão, cobertura e medida-F pode-se concluir que o incremento da variável *w* teve efeito positivo nos valores retornados por estas medidas. Contudo, a variação destes valores não alcançou 0,1 em uma escala de 0 a 1, o que permite concluir que em cenários de aplicação da abordagem proposta, esta variável possa ser configurada com o valor mínimo, resultando em um número menor de solicitações de *feedback* ao usuário. A Figura 5.8 apresenta o gráfico da Figura 5.5 com os valores obtidos em termos da medida-F apresentados em uma escala de 0 a 1.

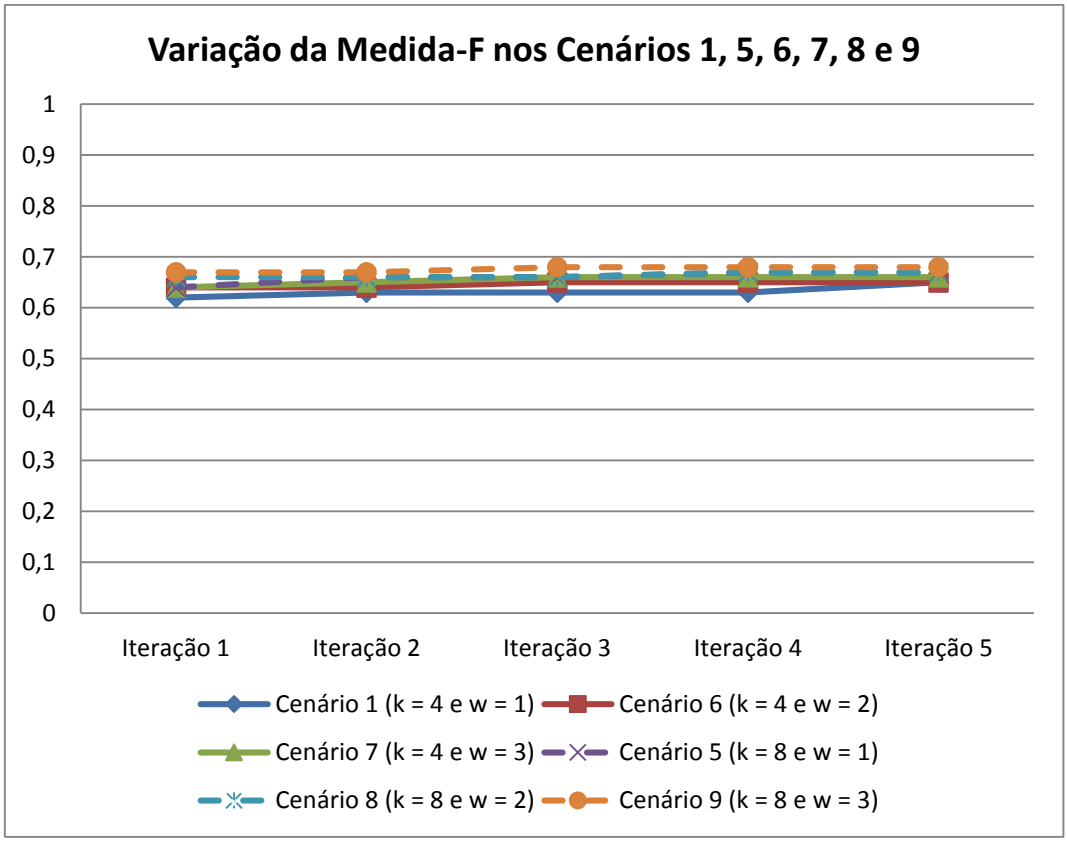

Figura 5.8 – Medida-F nos cenários 1, 5, 6, 7, 8 e 9 em escala de 0 a 1

5.3 Avaliação dos Resultados

Conforme concluído a partir da análise dos cenários executados, os resultados obtidos pela abordagem proposta são suscetíveis à configuração das variáveis *k* e *w*. Portanto, para comparar o desempenho da abordagem proposta em relação ao estado da arte dos sistemas interativos de alinhamento de ontologias, representados pelos sistemas avaliados na trilha de *interactive matching* da OAEI, foi necessário selecionar um dos cenários definidos para avaliação da abordagem.

Considerando o objetivo deste trabalho, que consiste na redução do número de *feedbacks* do usuário e aumento do efeito destes *feedbacks* ao longo das iterações, foi escolhido o cenário 1, e os resultados obtidos na primeira iteração. A Tabela 5.3 apresenta os resultados (em termos de precisão, cobertura, medida-F e número de *feedbacks* solicitados) obtidos pelos sistemas avaliados na trilha de *interactive matching* da OAEI (campanha de 2014), bem como os resultados obtidos pela abordagem proposta neste trabalho.

Tabela 5.3 – Sistemas de alinhamento de ontologia avaliados na trilha *Interactive Matching* [44]

Sistema de Alinhamento	**Precisão**	**Cobertura**	**Medida-F**	**Número Feedbacks**
JARVIS	0,89	0,53	0,62	4,7
AML [45]	0,91	0,73	0,80	6,9
Hertuda [46]	0,79	0,50	0,58	12,3
LogMap [47]	0,89	0,64	0,73	4,1
WeSeE [48]	0,73	0,40	0,47	5,5

A abordagem proposta (JARVIS) obteve uma medida-F média de 0,62. Para alcançar este valor, foram solicitados, em média, 4,7 *feedbacks* (pares de entidades). Comparando este valor com os resultados obtidos pelos sistemas avaliados pela OAEI, é possível concluir que a abordagem proposta apresentou resultado compatível com os sistemas avaliados, que a posicionou entre as três melhores abordagens para alinhamento interativo de ontologias.

A abordagem proposta neste trabalho realiza todo o processo de aprendizado (*active learning*) explorando características das ontologias a serem alinhadas. Para realizar os experimentos apresentados neste capítulo, foram consideradas medidas de similaridade baseadas em *string* e semântica. Ao explorar apenas estas características das ontologias, a capacidade de alinhamento da abordagem ficou restrita, o que afetou diretamente o resultado obtido. Para obter resultados melhores, é preciso explorar outras características das ontologias, como o tipo de entidade a ser alinhada, a sua estrutura (relacionamentos e atributos) e suas instâncias.

No capítulo seguinte serão apresentadas as características dos sistemas de alinhamento de ontologias participantes da trilha *interactive matching*, bem como outros trabalhos relacionados que tem como característica a participação do usuário no processo de alinhamento de ontologias.

6 Trabalhos Relacionados

Este capítulo tem por objetivo apresentar o estado da arte das abordagens de alinhamento de ontologias que buscam incluir o usuário neste processo com a finalidade de melhorar a qualidade dos alinhamentos. Finalizando este capítulo, será apresentado um comparativo destas abordagens com a abordagem proposta neste trabalho.

6.1 Descrição das Abordagens Relacionadas

Após a revisão do estado da arte das abordagens de alinhamento de ontologias, foi selecionado um conjunto de propostas que consideram a participação do usuário neste processo como forma de melhorar a qualidade dos alinhamentos. As abordagens descritas a seguir têm como característica a solicitação de *feedbacks* do usuário sobre pares de entidades, apresentando diferentes estratégias para seleção destes pares de entidades e para a propagação do efeito do *feedback* do usuário para outros pares de entidades.

Em [49], Duan *et al.* apresentam um conjunto de técnicas que empregam o *feedback* do usuário para customizar o processo de alinhamento de ontologias. Este trabalho explora características léxicas, semânticas e estruturais das ontologias em um processo chamado de propagação estrutural supervisionada iterativa. Estas características são exploradas aplicando medidas de similaridades associadas a pesos, que variam de acordo com a importância da medida na determinação das correspondências. Estes pesos são aprendidos através de uma abordagem de aprendizado supervisionada utilizando regressão logística, em que as instâncias de treinamento são geradas através de um arquivo de *gold standard*. Este arquivo é fornecido pelo usuário e nele são especificados pares de entidades correspondentes. A

propagação do efeito da participação do usuário é realizada de forma iterativa explorando a estrutura da ontologia. Para este fim, são empregadas medidas estruturais agregadas que determinam similaridades máxima e média de um par de entidades considerando outras entidades conectadas a elas através de um relacionamento identificado por um rótulo, como por exemplo, *subclassOf*. A cada iteração realizada pela abordagem, os pesos são aprendidos e as correspondências candidatas são estendidas para incluir novos pares de entidades que tenham pelo menos uma medida de similaridade estrutural diferente de zero.

