Pós-Graduação em Ciência da Computação

Gabrielle Karine Canalle

UMA ESTRATÉGIA PARA SELEÇÃO DE ATRIBUTOS RELEVANTES NO PROCESSO DE RESOLUÇÃO DE ENTIDADES

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
<www.cin.ufpe.br/~posgraduacao>

RECIFE
2016

Gabrielle Karine Canalle

UMA ESTRATÉGIA PARA SELEÇÃO DE ATRIBUTOS RELEVANTES NO PROCESSO DE RESOLUÇÃO DE ENTIDADES

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: *Ana Carolina Salgado*
Co-Orientadora: *Bernadette Farias Lóscio*

RECIFE
2016

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

C212e Canalle, Gabrielle Karine
Uma estratégia para seleção de atributos relevantes no processo de resolução de entidades / Gabrielle Karine Canalle. – 2016.
87 f.: il., fig., tab.

Orientadora: Ana Carolina Salgado.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2016.
Inclui referências.

1. Banco de dados. 2. Integração de dados. 3. Resolução de entidades. I. Salgado, Ana Carolina (orientadora). II. Título.

025.04 CDD (23. ed.) UFPE- MEI 2016-145

Gabrielle Karine Canalle

Uma Estratégia para Seleção de Atributos Relevantes no Processo de Resolução de Entidades

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 22/08/2016

_____

Orientador: Prof. Dr. Ana Carolina Brandão Salgado

BANCA EXAMINADORA

_____

Prof. Dr. Luciano de Andrade Barbosa
Centro de Informática / UFPE

_____

Prof. Dr. Carlos Eduardo Santos Pires
Departamento de Sistemas e Computação/UFCG

_____

Profa. Dra. Bernadette Farias Lóscio
Centro de Informática / UFPE
(**Co-Orientadora**)

Dedico este trabalho à minha mãe Marinez, meu pai Valério, meu irmão Gabriel, minha madrinha Lucirene, amigos, e professores que de alguma maneira contribuíram para sua realização.

Agradecimentos

Agradeço primeiramente a Deus por sempre estar guiando meus passos, me possibilitando realizar meus sonhos e descobrir os rumos do destino que ele tem reservado à mim.

Agradeço à minha mãe, que passou noites acordada para que eu dormisse e sonhasse, que sempre me apoiou e acreditou em mim, e que sofreu de saudade quando, para realizar meus sonhos, precisei tomar a difícil decisão de ficar longe. Pelo carinho, amor e cuidado que tem por mim, e pela paciência para me acalmar nos momentos de desespero, mesmo à distância. Ao meu pai, que, com seu infinito cuidado, me trouxe para iniciar minha caminhada em busca desse sonho. Obrigada pai! Por me apoiar incondicionalmente e fazer o possível para realizar todos os meus sonhos. Ao meu irmão Gabriel, que quando chegou me despertou um sentimento que eu não conhecia. Sentimento gigante, o amor de irmão. Agradeço a toda minha família, minha base, por todo apoio, e pela compreensão em não me ter por perto. À minha madrinha-mãe, Lu, pelas companhia via Skype, pelos conselhos e por tudo que já fez por mim até hoje. À minha querida tia Nena, que tanto amo. Família, se cheguei até aqui, é por vocês! Amo vocês!

Ao meu companheiro e namorado, Renã, por todas as vezes que teve que aguentar minhas lamúrias, falta de paciência, minhas crises de "O que vai ser depois do mestrado?", e por todas as vezes que me incentivou a não desistir e a buscar meus objetivos. Por estar ao meu lado nos momentos bons e também nos ruins, e não me abandonar nunca. Obrigada por todo cuidado e amor!

Às minhas orientadoras Ana Carolina Salgado e Bernadette Farias Lóscio. Obrigada por acreditarem no meu potencial, por me estimularem a ir além, e sempre me afirmarem diante da minha dúvida: "Sim Gabi, isso dá um mestrado". Meus mais sinceros agradecimentos a vocês, por acolherem carinhosamente essa Paranaense desconhecida. Registro aqui minha profunda admiração pelas pessoas e profissionais que vocês são. Se depender de mim, não largo vocês tão cedo.

Não posso deixar de agradecer meu orientador da graduação e amigo, professor Everton Araújo. Se não fosse você me incentivando a buscar esse caminho, nada disso seria possível. Obrigada professor, por me aconselhar a confiar mais na minha capacidade e por sempre acreditar que posso mais. Meu sincero agradecimento.

Agradeço ao meu amigo Neto, que desde que cheguei aqui em Recife nunca hesitou em me auxiliar no que precisei. Sempre esteve ao meu lado, sorrindo ou chorando (ou em desespero), tanto para me consolar quanto para comemorar comigo. Amigo que levo do mestrado para a vida. Obrigada migo!

Aos meus amigos da "Toca dos Dados": Marcelo, Diego, Priscilla e Danusa, os que mais convivi. Em especial à Marcelo e sua esposa, Vanessa, por terem compartilhado comigo sua casa, e principalmente o Maguila e a Nina. Levo em meu coração a gratidão e a amizade que

tenho por vocês. Diego, obrigada por toda ajuda. Priscilla e Danusa, adorei conhecer melhor vocês. São especiais pra mim. Obrigada pela bela amizade.

Quero agradecer também às minhas amigas Suilan e Henriete. À Suilan por sempre me colocar em suas orações, por me acolher como uma filha e me emprestar sua família, e por estar sempre por perto. À Henriete, minha amiga desde a graduação, que mesmo longe não deixou de me apoiar e me ouvir quando precisei. Obrigada amiga, pela amizade e apoio.

Agradeço ainda à todos aqueles que não mencionei aqui, mas que de alguma forma contribuíram para realização deste trabalho. Obrigada!

É muito melhor lançar-se em busca de conquistas grandiosas, mesmo expondo-se ao fracasso, do que alinhar-se com os pobres de espírito, que nem gozam muito nem sofrem muito, porque vivem numa penumbra cinzenta onde não conhecem nem vitória, nem derrota.

—THEODORE ROOSEVELT

Resumo

Integração de Dados é um processo essencial quando deseja-se obter uma visão unificada de dados armazenados em fontes de dados autônomas, heterogêneas e distribuídas. Uma etapa crucial desse processo é a Resolução de Entidades, que consiste em identificar instâncias que se referem à mesma entidade do mundo real. A Resolução de Entidades se subdivide em várias fases, incluindo uma fase de comparação entre pares de instâncias. Nesta fase, são utilizadas funções que avaliam a similaridade entre os valores dos atributos que descrevem as instâncias. É importante notar que a qualidade do resultado do processo de Resolução de Entidades é diretamente afetada pelo conjunto de atributos selecionados para a fase de comparação de instâncias. Contudo, selecionar tais atributos pode ser um grande desafio, devido ao grande número de atributos que descrevem as instâncias ou à baixa relevância de alguns atributos para o processo de Resolução de Entidades. Na literatura existem alguns trabalhos que abordam esse problema. Em sua maioria, as abordagens propostas para seleção de atributos utilizam aprendizagem de máquina. No entanto, além da necessidade de um conjunto de treinamento, cuja definição é uma tarefa difícil, principalmente em cenários de grandes volumes de dados, a aprendizagem de máquina é um processo custoso. Neste contexto, este trabalho propõe uma estratégia para seleção de atributos relevantes a serem considerados na fase de comparação de instâncias do processo de Resolução de Entidades. A estratégia proposta considera critérios relacionados aos dados, tais como a densidade e repetição de valores de cada atributo, e critérios relacionados às fontes, tal como a confiabilidade, para avaliar a relevância de um atributo para a fase de comparação de instâncias. Um atributo é considerado relevante se contribui positivamente para a identificação de correspondências verdadeiras, e irrelevante se contribui na identificação de correspondências erradas (falsos positivos e falsos negativos). Em experimentos realizados, utilizando a estratégia proposta, foi possível alcançar bons resultados na comparação de instâncias do processo de Resolução de Entidades, ou seja, os atributos dados como relevantes foram aqueles que contribuíram para encontrar o maior número de correspondências verdadeiras, com o menor número de correspondências erradas.