No trabalho de Shi *et al.* [7], é apresentado um *active learning framework* para alinhamento de ontologias, que tem por objetivos identificar os pares de entidades mais informativos, para consulta ao usuário, e propagar a correção através da estrutura da ontologia, melhorando a acurácia do alinhamento. De acordo com a abordagem, os pares de entidades mais informativos são aqueles identificados como possíveis erros (*error match*). Estes possíveis erros são medidos pela taxa de erro, calculada combinando três medidas: *confidence*, *similarity distance* e *contention point*. *Confidence* mede a efetividade de um método de alinhamento (medida de similaridade) em relação à correção das correspondências, *similarity distance* é calculada considerando os valores de similaridade obtidos dos mapeamentos de uma entidade da ontologia origem para duas ou mais entidades na ontologia destino e *contention point* é determinado considerando o desacordo entre os métodos de alinhamento (medida de similaridade) empregados. A seleção de pares de entidades para *feedback* do usuário é realizada considerando a taxa de erro e a taxa de propagação. A taxa de propagação considera o efeito de uma correspondência possivelmente incorreta em outras e é determinada utilizando uma estrutura chamada grafo de propagação, baseada no algoritmo *similarity flooding*. Neste grafo, os pares de entidades são representados por nós e os relacionamentos entre entidades por arestas duplas com pesos, que indicam o quanto a similaridade de um par de entidades propaga para outro. A propagação do efeito do *feedback* é realizada em um processo iterativo utilizando o grafo de propagação, e a cada confirmação do usuário os valores de similaridade dos pares de entidades interligados no grafo são decrementados ou incrementados, caso o usuário confirme o par de entidades como uma correspondência incorreta ou não.

Em [32], To *et al.* propõem um framework de aprendizado de máquina adaptativo que emprega múltiplas estratégias de aprendizado de máquina e a interação com o

usuário para melhorar a qualidade do alinhamento. No trabalho de To *et al.*, são considerados dois tipos de interação com o usuário: pré-alinhamento e *feedback*. Dependendo do cenário apresentado, a abordagem seleciona a estratégia de aprendizado: supervisionado, se existir um pré-alinhamento com muitos dados rotulados; ou semi-supervisionado, se o usuário desejar interagir com o sistema fornecendo *feedbacks*. Nos dois casos, vetores de similaridade são extraídos das instâncias de treinamento fornecidas pelo usuário e utilizadas, pelo algoritmo *Naive Bayes*, para treinar um modelo de classificação. No método semi-supervisionado com *feedback* relevante, inicialmente são selecionados pares de entidades para classificação manual do usuário, aplicando a técnica de *clustering*. Estes pares classificados são utilizados para treinar o classificador, que uma vez gerado, é utilizado para classificar os demais pares de entidades. A cada iteração realizada, são selecionados, para confirmação do usuário, pares de entidades em que ocorre a menor confiança na classificação. Após esta confirmação, estes pares de entidades são adicionados ao conjunto de treinamento e utilizados na iteração seguinte. O processo se repete por um determinado número de iterações e então é alterado para o modo automático que utiliza os pares de maior confiança para treinar o classificador.

A abordagem apresentada por Wagner *et al.* [50], estende abordagens tradicionais de alinhamento de ontologias incluindo propriedades incrementais, interativas e iterativas. A cada iteração realizada são geradas partições das ontologias a serem alinhadas. As entidades que compõem cada partição são automaticamente selecionadas com base na proximidade, definida nesta abordagem como o número mínimo de relacionamentos que ligam estas entidades. Definidas estas partições, algoritmos de alinhamento são utilizados para gerar correspondências entre entidades, gerando um alinhamento entre as partições das ontologias. O usuário fornece *feedbacks* sobre todos os pares de entidades identificados como correspondentes pelo algoritmo de alinhamento previamente selecionado, rejeitando estes pares ou mesmo adicionando pares não identificados. O resultado da participação do usuário é um conjunto de correspondências revisado, que é armazenado em um repositório para utilização nas interações seguintes.

Em Cruz *et al.* [22], é apresentada uma abordagem interativa para o processo de alinhamento de ontologias, que incorpora o *feedback* do usuário em um conjunto de iterações. A abordagem é baseada na ideia de *signature vectors*, que são elementos

chave para seleção de pares de entidades que serão apresentados ao usuário, bem como para a propagação do *feedback* para outras correspondências. A seleção de pares de entidades para *feedback* do usuário é realizada considerando a *disagreement metric*, definida como a variância dos valores de similaridade no *signature vectors*. O valor desta medida diminui à medida que os *matchers* (Medidas de similaridade) concordam com a correspondência entre as entidades, e aumenta em caso contrário. Pares de entidades são agrupados em *clusters*, de acordo com seu *signature vectors*, sendo que estes *clusters* são delimitados por um *threshold*. A cada iteração os *k* pares de entidades de maior *disagreement metric* são selecionados para *feedback* do usuário e este *feedback* é propagado para pares de entidades similares mantidos em um *cluster*. Esta propagação é realizada incrementando (ou decrementando) o valor de similaridade dos pares de entidades contidos no mesmo *cluster*, quando o par de entidades é confirmado (ou não) como correspondente. O valor de similaridade dos pares de entidades no *cluster* é determinado a partir de uma função linear que utiliza pesos para indicar a qualidade da medida de similaridade empregada.

WeSeE [48] é um sistema para alinhamento de ontologias baseado em elementos, que utiliza informação da web para alinhar ontologias. Este sistema utiliza um componente de busca na web para obter documentos relevantes que tenham associação com os conceitos das ontologias a serem alinhadas. Para cada conceito é realizado uma busca utilizando o fragmento da *URI*, *label* e comentário, sendo gerados três documentos respectivamente. A similaridade de cada par de conceitos é calculada como a similaridade máxima considerando todos os documentos gerados para estes conceitos. WeSeE utiliza um método básico para alinhamento interativo, em que busca definir um *threshold* para selecionar o alinhamento final de um conjunto de correspondências candidatas. Internamente o sistema trabalha com um *threshold* duplo (máximo e mínimo) que é atualizado à medida que o usuário é solicitado a fornecer *feedbacks* sobre pares de entidades dentro de um processo iterativo. A seleção é realizada considerando um *threshold* médio e pares de entidades que possuem valor de confiança igual ao deste *threshold* são escolhidos para *feedback*. A cada iteração os *thresholds* máximo e mínimo são atualizados propagando o efeito do *feedback*. Este processo é executado enquanto o *threshold* máximo for maior que o *threshold* mínimo. Quando esta condição não for mais alcançada, o alinhamento final é gerado considerando o *threshold* obtido e os *feedbacks* do usuário.

AgreementMakerLight [45] [51] é um framework automático para alinhamento de ontologias derivado do sistema AgreementMaker. Internamente, AML aplica três algoritmos de alinhamento que exploram os elementos das ontologias: *Lexical Matcher*, *Word Matcher* e *Parametric String Matcher*. A variação AML-BK utiliza recursos externos como *background knowledge,* além de explorar os elementos das ontologias aplicando os algoritmos anteriores. O algoritmo *Wordnet Matcher* consulta a base de dados da *Wordnet* buscando por sinônimos para os nomes das entidades, posteriormente utilizados pelo *Lexical Matcher* na determinação de possíveis correspondências. Na abordagem interativa, AML emprega um algoritmo de seleção interativa que solicita ao usuário *feedbacks* sobre pares de entidades em caso de conflito, avaliando seus relacionamentos, ou similaridade abaixo de um dado *threshold*, até que um determinado número de respostas negativas seja atingido.