Palavras-chave: Integração de Dados. Resolução de Entidades. Seleção de Atributos

Abstract

Data integration is an essential task for achieving a unified view of data stored in autonomous, heterogeneous and distributed sources. A key step in this process is Entity Resolution, which consists of identifying instances that refer to the same real-world entity. Entity Resolution can be subdivided into several stages, including a comparison step between instance pairs. In this step, functions that check the similarity between values of attributes are used to discover equivalent instances. It is important to note that the quality of the result of the entity resolution process is directly affected by the set of selected attributes used to compare the instances. However, selecting such attributes can be challenging, due to either the large number of attributes that describes an instance or to the low relevance of some attributes regarding to the entity resolution process. In the literature, there are some approaches that investigated this problem. Most of them employ machine learning techniques for selecting relevant attributes. Usually, these techniques are computationally costly and also have the necessity of defining a training set, which requirements are non-trivial, mainly in large volumes of data scenarios. In this context, this work proposes a strategy for selecting relevant attributes to be considered in the instance comparison phase of the process of Entity Resolution. The proposed strategy considers criteria related to data, such as density and repetition of values of each attribute, and related to sources, such as reliability, to evaluate the relevance of the attributes. An attribute is considered relevant if contributes positively for the identification of true matches, and irrelevant if contributes for the identification of incorrect matches (false positives and false negatives). In our experiments, the proposed strategy achieved good results for the Entity Resolution process. That is, the attributes classified as relevant were the ones that contributed to find the greatest number of true matches with a few incorrect matches.

Keywords: Data Integration. Entity Resolution. Attribute Selection

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

CSV	Comma-separated values	69
FS	Fellegi-Sunter	43
GAV	Global-as-view	22
GLAV	Global-local-as-view	22
GNU	General public license	69
JSON	JavaScript Object Notation	69
LAV	Local-as-view	22
MD	Matching Dependencies	41
OWL	Web Ontology Language	21
RCK	Relative Candidate Keys	41
SN	Sorted Neighborhood	43
TF-IDF	Term Frequency-Inverse Document Frequency	24
UDD	Unsupervised Duplicate Detection	39
UP	Universidade de Potsdam	69
WCSS	Weighted Component Similarity Summing	39
XML	EXtensible Markup Language	21

Sumário

1 Introdução

Neste capítulo apresentamos uma introdução sobre o problema de Integração de Dados, com foco na etapa de Resolução de Entidades, onde a estratégia de seleção de atributos proposta neste trabalho está inserida. Demonstramos a necessidade de estratégias de seleção de atributos na Resolução de Entidades por meio de um exemplo. Nosso trabalho propõe uma estratégia de seleção de atributos relevantes para o processo de Resolução de Entidades, e para isso avalia critérios de qualidade relacionados aos dados e às fontes. Nossa estratégia visa que, ao final do processo de Resolução de Entidades, seja encontrado o maior número possível de correspondências verdadeiras com o menor número de correspondências erradas. Inicialmente, na Seção 1.1 apresentamos uma justificativa e motivação para este trabalho. Na Seção 1.2 é descrita a caracterização do problema. A Seção 1.3 apresenta os objetivos deste trabalho. Por fim, a estrutura da dissertação é vista na Seção 1.4.

1.1 Motivação

A crescente facilidade de geração e compartilhamento de dados tem contribuído para um crescimento acelerado no volume de dados disponíveis em meio digital. Entretanto, esse crescimento tem ocorrido de forma descontrolada, de tal forma que muitos dados contêm valores errôneos, ausentes ou duplicados, o que pode dificultar a sua utilização. Apesar disso, existe uma demanda cada vez maior por soluções de integração de dados distribuídos em fontes de dados distintas. Exemplos disso são os Web Sites de comparação de preços[1], como o Buscapé[2] e o BondFaro[3].

As soluções de integração de dados visam combinar dados residentes em diferentes fontes provendo aos usuários uma visão unificada desses dados (LENZERINI, 2002). Uma etapa importante no processo de Integração de Dados é a de Resolução de Entidades (CHRISTEN, 2012), que busca identificar a equivalência entre instâncias que representam uma mesma entidade do mundo real (DONG; SRIVASTAVA, 2015).

[1] Esses sites combinam dados oriundos de mais de 500 lojas.
[2] http://www.buscape.com.br
[3] http://www.bondfaro.com.br

A Resolução de Entidades é composta por três etapas principais, incluindo uma etapa de comparação entre pares de instâncias. Durante essa etapa, é avaliada a similaridade entre os valores dos atributos que descrevem as instâncias que estão sendo comparadas.

Um dos principais desafios a serem enfrentados durante o processo de Resolução de Entidades diz respeito à escolha dos atributos a serem utilizados na etapa de comparação, já que a similaridade calculada entre as instâncias depende diretamente dos atributos que serão considerados. Sendo assim, a qualidade do resultado do processo de Resolução de Entidades é diretamente afetada pela relevância dos atributos selecionados para a etapa de comparação de instâncias.

Pode-se pensar que quanto maior o número de atributos considerados melhor será o resultado do processo, já que estaremos considerando o máximo de informação possível, mas atributos com baixa relevância não trazem ganho para o processo de resolução, e por vezes podem até diminuir a qualidade do resultado final. Um exemplo de atributo com baixa relevância para a comparação, é um atributo que contém um valor similar para a maioria das instâncias. Notoriamente, o atributo não contribuiria para o processo, podendo ainda prejudicar o desempenho do mesmo.

Geralmente, a escolha dos atributos é realizada de forma manual (WANG et al., 2011; KöPCKE; RAHM, 2008). Entretanto, uma grande quantidade de atributos, associada à falta de conhecimento prévio do domínio das fontes, pode fazer com que atributos que não contribuam de forma eficiente para a Resolução de Entidades sejam considerados.

Uma vez que a qualidade do resultado da Resolução de Entidades é diretamente afetada pelos atributos selecionados, torna-se fundamental o uso de estratégias capazes de selecionar os melhores atributos para a comparação entre instâncias.

1.2 Caracterização do Problema

Nosso trabalho está inserido em um cenário de Integração de múltiplas fontes de dados e, mais especificamente, na etapa de Resolução de Entidades. Para ilustrar esse cenário, e a necessidade de selecionar os melhores atributos a serem considerados na fase de comparação do processo de Resolução de Entidades, e como esses atributos impactam na qualidade do processo, suponha o exemplo abaixo.

Considere um serviço de biblioteca digital online, no domínio de Ciências da Computação, que integra dados de múltiplas fontes, tais como CiteSeerX[4] *e DBLP*[5]*, e possibilita a realização de pesquisas por título, autor, ou palavra-chave. Suponha que um usuário esteja interessado em buscar artigos sobre "Integração de Dados". O serviço de integração submete a consulta para as fontes de dados CiteSeerX e DBLP, e obtém um conjunto de artigos. Uma pequena fração desse resultado pode ser vista na Tabela 1.1.*

[4] http://citeseerx.ist.psu.edu
[5] http://dblp.uni-trier.de

Tabela 1.1: Resultado de uma Consulta nas Fontes de Dados CiteSeerX e DBLP.

ID Paper	ID	Fonte	Author	Title	Year	Venue	Pages
1	1	CiteSeerX	M. Lanzerini	Data Integration: A Theorical Perspective (2002)	2002	Symposium on Principles of Database Systems	NULL
	2	DBLP	Maurizio Lanzerini	Data Integration: A Theorical Perspective	2002	PODS 2002	233-246
2	3	DBLP	Guy Pierra	The PLIB ontology-based approach to data integration	2004	IFIP Congress Topical Sessions	13-18
3	4	CiteSeerX	Patrick Ziegler and Klaus R. Dittrich	Three decades of data integration - all problems solved	NULL	In 18th IFIP Computer Congress (WCC)	NULL
	5	DBLP	Patrick Ziegler and Klaus R. Dittrich	Three decades of data integration - All problems solved?	2004	IFIP Congress Topical Sessions	NULL

Neste exemplo, são apresentadas cinco instâncias, cada uma contendo um identificador próprio (*ID*). Essas instâncias são referentes a três artigos diferentes identificados pela coluna *ID Paper*. A fim de retornar o resultado integrado ao usuário, a Resolução de Entidades deve ser realizada. Para isso, na fase de comparação, os atributos que descrevem as instâncias são comparados. Como geralmente a seleção de atributos é feita de forma manual, ou considerando todos os atributos na fase de comparação, vamos supor as duas situações.

Suponha que a Resolução de Entidades será realizada considerando todos os atributos que descrevem as instâncias. Possivelmente, as instâncias 1 e 2 seriam dadas como não duplicadas, já que dos cinco atributos considerados, dois (*Venue* e *Pages*) possuem uma similaridade igual a 0. O mesmo aconteceria com as instâncias 4 e 5, em que os atributos *Year* e *Pages* também possuem similaridade igual a 0. Com isso, podemos observar que atributos contendo valores nulos afetam negativamente o processo de Resolução de Entidades. Isso acontece porque, em alguns algoritmos de Resolução de Entidades, um valor nulo na comparação ocasiona em uma similaridade igual a 0, podendo fazer com que duas instâncias sejam dadas como distintas mesmo sendo correspondentes. Sendo assim, a comparação considerando todos os atributos resultaria em correspondências erradas chamadas de Falso Negativo, já que são instâncias duplicadas dos Artigos 1 e 3 respectivamente, e que, possivelmente, o algoritmo de Resolução de Entidades consideraria como não correspondentes.

Agora, considere que um subconjunto de atributos foi selecionado aleatoriamente, sem considerar os valores dos atributos. Sendo o subconjunto composto pelos atributos *Year* e *Venue*, provavelmente, as instâncias 3 e 5 seriam consideradas duplicadas, já que os valores desses atributos possuem uma alta similaridade. Dessa forma, pode-se observar que atributos com valores repetidos também afetam negativamente o processo de Resolução de Entidades, já que um valor repetido na comparação ocasiona uma alta similaridade, o que pode fazer com que duas instâncias sejam dadas como correspondentes mesmo sendo distintas. Isso ocasionaria em uma correspondência errada, chamada de Falso Positivo, que se refere a um par de instâncias não correspondentes que foram dadas como correspondentes pelo algoritmo de Resolução de Entidades. Por outro lado, as instâncias 1 e 2, e 4 e 5 seriam dadas como não correspondentes, resultando em dois Falsos Negativos.

Podemos notar que utilizando todos os atributos, ou um subconjunto de atributos selecionado aleatoriamente, o resultado obtido por meio do processo de Resolução de Entidades

não seria tão satisfatório. Devido à isso, faz-se necessária uma estratégia de seleção de atributos em que somente os atributos relevantes para o processo de Resolução de Entidades sejam selecionados.

Dessa forma, este trabalho propõe uma estratégia que automatiza o processo de Seleção de atributos a fim de facilitar a identificação dos atributos a serem usados na comparação de instâncias. A seleção de atributos proposta considera critérios relacionados aos dados, tais como a densidade e repetição de valores de cada atributo, e critérios relacionados às fontes, tal como a confiabilidade, para avaliar a relevância de um atributo para a fase de comparação de instâncias. Um atributo é considerado relevante se contribui positivamente para a identificação de correspondências verdadeiras, e irrelevante se contribui na identificação de correspondências erradas (falsos positivos e falsos negativos). Com o uso da estratégia proposta, espera-se que ao final do processo de Resolução de Entidades seja obtido o maior número possível de correspondências verdadeiras com o menor número de correspondências erradas.

1.3 Objetivos

Este trabalho tem como principal objetivo propor uma estratégia de seleção de atributos relevantes para serem utilizados na fase de comparação de pares de instâncias do processo de Resolução de Entidades. A estratégia proposta utiliza dois tipos de critérios para avaliação da relevância dos atributos: i) critérios relacionados aos dados, para o cálculo da relevância individual e ii) critérios relacionados às fontes, para o cálculo da relevância global. Para isso, foram estabelecidos alguns objetivos específicos. São eles:

- Definição de um processo com o detalhamento das atividades da estratégia proposta;
- Definição dos critérios e das métricas para cálculo da relevância individual e global dos atributos;
- Formalização do problema de Seleção de Atributos para o processo de Resolução de Entidades;
- Implementação dos algoritmos de análise de relevância individual e global dos atributos;
- Realização de experimentos para avaliar a estratégia proposta.

1.4 Estrutura da Dissertação

A seguir, apresentamos a estrutura deste trabalho:

- O Capítulo 2 apresenta a Fundamentação Teórica referente aos conceitos relevantes para o desenvolvimento desta dissertação. Serão descritos assuntos referentes a

Integração de Dados com foco na etapa de Resolução de Entidades, e Seleção de Atributos.

- O Capítulo 3 apresenta os principais trabalhos relacionados, e uma tabela comparativa onde são elencadas as principais características e diferenciais de cada trabalho.
- O Capítulo 4 descreve a estratégia proposta para seleção de atributos no processo de Resolução de Entidades, apresentando os critérios definidos para avaliar a relevância dos atributos, e as fórmulas propostas para o cálculo de relevância individual e global dos atributos.
- O Capítulo 5 destaca a implementação da estratégia proposta. Também é apresentado um estudo de caso, no qual a estratégia proposta foi aplicada para analisar seu comportamento. Os experimentos e resultados obtidos são apresentados, e por fim, algumas discussões acerca dos resultados são realizadas.
- O Capítulo 6 apresenta algumas considerações sobre a pesquisa, as conclusões e os trabalhos futuros.