LogMap [47] é um sistema de alinhamento de ontologias altamente escalável com capacidades de raciocínio e reparação de inconsistências embutido. Seu algoritmo pode ser dividido em dois estágios principais [52]: *Computation of candidate mappings*, e *Assessment of candidate mappings*. No estágio *Computation of candidate mappings* é gerado um conjunto tipicamente grande de correspondências candidatas utilizando somente técnicas léxicas. Neste estágio são empregadas ainda, técnicas de modularização baseadas em lógica, utilizadas para fragmentar as ontologias capturando o significado dos conceitos. Já o estágio *Assessment of candidate mappings* tem por objetivo maximizar a precisão sem prejudicar a cobertura, progressivamente descartando correspondências que provavelmente são incorretas, bem como identificando as corretas. Técnicas léxicas, estruturais e baseadas no raciocínio são empregadas neste estágio. Caso o sistema esteja em modo interativo, e existam incertezas quanto a correspondência entre pares de entidades, o *feedback* do usuário é solicitado. Neste caso, as possíveis correspondências em que ocorre a incerteza são ordenadas de acordo com seus valores de similaridade e exibidas para confirmação do usuário, que escolhe entre aceitar ou rejeitar a possível correspondência. O efeito do *feedback* do usuário é propagado identificando possíveis correspondências ambíguas (A ≡ B é ambíguo a C ≡ D, se C = A ou D = B) e automaticamente aceitando ou rejeitando estas possíveis correspondências.

Hertuda [46] é um sistema de alinhamento de ontologias baseado em elementos e comparação de strings. Este sistema gera correspondências homogêneas, o que significa

que classes, *data properties* e *object properties* são processados separadamente. Três *thresholds* podem ser definidos independentemente um para cada tipo de elemento a ser alinhado. Uma matriz de similaridade é calculada através de um produto cartesiano dos elementos das ontologias utilizando a distância *Damerau–Levenshtein*. Hertuda utiliza o mesmo método para alinhamento interativo que o sistema WeSeE, em que *feedbacks* do usuário são utilizados para definir um *threshold* com o objetivo de selecionar um conjunto de correspondências e gerar um alinhamento final.

6.2 Análise das Abordagens Relacionadas

Conforme demonstrado anteriormente diversas abordagens apresentadas na literatura consideram a participação do usuário no processo de alinhamento com o objetivo de melhorar a qualidade dos alinhamentos. Nesta seção estas abordagens terão suas características comparadas com a abordagem proposta neste trabalho, considerando principalmente as estratégias utilizadas para seleção de pares de entidades e para propagação do efeito do *feedback*.

A abordagem proposta por Duan *et al.* não apresenta uma estratégia para seleção de pares de entidades para *feedback* do usuário, sendo baseada em um arquivo de *gold standard* fornecido como entrada. Contudo, Duan *et al.* destacam a necessidade de uma estratégia de seleção que permita a melhoria do aprendizado ao longo das iterações, o que é endereçado pela abordagem proposta neste trabalho, que realiza a seleção de pares de entidades considerando o desacordo entre classificadores. Em relação a estratégia de propagação do efeito do *feedback*, na proposta apresentada por Duan *et al.*, é utilizada regressão logística, que difere da abordagem proposta, que emprega a classificação.

Shi *et al.* apresentam uma abordagem para seleção de pares de entidades para *feedback* do usuário baseada na identificação de possíveis incorreções (taxa de erro) e no efeito destas incorreções nos vizinhos (taxa de propagação). Shi *et al.* destacam que ao focar apenas nos possíveis erros, a abordagem em determinados momentos não consegue melhorar os resultados em termos de cobertura, o que não é esperado na abordagem proposta neste trabalho que seleciona pares de entidades para *feedback* considerando o nível de informação (correspondentes ou não) medido pelo desacordo entre classificadores. Já em relação a estratégia de propagação adotada, na abordagem proposta por Shi *et al.*, é utilizando um grafo baseado no algoritmo *similarity flooding*

diferentemente da abordagem proposta neste trabalho que emprega a técnica de classificação.

Wagner *et al.* apresentam uma estratégia para seleção de pares de entidades para *feedback* que emprega particionamento de ontologias e diferentes algoritmos de alinhamento. Contudo, esta abordagem não apresenta uma estratégia para propagação do efeito do *feedback* para outros pares de entidades durante as iterações, diferentemente da abordagem proposta. Espera-se que com a capacidade de aprendizado, baseada na seleção de pares de entidades informativos e na propagação através de classificadores, implementada pela abordagem proposta neste trabalho o usuário forneça um número menor de *feedbacks* ao longo das iterações.

Em seu trabalho, Cruz *et al.* propõe uma estratégia para seleção de pares de entidades para *feedback* do usuário baseada na variância dos valores de similaridade obtidos a partir de algoritmos de alinhamento. Em relação a estratégia de propagação, esta abordagem propõe o incremento ou decremento da similaridade agregada agrupando pares de entidades semelhantes em um cluster. Estas estratégias diferem da apresentada neste trabalho que emprega seleção baseada no nível de informação e propagação baseada em classificação. Algoritmos de alinhamento, que exploram diferentes perspectivas das ontologias, tendem a apresentar muitas variações no valores de similaridades (para cada par de entidades possível), o que pode levar a um aumento no espaço de busca de pares de entidades para *feedback*. Portanto, espera-se que com a abordagem de seleção proposta neste trabalho, baseada no desacordo entre classificadores ocorra uma redução neste espaço de busca, resultando na redução no número de *feedbacks* necessários.

To *et al.* apresentam uma estratégia de seleção de pares de entidades, baseada na confiança do classificador *Naive Bayes*, o que difere da abordagem proposta neste trabalho. Ambas abordagens apresentam o alinhamento de ontologias no contexto de *active learning* e pelos cenários descritos trabalham com poucas instâncias de treinamento. Neste cenário, ao considerar a visão de apenas um classificador a abordagem pode estar suscetível a erros de classificação. Portanto, espera-se que com a abordagem de seleção proposta neste trabalho, baseada no desacordo de um conjunto de classificadores heterogêneos (*Naive Bayes, Random Forest* e *Multilayer Perceptron*), haja um aumento na qualidade dos alinhamentos (medida-F). Em relação a estratégia de propagação, ambas abordagens empregam a classificação, contudo na abordagem

proposta neste trabalho, esta classificação é realizada por um comitê de classificadores. A determinação da classe é realizada contabilizando os votos dos membros do comitê, sendo escolhida aquela que obteve o maior número de indicações. Esta característica é vantajosa por considerar a visão da maioria, o que pode contribuir para redução do número de correspondências incorretas.

WeSeE e Hertuda apresentam soluções automáticas para alinhamento de ontologias com variação para a participação do usuário, o que difere da abordagem proposta neste trabalho centrada na participação do usuário. Em modo interativo, estas abordagens selecionam pares de entidades utilizando um *threshold* médio. Já a propagação do *feedback* é realizada atualizando um *threshold* máximo e um *threshold* mínimo. A estratégia de determinar um *threshold* utilizando *feedbacks* do usuário difere da abordagem proposta neste trabalho baseada em *active learning*. Diferentes técnicas de alinhamento tendem a gerar variações nos valores de similaridade e ao considerar uma similaridade agregada e um *threshold* para seleção dos pares de entidades correspondentes, a abordagem pode aumentar a chance de erros, incluindo no alinhamento pares de entidades identificados como correspondentes quando não são. Espera-se que a abordagem proposta neste trabalho seja capaz de se adaptar melhor as variações de similaridade, utilizando para isto classificadores gerados com base no *feedback* do usuário.

LogMap também apresenta uma solução automática para alinhamento de ontologias com uma variação para a participação do usuário, o que difere da abordagem proposta. Esta abordagem adota uma estratégia, em modo interativo, que seleciona pares de entidades que não foram claramente incluídos ou excluídos pelas heurísticas automáticas. Em relação à propagação do efeito do *feedback*, esta abordagem adota heurísticas baseadas em conflitos e ambiguidades. Ao considerar apenas os pares de entidades que não foram claramente incluídos ou excluídos pelas heurísticas automáticas (incerteza) como candidatos ao *feedback* do usuário, LogMap pode gerar alinhamentos com pares de entidades supostamente correspondentes e que não são. Isto pode ocorrer, devido ao fato do *feedback* do usuário não ter seu efeito propagado para os pares de entidades considerados como correspondentes pelas heurísticas automáticas (certeza). Diferentemente, a abordagem proposta neste trabalho é baseada no *feedback* do usuário, ou seja, a cada interação realizada os classificadores são treinados e as correspondências candidatas classificadas. Espera-se que esta abordagem aproveite de

maneira mais efetiva o *feedback* do usuário, alcançando alinhamentos de melhor qualidade.