2

Fundamentação Teórica

Neste capítulo, é apresentada uma revisão bibliográfica acerca dos assuntos relacionados à estratégia proposta neste trabalho. Para isto, na Seção 2.1 é discutido o problema de Integração de Dados. O foco principal é na etapa de Resolução de Entidades (Subseção 2.1.2), detalhando as fases dessa etapa, métricas de avaliação de qualidade do processo de Resolução de Entidades e os conceitos relacionados. Em seguida, a Seção 2.2 descreve o processo de seleção de atributos na mineração de dados e seus objetivos (Subseção 2.2.1), e a seleção de atributos na Resolução de Entidades (Subseção 2.2.2). Por fim, na seção 2.3 são apresentadas as conclusões deste capítulo

2.1 Integração de Dados

O surgimento da Internet propiciou um grande crescimento nas bases de dados distribuídas, heterogêneas e autônomas. Publicar e acessar informações se tornaram processos simples, resultando em um crescimento acelerado no volume de dados disponíveis em meio digital. Esse crescimento ocorreu de forma desestruturada, causando dificuldade na recuperação dessas informações, e fazendo com que o usuário nem sempre obtenha uma resposta satisfatória em suas buscas. Além disso, uma grande porção desses dados contém valores errôneos, ausentes ou duplicados. Com isso, a preocupação e necessidade por soluções de Integração de Dados se torna cada vez maior.

O problema de Integração de Dados tem sido reconhecido como criticamente importante e demandado muitas pesquisas por mais de uma década (SHETH; LARSON, 1990; ABITEBOUL et al., 2003; HALEVY, 2003). A Integração de Dados já foi muito discutida no contexto de bancos de dados convencionais, e consecutivamente ampliada, abordando também o contexto da Web, onde é comum serem encontrados dados disponibilizados em várias fontes de dados distintas, que podem ser banco de dados relacionais, orientado a objetos, planilhas, páginas *Web*, entre outros.

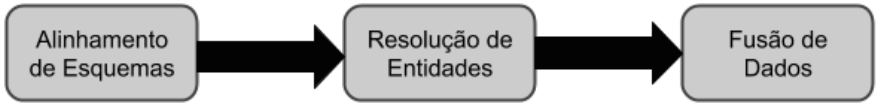
O problema consiste em criar uma visão unificada dessas fontes de dados, para que o usuário, ao submeter consultas a essa visão, receba respostas satisfatórias, que englobem dados oriundos de todas as fontes capazes de responderem as consultas.

Buscando esse objetivo, várias pesquisas e soluções de integração tem sido propostas na literatura como ferramentas para oferecer acesso uniforme a dados distribuídos em fontes heterogêneas e autônomas. (DRAPER; HALEVY; WELD, 2001; IVES et al., 1999; PATRAO et al., 2013) A principal tarefa de um sistema de Integração de Dados é fornecer uma interface uniforme para responder consultas que normalmente requerem extração e combinação de dados originários de múltiplas fontes distintas (LEVY, 1999).

Segundo Lenzerini (2002), "A Integração de Dados é o processo de combinar os dados residentes em diferentes fontes provendo aos usuários uma visão unificada desses dados", ou seja, o objetivo é liberar o usuário de ter que localizar as fontes e manipular cada uma isoladamente, combinando manualmente os dados vindos dessas diversas fontes.

O grande número de pesquisas e soluções propostas para o problema de integração pode induzir que essa é uma área de pesquisa já resolvida, mas pelo contrário, é uma área cada vez mais importante e onde não existe uma solução geral que seja adequada ou que se ajuste aos diversos problemas de integração, o que pode-se constatar pela constante pesquisa e surgimento de novas propostas.

Figura 2.1: Etapas da Integração de Dados.



Adaptado de: Dong e Srivastava (2015)

Segundo Dong e Srivastava (2015), a Integração de Dados é um processo complexo e consiste em três etapas principais, como mostra a Figura 2.1. A etapa de Alinhamento de Esquemas é responsável por encontrar correspondências entre os elementos semanticamente correspondentes dos esquemas participantes do processo, e a partir dessas correspondências gerar mapeamentos entre esses elementos. A partir desses mapeamentos, a etapa de Resolução de Entidades tem como objetivo encontrar instâncias que se referem a uma mesma entidade do mundo real. Por fim, sabendo as instâncias que correspondem a uma mesma entidade, a Fusão de Dados é responsável por criar uma representação única para cada entidade do mundo real.

A seguir, abordaremos cada etapa, detalhando seus objetivos, com maior foco na etapa de Resolução de Entidades.

2.1.1 Alinhamento de Esquemas

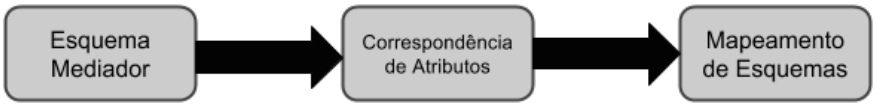
Um esquema é uma estrutura formal que representa um artefato projetado, como um esquema relacional, um esquema *EXtensible Markup Language (XML)*, um esquema representado em *Web Ontology Language (OWL)*, entre outros. A correspondência é a relação entre um ou mais elementos de um esquema e um ou mais elementos do outro (BERNSTEIN; MADHAVAN; RAHM, 2011).

O alinhamento de esquemas, segundo Rahm e Bernstein (2001), pode ser definido como: dados dois ou mais esquemas como entrada o resultado irá produzir um mapeamento entre elementos dos esquemas que se correspondem semanticamente.

O objetivo principal do alinhamento de esquemas é tornar os esquemas compatíveis para o processo de Integração de Dados, identificando e resolvendo conflitos entre esquemas de fontes de dados heterogêneas.

A abordagem tradicional de alinhamento de esquemas é dividida em três fases: criação de um esquema mediador, correspondência de atributos e mapeamento de esquemas (DONG; SRIVASTAVA, 2015), como mostra Figura 2.2.

Figura 2.2: Etapas do Alinhamento de Esquemas.



Adaptado de: Dong e Srivastava (2015)

Na primeira etapa do alinhamento de esquemas um esquema mediador é criado para fornecer uma visão unificada e virtual das fontes e capturar os aspectos mais evidentes do domínio que está sendo considerado. Muitas vezes a criação do esquema mediador é feita manualmente.

Em seguida, os atributos em cada esquema de origem são correspondidos com os atributos correspondentes no esquema mediador. Para isso, muitas técnicas foram propostas. (RAHM; BERNSTEIN, 2001; BELLAHSENE; BONIFATI; RAHM, 2011)

Por fim, na última etapa, de acordo com as correspondências encontradas entre os atributos na etapa anterior, são criados os mapeamentos entre cada esquema de origem e o esquema mediador. Esses mapeamentos especificam as relações semânticas entre os elementos dos esquemas (DONG; SRIVASTAVA, 2015).

Existem três tipos de mapeamentos de esquema: *Global-as-view (GAV)*, *Local-as-view (LAV)* e *Global-local-as-view (GLAV)*. No GAV, para cada componente do esquema global é escrita uma consulta sobre os esquemas locais. No LAV, ao invés de escrever consultas que definem como as entidades do esquema global são obtidas, são definidas consultas que descrevem como obter a extensão das fontes de dados a partir do esquema global. E o GLAV é uma junção entre os dois primeiros. Foram propostas ferramentas semi-automáticas para criação dos mapeamentos, de acordo com os resultados da correspondência de atributos (FAGIN et al., 2009).

Para simplificar o processo de alinhamento de esquemas, algumas ferramentas semi-automáticas foram propostas na literatura como, a Automed (JASPER et al., 2003), e a Clio (FAGIN et al., 2009).

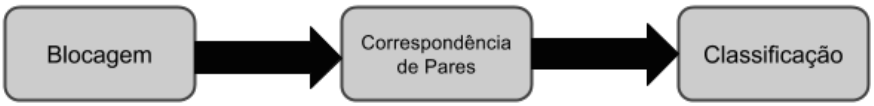
2.1.2 Resolução de Entidades

O problema de Resolução de Entidades atraiu a atenção de pesquisadores de diferentes áreas, incluindo Bancos de Dados, Mineração de Dados, Inteligência Artificial e Processamento de Linguagem Natural, e foi estudada de forma independente nesses diferentes domínios. Devido a isso, foram atribuídos diversos nomes ao problema de Resolução de Entidades, tais como "*Record Linkage*" (GU et al., 2003), "*Merge/Purge*" (HERNÁNDEZ; STOLFO, 1998), "*Reference Conciliation*" (DONG; HALEVY; MADHAVAN, 2005), "*Entity Resolution*" (BHATTACHARYA; GETOOR, 2007) ou "*Duplicate Detection*" (DRAISBACH; NAUMANN, 2010).

O objetivo da Resolução de Entidades é identificar múltiplas instâncias referentes ao mesmo objeto do mundo real (CHEN et al., 2012; DONG; SRIVASTAVA, 2015). Inicialmente, técnicas de blocagem são utilizadas, afim de evitar o produto cartesiano das comparações. Depois, a identificação de correspondências entre instâncias é realizada a partir da comparação entre os valores dos atributos que as descrevem. Depois, os algoritmos de clusterização/classificação agrupam as instâncias duplicadas de acordo com as similaridades calculadas. No caso ideal, as instâncias que estão no mesmo grupo são referentes à mesma entidade do mundo real.

O processo de Resolução de Entidades pode ser classificado em dois tipos, determinístico e probabilístico. Na Resolução de Entidades determinística, um critério é aplicado a cada par de instâncias, se o par satisfizer o critério é duplicado, senão não é. Existe ainda uma terceira classificação, onde um par de instâncias é dado como potencialmente correspondente. Pares de instâncias classificados como potencialmente correspondentes precisam passar por uma revisão manual, que classifica o por como correspondente ou não. Na Resolução de Entidades probabilística um algoritmo calcula um peso para cada par de instâncias, e esse peso reflete a probabilidade de os registros serem duplicados ou não.

Figura 2.3: Etapas da Resolução de Entidades.



Adaptado de: Dong e Srivastava (2015)

A Resolução de Entidades consiste de três etapas principais, que são: blocagem, correspondência de pares e classificação, como mostra a Figura 2.3. As etapas da Resolução de Entidades serão detalhadas a seguir.

2.1.2.1 Blocagem

A identificação de instâncias similares é idealmente realizada por meio de uma comparação de cada um dos registros contra todos os demais, o que se torna inviável quando o número de registros é muito grande (CHRISTEN; GOISER, 2007). Visando solucionar este problema, métodos de blocagem de dados têm sido desenvolvidos (MCCALLUM; NIGAM; UNGAR,

2000; COHEN; RAVIKUMAR; FIENBERG, 2003; BILENKO; KAMATH; MOONEY, 2006; EVANGELISTA et al., 2010).

Se a comparação entre instâncias for realizada sem qualquer técnica de blocagem, uma instância é comparada a todas as outras, gerando um total de $n*(n-1)/2$ comparações, onde n é o numero total de instâncias (CHRISTEN; GOISER, 2007). Portanto, as técnicas de blocagem são utilizadas para reduzir o número de comparações entre as instâncias, particionando o conjunto de registros de entrada em blocos. Isso restringe que a etapa de correspondência de pares ocorra apenas nos pares candidatos à correspondência, evitando que pares pouco prováveis de corresponder sejam comparados.

A blocagem foi proposta como uma estratégia para diminuir o número de comparações entre pares, tornando o processo de Resolução de Entidades escalável e possível para grandes conjuntos de dados. A blocagem padrão consiste em separar o conjunto total de dados identificados por uma chave de blocagem onde, preferencialmente, todas as instâncias referentes a uma mesma entidade estejam no mesmo bloco.

Para exemplificar, considere um conjunto de dados contendo instâncias de pessoas. Como chave de blocagem foi atribuído o atributo sobrenome, então todas as instâncias com o mesmo sobrenome foram inseridas no mesmo bloco. Deste modo, apenas as instâncias com o mesmo valor de chave de blocagem seriam comparadas detalhadamente na etapa de correspondência de pares, diminuindo drasticamente o número de comparações.

Genericamente, os métodos de blocagem podem ser classificadas em dois grupos de acordo com a abordagem usada para organizar os registros em blocos: (i) métodos estáticos, (ii) métodos dinâmicos. Nos métodos estáticos, o processo de agrupamento das instâncias não leva em consideração as características encontradas nos dados. Essa técnica se baseia em regras de agrupamento predefinidas, e essas regras não mudam de acordo com as características encontradas nos dados. Um exemplo de técnica de blocagem estática é o *Canopy Blocking* (MCCALLUM; NIGAM; UNGAR, 2000). Os métodos mais recentes de blocagem são baseados em Aprendizagem de Máquina *(Machine Learning)*, para determinar a melhor função de blocagem para o agrupamento e comparação de instâncias e, por isso, são conhecidos como dinâmicos e adaptativos. Um exemplo de técnica de blocagem dinãmica é o *DNF Blocking* (BILENKO; KAMATH; MOONEY, 2006) (EVANGELISTA et al., 2010).

Foram propostos vários métodos para blocagem de dados na litertura. O *DNF Blocking* (BILENKO; KAMATH; MOONEY, 2006), baseado em aprendizagem de máquina, utiliza pares de instâncias classificadas como verdadeiras, caso representem a mesma entidade, ou falsas, caso sejam de entidades distintas. Essas instâncias são utilizadas para a escolha de regras que produzam os melhores agrupamentos na blocagem. O *Soft Term Frequency-Inverse Document Frequency (TF-IDF)*, foi proposto inicialmente como uma medida de similaridade de *Strings* e, posteriormente, adaptado para a blocagem. O método de blocagem *Canopy Blocking* (MCCALLUM; NIGAM; UNGAR, 2000), é composto de duas etapas. Na primeira etapa as instâncias são agrupadas com um baixo custo de processamento, e a segunda etapa tem o objetivo

de refinar os resultados da primeira etapa. O método BGP (EVANGELISTA et al., 2010) também utiliza Aprendizagem de Máquina para descobrir boas chaves de blocagem. Este método é baseado em programação genética, permitindo o uso de regras mais flexíveis e um maior número de regras para a definição de funções de blocagem.

2.1.2.2 Correspondência entre Pares

A correspondência entre pares é a principal fase do processo de Resolução de Entidades. Esta fase tem como objetivo realizar a comparação detalhada de um par de instâncias. Geralmente, a semelhança entre duas instâncias é calculada comparando-se vários atributos, e quanto maior a similaridade dos atributos das instâncias, mais provável elas são de corresponderem à mesma entidade do mundo real.

Para isso, várias técnicas de correspondência têm sido propostas na literatura. As técnicas baseadas em regras são as mais comumente utilizadas no processo de comparação de pares, e utilizam conhecimento de domínio para classificar se um par de instâncias é ou não duplicado. Uma vantagem dessa abordagem é que as regras podem ser adaptadas para lidar com diversos cenários, mas uma desvantagem é que requer um considerável conhecimento de domínio, bem como conhecimento sobre os dados para formular as regras de correspondência (DONG; SRIVASTAVA, 2015).

As técnicas baseadas em classificação, criam um classificador utilizando um conjunto de treinamento contendo exemplos positivos e negativos, e o classificador decide se um par de instâncias é ou não duplicado. Sua vantagem é que não exige conhecimento do domínio. A desvantagem é que essa técnica muitas vezes exige um grande número de exemplos de treinamento para treinar com precisão o classificador.