Seguindo a mesma linha de WeSeE, Hertuda e LogMap, AML e AML-BK também apresentam uma abordagem automática para alinhamento de ontologias com uma variação para a participação do usuário. A estratégia de seleção de pares de entidades para *feedback* do usuário é baseada na identificação de possíveis conflitos (estruturais), e na utilização de um *threshold*, o que a difere da abordagem proposta, que seleciona pares de entidade mais informativos. AML e AML-BK não apresentam uma estratégia para propagação do efeito do *feedback* para outros pares de entidades, diferentemente da abordagem proposta.

Para efeito de comparação entre as abordagens apresentadas neste capítulo e a abordagem proposta neste trabalho, a Tabela 6.1 apresenta um resumo das estratégias de seleção de pares de entidades e propagação do *feedback* anteriormente descritas.

Tabela 6.1 – Matriz de trabalhos relacionados

Trabalho	**Estratégia de seleção de pares de entidades para feedback do usuário**	**Estratégia de propagação do efeito do feedback do usuário**
Jarvis	Desacordo entre classificadores.	Classificação.
Duan *et al.*	Não possui.	Regressão Logística.
Shi *et.al.*	Taxa de erro e taxa de propagação.	Grafo Propagação.
To *et. al.*	Confiança quanto a classificação.	Classificação.
Cruz *et. al.*	Variância dos valores de similaridade	Função linear
WeSeE	Threshold médio.	Threshold mínimo e máximo.
Wagner *et al.*	Particionamento da ontologia e correspondências geradas por um algoritmo	Não possui.
Hertuda	Threshold médio.	Threshold mínimo e máximo.
LogMap	Incerteza das heurísticas automáticas.	Heurísticas de ambiguidade e conflito.
AML e AML-BK	Conflitos estruturais e threshold.	Não possui.

No capítulo 5, os resultados das abordagens apresentadas por WeSeE, Hertuda, LogMap, AML e AML-BK são comparados, em termos dos valores de precisão,

cobertura e medida-F, à abordagem proposta neste trabalho seguindo as orientações da trilha de *interactive matching* da OAEI.

No capítulo seguinte serão apresentadas as conclusões sobre a abordagem proposta, bem como os pontos a serem evoluídos em trabalhos futuros.

7 Conclusão

Este capítulo apresenta a conclusão deste trabalho e suas contribuições, bem como as limitações da abordagem proposta e trabalhos futuros.

7.1 Considerações Finais

O número crescente de ontologias heterogêneas que muitas vezes descrevem o mesmo domínio de interesse tem se apresentado como um desafio para a interoperabilidade de dados. Este desafio vem sendo enfrentado pela área de estudo de Alinhamento de Ontologias, que tem na identificação de correspondências entre entidades a solução para o problema da heterogeneidade ontológica.

Os estudos na área de Alinhamento de Ontologias vêm avançando ao longo dos anos e novos desafios têm surgido diante das novas demandas. Dentre estes novos desafios, está o endereçado por este trabalho que trata da busca por soluções interativas para o alinhamento de ontologias com o objetivo de melhorar a qualidade dos alinhamentos gerados.

Conforme discutido durante este trabalho, diversas abordagens consideram a solicitação de *feedbacks* relevantes sobre pares de entidades como forma de envolvimento do usuário no processo de alinhamento. Estas abordagens interativas têm em comum a busca pelo aumento do efeito da participação do usuário, reduzindo o seu esforço ao longo das interações.

Neste sentido, este trabalho apresentou uma abordagem interativa para alinhamento de ontologias baseada na estratégia de seleção conhecida como *query-by-committee*. A estratégia *query-by-committe* busca selecionar instâncias informativas, no contexto deste trabalho, pares de entidades, considerando o desacordo (na classificação) entre membros de um comitê de classificadores.

Os classificadores que compõem o comitê são treinados utilizando um conjunto inicial de instâncias classificadas. Na abordagem apresentada neste trabalho, estas instâncias foram representadas por pares de entidades das ontologias selecionados utilizando a heurística do algoritmo Farthest First e classificados pelo usuário.

O desacordo entre os classificadores foi mensurado empregando a medida *vote entropy* proposta por Dagan e Engelson. A cada iteração realizada utilizando a abordagem proposta foram selecionados os pares de maior *vote entropy* e menor distância euclidiana média, sendo que esta última foi empregada com o objetivo de evitar a consulta de *outliers*.

Para avaliar a abordagem proposta neste trabalho, foi executado um conjunto de experimentos que simularam diferentes cenários com variações para o número de pares selecionados para *feedback* do usuário nas diferentes situações em que são necessários. Estes experimentos foram realizados utilizando as ontologias do *conference data set*, empregado pela OAEI na avaliação de abordagens interativas (*interactive matching*).

Os resultados obtidos pela abordagem proposta nos experimentos realizados foram comparados com os sistemas avaliados pela OAEI na trilha *interactive matching*. Estes resultados demonstraram que a abordagem apresentada é promissora, pois os alinhamentos gerados apresentaram qualidade compatível com o estado da arte das soluções avaliadas pela OAEI nesta trilha, considerando as medidas de precisão, cobertura e medida-F.

7.2 Limitações e Trabalho Futuros

Por se tratar de uma abordagem de alinhamento de ontologias baseada em *feedbacks* do usuário, a principal limitação deste trabalho está relacionada a dependência de um usuário especialistas no domínio que responda as solicitações para classificação de pares de entidades durante o processo de alinhamento.

Outra limitação da abordagem proposta está relacionada a configuração adequada das variáveis *k* e *w*. Estas variáveis representam o número de *feedbacks* solicitados ao usuário em diferentes situações durante o processo e conforme visto no capítulo 5, os valores assumidos por elas afetam os resultados obtidos pela abordagem proposta. Seguindo a linha de explorar apenas o conhecimento do domínio de um usuário, uma nova versão da abordagem proposta deve considerar a definição de um mecanismo que

determine os valores das variáveis *k* e *w* automaticamente sem a necessidade de intervenção do usuário.

Para realização dos experimentos de avaliação da abordagem proposta foi empregado um conjunto de medidas de similaridade baseadas em *string* e medidas semânticas. Em trabalhos futuros, é necessário estender esta abordagem, tornando-a capaz de explorar outros aspectos das ontologias, incluindo a estrutura (interna e relações) e possíveis instâncias.

Um dos avanços nesta pesquisa está relacionado a sua adaptação para suportar um comitê de usuários. A ideia é implementar mecanismos que lidem com o possível desacordo entre especialistas no domínio quanto a correspondência de um par de entidades. Este desacordo entre especialistas pode ser visto como uma oportunidade para a melhoria da qualidade dos alinhamentos.

Concluindo, no futuro esta abordagem deve ser evoluída de forma a considerar uma estrutura capaz de armazenar as experiências, obtidas ao longo do tempo com os *feedbacks* do usuário, e reutilizar estas experiências em novos alinhamentos.

Referências

[1] BERNERS-LEE, T., HENDLER, J., LASSILA, O. *The Semantic Web*. *Scientific American*. 2001.

[2] D'AQUIN, M., NOY, N. F. *Where to publish and find ontologies? A survey of ontology libraries*. *Journal of Web Semantics*. 2012.

[3] EUZENAT, J., SHVAIKO, P. *Ontology matching, 2nd Edition*. 2013. p. 333

[4] EUZENAT, J., SHVAIKO, P. *Ontology matching*. 2007. p. 333

[5] EHRIG, M. Ontology Alignment - Bridging the Semantic Gap. In: MANAGEMENT. 2005.

[6] SHVAIKO, P., EUZENAT, J. Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*, v. 25, p. 158–176, doi:10.1109/TKDE.2011.253, 2013.

[7] SHI, F. et al. Actively Learning Ontology Matching via User Interaction. (A. Bernstein et al., Eds.)In: INTERNATIONAL SEMANTIC WEB CONFERENCE. *Anais*..., Lecture Notes in Computer Science. [S.l.]: Springer, 2009.