Técnicas baseadas em funções de similaridade também são muito utilizadas. Essas técnicas se baseiam em aplicar métricas para calcular a similaridade entre os valores dos atributos que descrevem as instâncias. Essas técnicas podem ser classificadas como baseadas em caracteres, em *tokens*, ou híbridas.

Entre as funções baseadas em caracteres a mais conhecida é a função de similaridade por distância de edição (*edit distance*), ou distância de Levenshtein (LEVENSHTEIN, 1966), em que dadas duas cadeias de caracteres é retornado um valor de acordo com as operações de edição, com respectivos pesos. As operações de edição podem ser de substituição, deleção e adição de caracteres. Também nesse grupo encontra-se a função de similaridade Jaro (JARO, 1989) que é baseada na ordem e número de caracteres comuns entre duas cadeias de caracteres. A Jaro conta o número de caracteres que são comuns em duas *strings* dentro de uma determinada janela de caracteres, e o número de transposições entre caracteres. Com base nisso a similaridade Jaro é calculada (CHRISTEN, 2012).

Das funções baseadas em *tokens* vale a pena destacar a *Cosine Similarity* e a Jaccard. A *Cosine Similarity* (RIBEIRO-NETO; BAEZA-YATES, 1999), mede a similaridade entre duas cadeias de caracteres utilizando um *Vector Space Model*, modelo algébrico utilizado pela primeira

vez nos anos 60. A função de similaridade Jaccard (JACCARD, 1901), utiliza o coeficiente de Jaccard e mede a similaridade entre dois conjuntos dividindo a interseção dos conjuntos pela sua união.

Entre as funções de similaridade híbridas, destaca-se a proposta por Monge e Elkan (1996), que mescla funções de caracteres e *tokens* para calcular a similaridade entre cadeias de caracteres.

2.1.2.3 Classificação

A classificação dos pares de instâncias se dá com base no valor de similaridade obtido por meio da fase de correspondência de pares. A partir da comparação, os pares são classificados como correspondentes ou não correspondente. Em algumas abordagens, os pares também podem ser classificados como potencialmente correspondentes. Se um par é dado como correspondente, significa que as instâncias são duplicadas, se é dado como não correspondente, as instâncias não são duplicadas, e se for classificado como potencialmente correspondente, significa que são instâncias possivelmente correspondentes, mas que é necessária uma revisão manual para decidir se elas correspondem ou não.

A classificação pode ser dividida em duas abordagens: não supervisionada ou supervisionada. Na abordagem não supervisionada, os pares são classificados com base apenas na similaridade calculada entre eles na fase de correspondência, sem ter acesso a quaisquer características de pares de instâncias verdadeiramente correspondentes e não correspondentes. Na abordagem supervisionada, é utilizado um conjunto de treinamento com pares verdadeiramente correspondentes e não correspondentes, formando um classificador supervisionado (CHRISTEN, 2012).

A forma mais simples de classificar os pares de instâncias como correspondentes ou não, é aplicar um limiar *(threshold)* de similaridade. A seleção do limiar pode ser feita manualmente, ou ser aprendido, utilizando um conjunto de treinamento. Outra abordagem comumente utilizada, é a classificação baseada em regras. São empregadas regras, onde de acordo com elas os pares são classificados como correspondentes ou não. A classificação baseada em regras é aplicada sobre os valores de similaridade entre os atributos das instâncias, calculados na etapa de correspondência. As regras são criadas com valores de similaridade de determinados atributos, combinados com conjunções (e), disjunções (ou) e negações (não).

As técnicas de classificação citadas acima são vistas como classificação tradicional. Uma abordagem alternativa, é utilizar a classificação baseada em grupos *(clustering)*, onde cada grupo *(cluster)* é composto de instâncias que se referem à mesma entidade (CHRISTEN, 2012). O agrupamento é o processo de agrupar dados que são semelhantes uns aos outros de acordo com algum critério (HAN; KAMBER; PEI, 2011). Geralmente é realizado de forma não supervisionada, portanto, não necessita de um conjunto de treinamento.

Deste modo, a etapa de agrupamento tem como objetivo particionar o conjunto de todas as instâncias, de acordo com as similaridades calculadas, de modo que cada partição se refira

a uma entidade distinta e, no caso ideal, todas as instâncias referentes a essa entidade estejam contidas nessa partição.

Muitos algoritmos de agrupamento foram desenvolvidos nas áreas de Estatística, Mineração de Dados e Aprendizagem de Máquina (HERNÁNDEZ; STOLFO, 1998; ASLAM; PELEKHOV; RUS, 2004; BANSAL; BLUM; CHAWLA, 2004). Em Hassanzadeh et al. (2009) foi realizado um estudo comparativo da acurácia de alguns desses algoritmos para o processo de Resolução de Entidades, avaliando a qualidade dos *clusters* gerados pelo processo por meio do framework *Stringer*, proposto por eles.

2.1.2.4 Avaliação da Qualidade do Processo de Resolução de Entidades

A qualidade do processo de Resolução de Entidades pode ser medida utilizando as seguintes dimensões.

- **Verdadeiro positivo**: pares de instâncias classificadas como correspondentes e que são verdadeiramente correspondentes.
- **Falso positivo**: pares de instâncias classificadas como correspondentes mas que não são verdadeiramente correspondentes.
- **Verdadeiro negativo**: pares de instâncias classificadas como não correspondentes e são verdadeiramente não correspondentes.
- **Falso negativo**: pares de instâncias classificadas como não correspondentes mas que são verdadeiramente correspondentes.

Com base no número de verdadeiros positivos **(TP)**, verdadeiros negativos **(TN)**, falsos positivos **(FP)** e falsos negativos **(FN)**, diferentes medidas de qualidade podem ser calculadas (CHRISTEN; GOISER, 2007). As medidas mais utilizadas para avaliar a qualidade do processo de Resolução de Entidades são apresentadas a seguir.

Precisão (*Precision*): É a proporção de correspondências que são classificadas como verdadeiras. Na precisão não se inclui o número de verdadeiros negativos, evitando o problema de desequilíbrio que a medida de acurácia possui. É calculada de acordo com a Equação 2.1.

$$prec = \frac{TP}{TP+FP} \tag{2.1}$$

Revocação (*Recall*): É semelhante a precisão, já que não utiliza o número de verdadeiros negativos. Mede quantos dos pares verdadeiramente correspondentes foram classificados como correspondentes. O cálculo é feito conforme a Equação 2.2.

$$rec = \frac{TP}{TP+FN} \tag{2.2}$$

Medida F (*F-measure*): Também chamada de *F-score*, a medida calcula a média harmônica entre os valores de precisão e cobertura. O valor de *F-measure* é obtido por meio da Equação 2.3.

$$f - measure = 2\left(\frac{prec * rec}{prec + rec}\right) \tag{2.3}$$

Especificidade (*Specifity*): Esta medida é também conhecida como taxa de Verdadeiros negativos, e se calcula conforme a Equação 2.4.

$$spec = \frac{TN}{TN + FP} \tag{2.4}$$

Segundo Christen (2012), as medidas de precisão, revocação e *F-measure* têm sido muito utilizadas, e ganharam popularidade nos últimos. A medida de *F-measure* foi reconhecida como uma métrica padrão, e recomendada como a melhor métrica para avaliar a qualidade do processo de Resolução de Entidades.