[8] NGO, D., BELLAHSENE, Z. YAM ++ - A combination of graph matching and machine learning approach to ontology alignment task. *Journal of Web Semantic The Semantic Web Challenge 2011 Special Issue (2012) 16*, 2012.

[9] ICHISE, R. Machine Learning Approach for Ontology Mapping Using Multiple Concept Similarity Measures. In: SEVENTH IEEE/ACIS INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION SCIENCE (ICIS 2008).

Disponível em: <http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4529843>, 2008.

[10] SETTLES, B. Active Learning Literature Survey. *Machine Learning*, v. 15, p. 201–221, doi:10.1.1.167.4245, 2010.

[11] RECKER, J. *Scientific Research in Information Systems: A Beginner's Guide*. [S.l.]: Springer Publishing Company, Incorporated, 2012.

[12] STUDER, R., BENJAMINS, V. R., FENSEL, D. *Knowledge engineering: Principles and methods*. *Data & Knowledge Engineering*. 1998.

[13] GUARINO, N., OBERLE, D., STAAB, S. What Is an Ontology? *Handbook on Ontologies*. 2009. p. 1–17.

[14] BAADER, F. et al. EDS*The Description Logic Handbook: Theory, Implementation, and Applications*. New York, NY, USA: Cambridge University Press, 2003.

[15] PETRAKIS, E. G. M. et al. Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies. In: 4 TH WORKSHOP ON MULTIMEDIA SEMANTICS (WMS06. *Anais*... [S.l: s.n.], 1998.

[16] CHEATHAM, M., HITZLER, P. String Similarity Metrics for Ontology Alignment. (H. Alani et al., Eds.)In: INTERNATIONAL SEMANTIC WEB CONFERENCE (2). *Anais*..., Lecture Notes in Computer Science. [S.l.]: Springer. Disponível em: <http://dblp.uni-trier.de/db/conf/semweb/iswc2013-2.html#CheathamH13>, 2013.

[17] COHEN, W. W., RAVIKUMAR, P. D., FIENBERG, S. E. A Comparison of String Distance Metrics for Name-Matching Tasks. In: PROCEEDINGS OF IJCAI-03 WORKSHOP ON INFORMATION INTEGRATION ON THE WEB. 2003.

[18] LIN, F., SANDKUHL, K. A survey of exploiting WordNet in ontology matching. In: IFIP INTERNATIONAL FEDERATION FOR INFORMATION PROCESSING. 2008.

[19] WU, Z., PALMER, M. Verb Semantics And Lexical Selection. In: PROC. OF THE 32ND ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. 1994.

[20] LIN, D. An Information-Theoretic Definition of Similarity. In: PROCEEDINGS OF ICML. 1998.

[21] JIANG, J. J., CONRATH, D. W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *CoRR*, v. cmp-lg/970, 1997.

[22] CRUZ, I. F., STROE, C., PALMONARI, M. Interactive User Feedback in Ontology Matching Using Signature Vectors. (A. Kementsietsidis & M. A. V. Salles, Eds.)In: ICDE. *Anais*... [S.l.]: IEEE Computer Society, 2012.

[23] MITCHELL, T. M. *Machine Learning*. 1997. v. 4p. 432

[24] MOHRI, M., ROSTAMIZADEH, A., TALWALKAR, A. *Foundations of Machine Learning*. [S.l.]: The MIT Press, 2012.

[25] WITTEN, I. H., FRANK, E., HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques (Google eBook)*. 2011. p. 664

[26] CICHOSZ, P. *Data mining algorithms : explained using R*. [S.l.]: Wiley, 2015.

[27] TAN, P.-N.;; STEINBACH, M. e KUMAR, V. *Introduction to Data Mining*. [S.l.]: Addison Wesley, 2006.

[28] FRIEDMAN, N. et al. Bayesian Network Classifiers. In: MACHINE LEARNING. 1997.

[29] BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, doi:10.1023/A:1010933404324, 2001.

[30] ZHANG, Q. J., GUPTA, K. C. *Neural Networks for RF and Microwave Design*. [S.l.]: Artech House, 2000.

[31] SILVA, A. A. *ATHENAS: Uma Avaliação Experimental da Combinação de Métricas de Similaridade para o Alinhamento de Ontologias através de Mineração de Dados*. 2013.

[32] TO H., I. R., LE, H. An Adaptive Machine Learning Framework with User Interaction for Ontology Matching. In: TWENTY-FIRST INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. 2009.

[33] FRIEDMAN, N., KOHAVI, R. Handbook of Data Mining and Knowledge Discovery. In: KLÖSGEN, W.; ZYTKOW, J. M. (Eds.). New York, NY, USA: Oxford University Press, Inc., 2002. p. 282–288.

[34] UTGOFF, P. E., UTGOFF, P. Incremental Induction of Decision Trees. *Machine Learning*, v. 4, p. 161–186, doi:10.1.1.10.181, 1989.

[35] QUINLAN, J. R. *Induction of decision trees*. *Machine Learning*. 1986.

[36] LIAW, A., WIENER, M. Classification and Regression by randomForest. *R news*, v. 2, p. 18–22, doi:10.1177/154405910408300516, 2002.

[37] RIEDMILLER, M. *Advanced supervised learning in multi-layer perceptrons — From backpropagation to adaptive learning algorithms*. *Computer Standards & Interfaces*. 1994.

[38] DASGUPTA, S. Performance Guarantees for Hierarchical Clustering. In: 15TH ANNUAL CONFERENCE ON COMPUTATIONAL LEARNING THEORY. Springer, 2002.

[39] SETTLES, B., CRAVEN, M. An analysis of active learning strategies for sequence labeling tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP 08*, p. 1070, doi:10.3115/1613715.1613855, 2008.

[40] MELVILLE, P., MOONEY, R. J. Diverse ensembles for active learning. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING. *Anais*... [S.l: s.n.]. Disponível em: <http://portal.acm.org/citation.cfm?id=1015385&dl=>, 2004.

[41] DAGAN, I., ENGELSON, S. P. Committee-Based Sampling For Training Probabilistic Classifiers. In: IN PROCEEDINGS OF THE TWELFTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING. 1995.

[42] DAVID, J. et al. The Alignment API 4.0. *Semantic Web*, v. 2, n. 1, p. 3–10, 2011.

[43] HALL, M. et al. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, v. 11, n. 1, p. 10–18, doi:10.1145/1656274.1656278, 2009.

[44] DRAGISIC, Z. et al. Results of the Ontology Alignment Evaluation Initiative 2014. In: PROCEEDINGS OF THE 9TH INTERNATIONAL WORKSHOP ON ONTOLOGY MATCHING COLLOCATED WITH THE 13TH INTERNATIONAL SEMANTIC WEB CONFERENCE {(ISWC} 2014), RIVA DEL GARDA, TRENTINO, ITALY, OCTOBER 20, 2014. Anais... [S.l: s.n.]. Disponível em: <http://ceur-ws.org/Vol-1317/oaei14_paper0.pdf>, 2014.

[45] FARIA, D. et al. AgreementMakerLight results for OAEI 2013. In: OM. *Anais*... [S.l: s.n.], 2013.

[46] HERTLING, S. Hertuda results for {OAEI} 2012. In: PROCEEDINGS OF THE 7TH INTERNATIONAL WORKSHOP ON ONTOLOGY MATCHING, BOSTON, MA, USA, NOVEMBER 11, 2012. Disponível em: <http://ceur-ws.org/Vol-946/oaei12_paper4.pdf>, 2012.

[47] JIMÉNEZ-RUIZ, E., GRAU, B. C., HORROCKS, I. LogMap and LogMapLt results for OAEI 2013. In: OM. 2013.

[48] PAULHEIM, H., HERTLING, S. WeSeE-Match results for OAEI 2013. In: OM. 2013.

[49] DUAN, S., FOKOUE, A., SRINIVAS, K. One size does not fit all : Customizing Ontology Alignment Using User Feedback. *ISWC*, v. 6496, p. 177–192, doi:10.1007/978-3-642-17746-0, 2010.

[50] WAGNER, F., MACÊDO, J. A. F. DE, LÓSCIO, B. F. An incremental and user feedback-based ontology matching approach. In: IIWAS. 2011.

[51] FARIA, D. et al. The AgreementMakerLight Ontology Matching System. In: OTM CONFERENCES. 2013.

[52] JIMÉNEZ-RUIZ, E., GRAU, B. C., ZHOU, Y. LogMap 2.0: towards logic-based, scalable and interactive ontology matching. In: SWAT4LS. 2011.

Apêndice A. RESULTADOS DOS CENÁRIOS AVALIADOS

Tabela 0.1 – Resultados do cenário 1 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	4	1	0,56	0,70	0,47	1	0,50	0,67	0,40	1	0,50	0,67	0,40	1	0,56	0,70	0,47	1	0,52	0,75	0,40
Cmt	confOf	4	1	0,38	0,80	0,25	0				0				0				0			
Cmt	Edas	4	1	0,67	0,73	0,62	1	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62
Cmt	Ekaw	4	1	0,62	1	0,45	0				0				0				0			
Cmt	Iasted	4	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	4	1	0,75	0,75	0,75	1	0,82	0,90	0,75	1	0,82	0,90	0,75	1	0,82	0,90	0,75	1	0,86	1	0,75
Conference	confOf	4	1	0,61	0,88	0,47	1	0,61	0,88	0,47	1	0,61	0,88	0,47	1	0,61	0,88	0,47	0			
Conference	Edas	4	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41
Conference	Ekaw	4	1	0,44	0,73	0,32	1	0,44	0,73	0,32	0				0				0			
Conference	Iasted	4	1	0,42	0,80	0,29	0				0				0				0			
Conference	Sigkdd	4	1	0,65	0,58	0,73	1	0,71	0,69	0,73	1	0,72	0,90	0,60	1	0,67	0,67	0,67	1	0,69	0,71	0,67
ConfOf	Edas	4	1	0,65	0,83	0,53	1	0,60	0,82	0,47	1	0,62	0,90	0,47	0				0			
ConfOf	Ekaw	4	1	0,73	0,92	0,60	1	0,73	0,92	0,60	1	0,83	0,94	0,75	1	0,73	0,92	0,60	1	0,76	0,93	0,65
ConfOf	Iasted	4	0	0,62	1	0,44	0				0				0				0			
ConfOf	Sigkdd	4	0	0,73	1	0,57	0				0				0				0			
Edas	Ekaw	4	1	0,56	0,77	0,43	1	0,56	0,77	0,43	1	0,56	0,77	0,43	1	0,56	0,77	0,43	1	0,56	0,77	0,43

Edas	Iasted	4	1	0,52	0,88	0,37	1	0,52	0,88	0,37	0				0				0			
Edas	Sigkdd	4	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	4	1	0,75	1	0,60	1	0,75	1	0,60	0				0				0			
Ekaw	Sigkdd	4	0	0,78	1	0,64	0				0				0				0			
Iasted	Sigkdd	4	1	0,56	0,43	0,80	1	0,55	0,41	0,80	1	0,56	0,43	0,80	1	0,59	0,46	0,80	1	0,86	0,92	0,80

Tabela 0.2 – Resultados do cenário 2 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	5	1	0,45	0,71	0,33	1	0,45	0,71	0,33	1	0,45	0,71	0,33	0				0			
Cmt	confOf	5	1	0,38	0,80	0,25	0				0				0				0			
Cmt	Edas	5	1	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62
Cmt	Ekaw	5	0	0,62	1	0,45	0				0				0				0			
Cmt	Iasted	5	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	5	1	0,82	0,90	0,75	1	0,82	0,90	0,75	1	0,82	0,90	0,75	1	0,82	0,90	0,75	1	0,82	0,90	0,75
Conference	confOf	5	0	0,61	0,88	0,47	0				0				0				0			
Conference	Edas	5	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	0			
Conference	Ekaw	5	1	0,44	0,73	0,32	1	0,44	0,73	0,32	0				0				0			
Conference	Iasted	5	1	0,42	0,80	0,29	0				0				0				0			
Conference	Sigkdd	5	1	0,67	0,61	0,73	1	0,71	0,69	0,73	1	0,67	0,75	0,60	1	0,62	0,64	0,60	1	0,60	0,60	0,60
ConfOf	Edas	5	1	0,62	0,90	0,47	1	0,62	0,90	0,47	1	0,62	0,90	0,47	0				0			
ConfOf	Ekaw	5	1	0,65	0,91	0,50	1	0,80	0,93	0,70	1	0,81	0,88	0,75	1	0,81	0,88	0,75	1	0,81	0,88	0,75

ConfOf	Iasted	5	0	0,62	1	0,44	0				0				0				0			
ConfOf	Sigkdd	5	0	0,73	1	0,57	0				0				0				0			
Edas	Ekaw	5	1	0,54	0,71	0,43	1	0,56	0,77	0,43	1	0,56	0,77	0,43	0				0			
Edas	Iasted	5	1	0,52	0,88	0,37	1	0,52	0,88	0,37	1	0,52	0,88	0,37	1	0,52	0,88	0,37	0			
Edas	Sigkdd	5	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	5	1	0,36	0,24	0,70	1	0,70	0,70	0,70	1	0,78	0,88	0,70	1	0,82	1	0,70	1	0,82	1	0,70
Ekaw	Sigkdd	5	0	0,78	1	0,64	0				0				0				0			
Iasted	Sigkdd	5	1	0,65	0,55	0,80	1	0,57	0,42	0,87	1	0,61	0,45	0,93	1	0,62	0,48	0,87	1	0,62	0,48	0,87

Tabela 0.3 – Resultados do cenário 3 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	6	1	0,45	0,71	0,33	1	0,45	0,71	0,33	1	0,45	0,71	0,33	0				0			
Cmt	confOf	6	1	0,48	0,67	0,38	1	0,48	0,67	0,38	1	0,50	0,75	0,38	1	0,52	0,86	0,38	1	0,52	0,86	0,38
Cmt	Edas	6	1	0,76	1	0,62	1	0,76	1	0,62	1	0,70	1	0,54	1	0,76	1	0,62	0			
Cmt	Ekaw	6	0	0,62	1	0,45	0				0				0				0			
Cmt	Iasted	6	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	6	1	0,80	1	0,67	1	0,80	1	0,67	0				0				0			
Conference	confOf	6	1	0,61	0,88	0,47	1	0,61	0,88	0,47	1	0,61	0,88	0,47	0				0			
Conference	Edas	6	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	0				0			
Conference	Ekaw	6	1	0,44	0,73	0,32	1	0,44	0,73	0,32	1	0,44	0,73	0,32	1	0,44	0,73	0,32	1	0,44	0,73	0,32
Conference	Iasted	6	0	0,42	0,80	0,29	0				0				0				0			

Conference	Sigkdd	6	1	0,69	0,71	0,67	1	0,69	0,82	0,60	1	0,67	0,75	0,60	1	0,72	0,90	0,60	1	0,72	0,90	0,60
ConfOf	Edas	6	1	0,62	0,90	0,47	1	0,62	0,90	0,47	1	0,62	0,90	0,47	0				0			
ConfOf	Ekaw	6	1	0,76	0,93	0,65	1	0,81	0,88	0,75	1	0,81	0,88	0,75	1	0,81	0,88	0,75	1	0,81	0,88	0,75
ConfOf	Iasted	6	0	0,62	1	0,44	0				0				0				0			
ConfOf	Sigkdd	6	0	0,73	1	0,57	0				0				0				0			
Edas	Ekaw	6	1	0,58	0,73	0,48	1	0,59	0,79	0,48	1	0,59	0,79	0,48	1	0,59	0,79	0,48	1	0,59	0,79	0,48
Edas	Iasted	6	1	0,57	0,89	0,42	0				0				0				0			
Edas	Sigkdd	6	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	6	1	0,40	0,27	0,80	1	0,62	0,50	0,80	1	0,73	0,67	0,80	1	0,76	0,73	0,80	1	0,89	1	0,80
Ekaw	Sigkdd	6	0	0,78	1	0,64	0				0				0				0			
Iasted	Sigkdd	6	1	0,75	0,71	0,80	1	0,80	0,80	0,80	1	0,86	0,92	0,80	1	0,86	0,92	0,80	1	0,83	0,86	0,80