2.1.3 Fusão de Dados

A Fusão de Dados é a terceira das etapas principais do processo de Integração de Dados. Seu objetivo é combinar as instâncias que se referem à mesma entidade do mundo real (identificados pelo processo de Resolução de Entidades) fundindo-as em uma única representação (DONG; NAUMANN, 2009). Visando esse objetivo, a Fusão de Dados também é responsável por resolver os conflitos existentes entre as instâncias e aumentar a corretude nos dados integrados.

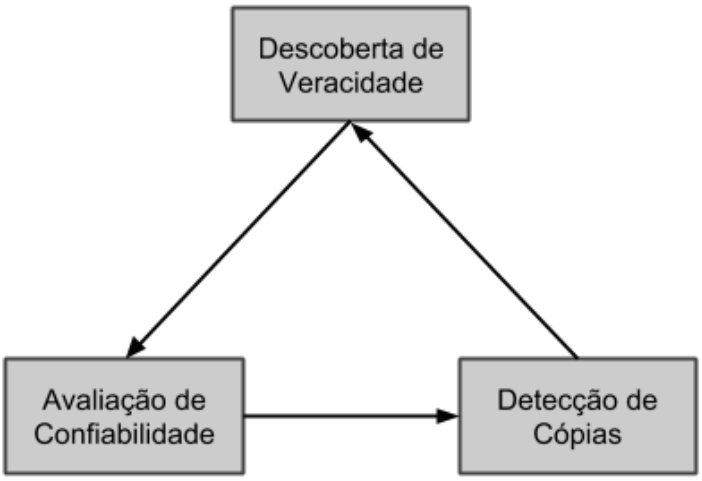
Os possíveis conflitos são classificados, segundo Dong e Naumann (2009), em dois tipos: incerteza e contradição. A incerteza é quando há um conflito entre um valor não-nulo e um ou mais valores nulos, todos utilizados para descrever a mesma propriedade de uma entidade real. Esse conflito é causado por informações faltantes em valores de atributos. O conflito de contradição ocorre quando dois ou mais valores não-nulos que descrevem a mesma propriedade de uma entidade são diferentes.

Para resolver os conflitos encontrados são propostas estratégias divididas em três tipos: estratégias para ignorar conflitos, estratégias para evitar conflitos e, por fim, estratégias para resolver conflitos, sendo a última a mais utilizada, fornecendo meios para resolver os conflitos individualmente.

Além das estratégias citadas, ainda podem ser utilizadas informações de qualidade das fontes para ajudar no processo de fusão, como a acurácia e a atualidade das fontes (DONG; SAHA; SRIVASTAVA, 2012).

Na Figura 2.4, as etapas da Fusão de Dados são mostradas: descoberta de veracidade, onde entre os valores conflitantes é descoberto qual é verdadeiro, e como base é tomado que o valor fornecido pelo maior número de fontes é verdadeiro; avaliação de confiabilidade, onde para cada fonte de dados é avaliada a confiabilidade; e detecção de cópias, onde é verificado se existe

Figura 2.4: Etapas da Fusão de Dados.



Adaptado de: Dong e Srivastava (2015)

cópia entre as fontes de dados.

2.2 Seleção de Atributos

A seleção de atributos tem como objetivo descobrir um subconjunto de atributos relevantes para uma tarefa alvo (DASH et al., 2002). O processo de seleção de atributos permite a ordenação dos atributos de acordo com algum critério de importância, ou seja, a redução de dimensionalidade do espaço de busca de atributos e a remoção de dados contendo ruídos, entre outros (LEE, 2005).

De maneira geral, busca-se com a seleção de atributos encontrar o melhor subconjunto de atributos de acordo com algum critério de qualidade. Idealmente, o melhor subconjunto contém o menor número de atributos que mais contribuem para o processo pretendido.

Diversas pesquisas foram desenvolvidas na tentativa de propor soluções para a seleção de atributos, principalmente nas áreas de Mineração de Dados e Aprendizagem de Máquina (DASH; LIU, 1997; DASH; LIU, 2000; DASH et al., 2002; DY; BRODLEY, 2004; ZHAO; LIU, 2007), que foram as primeiras a realizar pesquisas nessa área, pretendendo, além de melhorar a eficiência do processo de aprendizagem, diminuir a dimensionalidade das amostras, minimizando o custo computacional e também o espaço de armazenamento dos dados.

2.2.1 Seleção de Atributos na Mineração de Dados

Atualmente se associa a Mineração de Dados à busca do conhecimento compreensível, útil e surpreendente em bases de dados, e a aplicação dispensa a presença de um número significativo de atributos ou instâncias presentes das bases de dados originais, e que em certos

casos, não agregam em nada e até podem “atrapalhar” o processo de aprendizagem (BORGES, 2011).

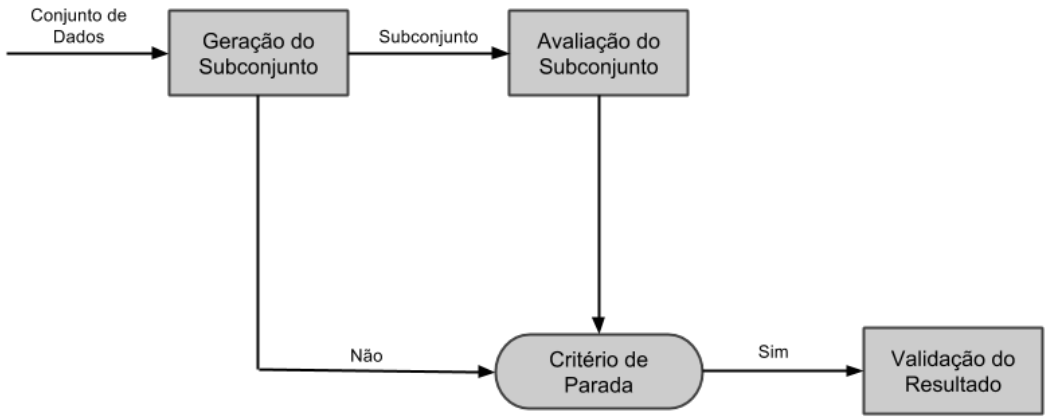
É comum pensar que quanto maior a quantidade de atributos, mais informações estariam disponíveis para o algoritmo de mineração de dados, porém, conjuntos de dados que contêm muitos atributos tendem a ficar mais esparsos, e isso traz uma dificuldade, conhecida como "maldição da dimensionalidade” (DY; BRODLEY, 2004).

A seleção de atributos na Mineração de Dados é um processo que tem como objetivo selecionar um subconjunto de M atributos do conjunto original de N atributos $(M \leq N)$, de modo que o espaço de atributos seja reduzido de acordo com um determinado critério (LIU; YU, 2005). A seleção de atributos é utilizada para melhorar a qualidade dos dados e modelos que são construídos, e torná-los mais compreensíveis. O desempenho de um algoritmo de aprendizagem é prejudicado tanto na velocidade (devido à dimensionalidade dos dados) quanto no resultado (devido às informações redundantes que podem confundir o algoritmo, não auxiliando na busca de um modelo correto pra o conhecimento) (MACEDO, 2012).

Se a tarefa alvo é a classificação, a seleção de atributos é útil para minimizar a taxa de erro dos classificadores, a complexidade do conhecimento a ser gerado por ele, e o número de atributos selecionados para compor a “nova” base (BORGES, 2011).