Tabela 0.4 – Resultados do cenário 4 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	7	1	0,52	0,58	0,47	1	0,48	0,60	0,40	1	0,48	0,60	0,40	1	0,50	0,67	0,40	1	0,56	0,70	0,47
Cmt	confOf	7	1	0,48	0,67	0,38	1	0,50	0,75	0,38	1	0,50	0,75	0,38	1	0,52	0,86	0,38	1	0,50	0,75	0,38
Cmt	Edas	7	1	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62
Cmt	Ekaw	7	0	0,62	1	0,45	0				0				0				0			
Cmt	Iasted	7	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	7	1	0,80	1	0,67	0				0				0				0			
Conference	confOf	7	1	0,61	0,88	0,47	1	0,61	0,88	0,47	0				0				0			

Conference	Edas	7	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	0			
Conference	Ekaw	7	1	0,49	0,75	0,36	1	0,44	0,73	0,32	1	0,44	0,73	0,32	1	0,44	0,73	0,32	1	0,44	0,73	0,32
Conference	Iasted	7	1	0,42	0,80	0,29	0				0				0				0			
Conference	Sigkdd	7	1	0,71	0,69	0,73	1	0,73	0,73	0,73	1	0,71	0,69	0,73	1	0,73	0,73	0,73	1	0,71	0,77	0,67
ConfOf	Edas	7	1	0,62	0,90	0,47	1	0,62	0,90	0,47	1	0,62	0,90	0,47	0				0			
ConfOf	Ekaw	7	1	0,73	0,92	0,60	1	0,76	0,93	0,65	1	0,78	0,88	0,70	1	0,84	0,89	0,80	0			
ConfOf	Iasted	7	1	0,71	1	0,56	0				0				0				0			
ConfOf	Sigkdd	7	1	0,73	1	0,57	0				0				0				0			
Edas	Ekaw	7	1	0,54	0,71	0,43	1	0,53	0,67	0,43	1	0,59	0,79	0,48	1	0,59	0,79	0,48	1	0,59	0,79	0,48
Edas	Iasted	7	1	0,57	0,89	0,42	0				0				0				0			
Edas	Sigkdd	7	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	7	1	0,46	0,32	0,80	1	0,41	0,28	0,80	1	0,67	0,57	0,80	1	0,89	1	0,80	1	0,89	1	0,80
Ekaw	Sigkdd	7	0	0,78	1	0,64	0				0				0				0			
Iasted	Sigkdd	7	1	0,80	0,80	0,80	1	0,86	0,92	0,80	1	0,86	0,92	0,80	1	0,86	0,92	0,80	0			

Tabela 0.5 – Resultados do cenário 5 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	8	1	0,52	0,75	0,40	1	0,44	0,50	0,40	1	0,50	0,67	0,40	1	0,50	0,67	0,40	1	0,52	0,75	0,40
Cmt	confOf	8	1	0,52	0,86	0,38	1	0,52	0,86	0,38	1	0,52	0,86	0,38	1	0,52	0,86	0,38	1	0,45	0,83	0,31
Cmt	Edas	8	1	0,76	1	0,62	1	0,76	1	0,62	0				0				0			
Cmt	Ekaw	8	0	0,62	1	0,45	0				0				0				0			

Cmt	Iasted	8	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	8	1	0,80	1	0,67	0				0				0				0			
Conference	confOf	8	1	0,61	0,88	0,47	1	0,61	0,88	0,47	1	0,61	0,88	0,47	1	0,61	0,88	0,47	0			
Conference	Edas	8	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41
Conference	Ekaw	8	1	0,44	0,73	0,32	1	0,44	0,73	0,32	1	0,49	0,75	0,36	0				0			
Conference	Iasted	8	1	0,42	0,80	0,29	0				0				0				0			
Conference	Sigkdd	8	1	0,73	0,73	0,73	1	0,73	0,73	0,73	1	0,72	0,90	0,60	1	0,72	0,90	0,60	1	0,69	0,82	0,60
ConfOf	Edas	8	1	0,62	0,90	0,47	1	0,62	0,90	0,47	0				0				0			
ConfOf	Ekaw	8	1	0,76	0,93	0,65	1	0,78	0,88	0,70	1	0,80	0,80	0,80	1	0,81	0,88	0,75	1	0,81	0,88	0,75
ConfOf	Iasted	8	1	0,71	1	0,56	1	0,71	1	0,56	0				0				0			
ConfOf	Sigkdd	8	1	0,73	1	0,57	1	0,73	1	0,57	0				0				0			
Edas	Ekaw	8	1	0,59	0,79	0,48	1	0,59	0,79	0,48	1	0,59	0,79	0,48	1	0,59	0,79	0,48	1	0,59	0,79	0,48
Edas	Iasted	8	1	0,52	0,67	0,42	1	0,57	0,89	0,42	1	0,57	0,89	0,42	1	0,57	0,89	0,42	1	0,57	0,89	0,42
Edas	Sigkdd	8	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	8	1	0,42	0,29	0,80	1	0,84	0,89	0,90	1	0,89	1	0,80	1	0,89	1	0,80	1	0,89	1	0,80
Ekaw	Sigkdd	8	0	0,78	1	0,64	0				0				0				0			
Iasted	Sigkdd	8	1	0,77	0,75	0,80	1	0,86	0,92	0,80	1	0,71	0,63	0,80	1	0,86	0,92	0,80	1	0,86	0,92	0,80

Tabela 0.6 – Resultados do cenário 6 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	4	2	0,52	0,75	0,40	2	0,52	0,75	0,40	2	0,52	0,75	0,40	2	0,52	0,75	0,40	1	0,52	0,75	0,40

Cmt	confOf	4	1	0,38	0,80	0,25	0				0				0				0			
Cmt	Edas	4	2	0,76	1	0,62	2	0,76	1	0,62	2	0,76	1	0,62	2	0,76	1	0,62	2	0,76	1	0,62
Cmt	Ekaw	4	2	0,62	1	0,45	0				0				0				0			
Cmt	Iasted	4	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	4	2	0,82	0,90	0,75	2	0,82	0,90	0,75	2	0,86	1	0,75	2	0,86	1	0,75	2	0,86	1	0,75
Conference	confOf	4	2	0,61	0,88	0,47	2	0,61	0,88	0,47	2	0,61	0,88	0,47	0				0			
Conference	Edas	4	2	0,56	0,88	0,41	2	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41
Conference	Ekaw	4	2	0,44	0,73	0,32	2	0,44	0,73	0,32	2	0,44	0,73	0,32	0				0			
Conference	Iasted	4	2	0,42	0,80	0,29	0				0				0				0			
Conference	Sigkdd	4	2	0,71	0,69	0,73	2	0,58	0,56	0,60	2	0,67	0,61	0,73	2	0,69	0,65	0,73	2	0,69	0,65	0,73
ConfOf	Edas	4	2	0,60	0,82	0,47	2	0,62	0,90	0,47	1	0,67	0,91	0,53	2	0,67	0,91	0,53	0			
ConfOf	Ekaw	4	2	0,60	0,90	0,45	2	0,78	0,88	0,70	2	0,78	0,88	0,70	2	0,78	0,88	0,70	2	0,83	0,94	0,75
ConfOf	Iasted	4	0	0,62	1	0,44	0				0				0				0			
ConfOf	Sigkdd	4	0	0,73	1	0,57	0				0				0				0			
Edas	Ekaw	4	2	0,56	0,77	0,43	2	0,56	0,77	0,43	2	0,56	0,77	0,43	2	0,56	0,77	0,43	0			
Edas	Iasted	4	2	0,52	0,88	0,37	0				0				0				0			
Edas	Sigkdd	4	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	4	2	0,75	1	0,60	1	0,75	1	0,60	0				0				0			
Ekaw	Sigkdd	4	0	0,78	1	0,64																
Iasted	Sigkdd	4	2	0,86	0,92	0,80	2	0,86	0,92	0,80	2	0,86	0,92	0,80	0				0			

Tabela 0.7 – Resultados do cenário 7 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	4	2	0,52	0,75	0,40	3	0,52	0,75	0,40	3	0,52	0,75	0,40	3	0,58	0,78	0,47	3	0,58	0,78	0,47
Cmt	confOf	4	1	0,38	0,80	0,25	0				0				0				0			
Cmt	Edas	4	3	0,76	1	0,62	3	0,73	0,89	0,62	3	0,76	1	0,62	3	0,76	1	0,62	3	0,76	1	0,62
Cmt	Ekaw	4	2	0,62	1	0,45	0				0				0				0			
Cmt	Iasted	4	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	4	3	0,82	0,90	0,75	1	0,86	1	0,75	2	0,86	1	0,75	0				0			
Conference	confOf	4	2	0,61	0,88	0,47	3	0,61	0,88	0,47	0				0				0			
Conference	Edas	4	3	0,56	0,88	0,41	1	0,56	0,88	0,41	1	0,56	0,88	0,41	0				0			
Conference	Ekaw	4	3	0,44	0,73	0,32	3	0,44	0,73	0,32	3	0,44	0,73	0,32	0				0			
Conference	Iasted	4	2	0,42	0,80	0,29	0				0				0				0			
Conference	Sigkdd	4	3	0,76	0,79	0,73	3	0,73	0,73	0,73	3	0,79	0,85	0,73	3	0,76	0,79	0,73	2	0,81	0,92	0,73
ConfOf	Edas	4	3	0,62	0,90	0,47	2	0,67	0,91	0,53	2	0,67	0,91	0,53	0				0			
ConfOf	Ekaw	4	3	0,55	0,89	0,40	3	0,76	0,82	0,70	3	0,81	0,88	0,75	3	0,87	0,89	0,85	3	0,89	0,94	0,85
ConfOf	Iasted	4	0	0,62	1	0,44	0				0				0				0			
ConfOf	Sigkdd	4	0	0,73	1	0,57	0				0				0				0			
Edas	Ekaw	4	3	0,56	0,77	0,43	3	0,56	0,77	0,43	3	0,56	0,77	0,43	2	0,56	0,77	0,43	0			
Edas	Iasted	4	3	0,52	0,88	0,37	0				0				0				0			
Edas	Sigkdd	4	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	4	3	0,75	1	0,60	1	0,75	1	0,60	0				0				0			

Ekaw	Sigkdd	4	0	0,78	1	0,64																
Iasted	Sigkdd	4	3	0,86	0,92	0,80	1	0,86	0,92	0,80	2	0,86	0,92	0,80	0				0			

Tabela 0.8 – Resultados do cenário 8 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	8	2	0,46	0,55	0,40	2	0,50	0,67	0,40	2	0,52	0,75	0,40	1	0,52	0,75	0,40	1	0,58	0,78	0,47
Cmt	confOf	8	2	0,52	0,86	0,38	2	0,45	0,83	0,31	2	0,45	0,83	0,31	2	0,45	0,83	0,31	0			
Cmt	Edas	8	2	0,76	1	0,62	0				0				0				0			
Cmt	Ekaw	8	0	0,62	1	0,45	0				0				0				0			
Cmt	Iasted	8	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	8	1	0,80	1	0,67	0				0				0				0			
Conference	confOf	8	2	0,61	0,88	0,47	1	0,61	0,88	0,47	1	0,61	0,88	0,47	0				0			
Conference	Edas	8	2	0,56	0,88	0,41	2	0,56	0,88	0,41	2	0,56	0,88	0,41	0				0			
Conference	Ekaw	8	2	0,44	0,73	0,32	2	0,44	0,73	0,32	0				0				0			
Conference	Iasted	8	2	0,42	0,80	0,29	0				0				0				0			
Conference	Sigkdd	8	2	0,69	0,71	0,67	2	0,69	0,82	0,60	2	0,72	0,90	0,60	2	0,72	0,90	0,60	1	0,72	0,90	0,60
ConfOf	Edas	8	2	0,67	0,91	0,53	0				0				0				0			
ConfOf	Ekaw	8	2	0,76	0,93	0,65	2	0,76	0,93	0,65	2	0,73	0,92	0,60	2	0,81	0,88	0,75	2	0,86	0,94	0,80
ConfOf	Iasted	8	1	0,71	1	0,56	2	0,71	1	0,56	0				0				0			
ConfOf	Sigkdd	8	2	0,73	1	0,57	0				0				0				0			
Edas	Ekaw	8	2	0,58	0,73	0,48	2	0,59	0,79	0,48	2	0,59	0,79	0,48	0				0			

Edas	Iasted	8	2	0,52	0,67	0,42	2	0,57	0,89	0,42	1	0,57	0,89	0,42	1	0,57	0,89	0,42	0			
Edas	Sigkdd	8	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	8	2	0,84	0,89	0,80	2	0,89	1	0,80	2	0,89	1	0,80	2	0,89	1	0,80	1	0,89	1	0,80
Ekaw	Sigkdd	8	0	0,78	1	0,64	0				0				0				0			
Iasted	Sigkdd	8	2	0,86	0,92	0,80	2	0,86	0,92	0,80	0				0				0			

Tabela 0.9 – Resultados do cenário 9 em termos de precisão (P), cobertura (C) e medida-F (F), no *Conference data set*

Ontologias		**Iteração 1**					**Iteração 2**				**Iteração 3**				**Iteração 4**				**Iteração 5**			
O	**O'**	***k***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***	***w***	***F***	***P***	***C***
Cmt	Conference	8	3	0,56	0,70	0,47	3	0,59	0,67	0,53	3	0,58	0,78	0,47	3	0,62	0,73	0,53	3	0,64	0,80	0,53
Cmt	confOf	8	3	0,52	0,86	0,38	3	0,45	0,83	0,31	3	0,45	0,83	0,31	3	0,52	0,86	0,38	1	0,52	0,86	0,38
Cmt	Edas	8	3	0,76	1	0,62	1	0,76	1	0,62	1	0,76	1	0,62	0				0			
Cmt	Ekaw	8	0	0,62	1	0,45	0				0				0				0			
Cmt	Iasted	8	0	0,89	0,80	1	0				0				0				0			
Cmt	Sigkdd	8	1	0,80	1	0,67	0				0				0				0			
Conference	confOf	8	2	0,61	0,88	0,47	1	0,61	0,88	0,47	1	0,61	0,88	0,47	0				0			
Conference	Edas	8	3	0,56	0,88	0,41	3	0,56	0,88	0,41	0				0				0			
Conference	Ekaw	8	3	0,44	0,73	0,32	1	0,44	0,73	0,32	0				0				0			
Conference	Iasted	8	2	0,42	0,80	0,29	0				0				0				0			
Conference	Sigkdd	8	3	0,69	0,71	0,67	3	0,76	0,79	0,73	3	0,81	0,92	0,73	2	0,81	0,92	0,73	1	0,81	0,92	0,73
ConfOf	Edas	8	2	0,67	0,91	0,53	0				0				0				0			
ConfOf	Ekaw	8	3	0,81	0,88	0,75	3	0,83	0,94	0,75	3	0,83	0,94	0,75	3	0,83	0,94	0,75	1	0,83	0,94	0,75

ConfOf	Iasted	8	1	0,71	1	0,56	2	0,71	1	0,56	0				0				0			
ConfOf	Sigkdd	8	3	0,73	1	0,57	0				0				0				0			
Edas	Ekaw	8	3	0,59	0,79	0,48	3	0,59	0,79	0,48	0				0				0			
Edas	Iasted	8	3	0,53	0,73	0,42	3	0,57	0,89	0,42	2	0,57	0,89	0,42	1	0,57	0,89	0,42	0			
Edas	Sigkdd	8	0	0,64	1	0,47	0				0				0				0			
Ekaw	Iasted	8	3	0,80	0,80	0,80	3	0,89	1	0,80	3	0,89	1	0,80	3	0,89	1	0,80	0			
Ekaw	Sigkdd	8	0	0,78	1	0,64	0				0				0				0			
Iasted	Sigkdd	8	3	0,86	0,92	0,80	1	0,86	0,92	0,80	1	0,86	0,92	0,80	0				0			