Segundo Dash e Liu (1997), o processo de seleção de atributos é composto por quatro etapas, como mostra a Figura 2.5.

Figura 2.5: Etapas do Processo de Seleção de Atributos.



Adaptado de: Dash e Liu (1997)

A partir de todos os atributos disponíveis, seleciona-se um subconjunto de variáveis relevantes com o apoio de um algoritmo de busca. Portanto, a geração dos subconjuntos é responsável por produzir subconjuntos de atributos baseada em uma estratégia de busca. Diferentes estratégias de busca são utilizadas, como busca exponencial e busca sequencial.

A busca exponencial faz todas as combinações possíveis entre os atributos para encontrar o subconjunto ótimo de atributos. Na busca sequencial os algoritmos podem ser sequenciais para frente ou para trás. Na sequencial para frente a busca inicia por um subconjunto vazio

de atributos, e inicialmente são avaliados subconjuntos com um atributo. Quando o melhor atributo é encontrado, ele é combinado com todos os outros disponíveis (em pares) e o melhor subconjunto é selecionado. A busca continua adicionando um atributo por vez até que não se consiga mais melhorar a qualidade do subconjunto. Na busca sequencial para trás, ao contrário, a busca inicia por um subconjunto de atributos e a cada iteração um atributo é removido até que não se consiga melhorar a qualidade do subconjunto encontrado.

Na etapa de avaliação do subconjunto, cada subconjunto é avaliado de acordo com um critério de avaliação. Os critérios de avaliação podem ser categorizados em duas abordagens: abordagens filtro e *wrapper*.

Na abordagem de filtro, o subconjunto é escolhido independentemente do algoritmo de mineração. A avaliação do subconjunto de atributos é realizada com base na avaliação individual de cada atributo. Para isso existem várias medidas de avaliação, dentre elas destacamos as medidas de informação, de dependência e de consistência, e são independentes do algoritmo utilizado (JOUVE; NICOLOYANNIS, 2005).

Na abordagem *wrapper*, a avaliação de qualidade do subconjunto de atributos é estimada analisando a qualidade do resultado do algoritmo utilizado, ou seja, a abordagem *wrapper* necessita pré-determinar um algoritmo de mineração e utiliza seu desempenho como critério de avaliação.

O critério de parada estabelece quando a seleção de atributos deve parar. De acordo com Liu e Yu (2005), quatro critérios podem ser utilizados:

- Todo o espaço de busca foi avaliado;
- O número alvo de atributos foi alcançado ou o número máximo de iterações;
- Adição (ou remoção) de atributos não produz um conjunto melhor;
- O conjunto suficientemente bom é encontrado, por exemplo, se o conjunto encontrado proporciona uma baixa taxa de erros.

A validação dos resultados é a ultima etapa da seleção de atributos e uma das formas de realizá-la é medindo diretamente o resultado, tendo um conhecimento prévio dos dados. No entanto, em alguns casos não se tem conhecimento prévio dos dados. Nesses casos, alguns métodos de monitoramento de mudança de desempenho podem ser utilizados (LIU; YU, 2005). Por exemplo, se utilizar uma taxa de erro de classificação como um indicador de desemprenho de uma tarefa de mineração, pode-se comparar o resultado anterior, utilizando todos os atributos, com o resultado obtido utilizando apenas os atributos selecionados.

Várias técnicas para seleção de atributos foram propostas na literatura (DASH; LIU, 2000; DASH et al., 2002; DY; BRODLEY, 2004; ZHAO; LIU, 2007). Dash e Liu (2000) propuseram uma técnica baseada em *wrapper*, que utiliza entropia[1] para identificar atributos sem

[1] Dash e Liu (2000) definem entropia a partir da similaridade entre dois atributos. A entropia apresenta valores baixos quando dois atributos têm alguma relacão, e valores altos quando esses atributos não estão relacionados.

importância e com ruído em um conjunto de dados. A relevância de um atributo é calculada a partir da entropia do conjunto de dados, eliminando o atributo em questão. Em Dash et al. Em Dash et al. (2002), o trabalho de Dash e Liu (2000) foi evoluído, e uma seleção de atributos baseada em filtros foi proposta para selecionar os atributos mais relevantes.

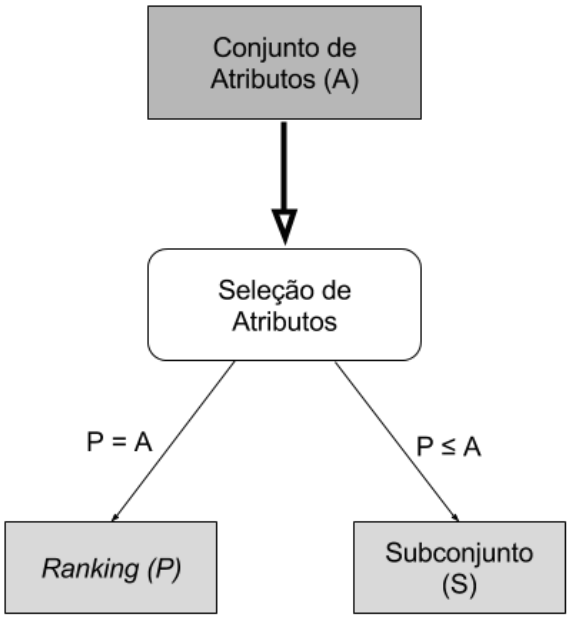
Outra técnica de seleção de atributos é o FSSEM *Feature Subset Selection using EM Clustering* (DY; BRODLEY, 2004), baseada em *wrapper*. Foi utilizada uma busca incremental, inicializando com zero atributos e, sequencialmente, adicionando um atributo de cada vez. Utiliza o algoritmo de agrupamento de dados *EM clustering*, e dois critérios se seleção: *Scatter Separability* e *Maximum Likelihood*.

Em Zhao e Liu (2007) é proposta a técnica SPEC (*Spectral Feature Selection*), que pode ser utilizada tanto para problemas supervisionados quanto não supervisionados.

2.2.2 Seleção de Atributos na Resolução de Entidades

Para a etapa de comparação do processo de Resolução de Entidades devem ser selecionados os atributos comuns a todas as instâncias envolvidas. As questões envolvidas no processo de seleção incluem: identificar os atributos comuns; verificar se o atributo tem informação suficiente para dar suporte a uma correspondência de qualidade; selecionar um subconjunto ótimo de atributos comuns (GU et al., 2003).

Figura 2.6: Tipos de Saída do Processo de Seleção de Atributos.



Não importa qual abordagem será utilizada no processo de Resolução de Entidades, em todas é necessário decidir quais atributos serão comparados. Os atributos selecionados para a etapa de comparação influenciam diretamente o processo de Resolução de Entidades, portanto, a escolha dos atributos é de extrema importância para se obter um resultado de qualidade.

Do ponto de vista da saída resultante da seleção de atributos, como se pode verificar na Figura 2.6, o processo pode fornecer o conjunto de atributos original ordenado *(ranking)*, ou um subconjunto de atributos relevantes, que está contido no conjunto de atributos original.

A seleção de atributos é uma etapa importante para o processo de Resolução de Entidades que afeta diretamente a qualidade do resultado do processo. Atributos contendo valores errôneos ou repetitivos, por exemplo, afetam o resultado do processo. Por exemplo, um campo como o sexo de uma pessoa que contém apenas dois possíveis valores (masculino e feminino) não fornece informações suficientes para identificar um par correspondente. Por outro lado, um atributo como sobrenome, fornece muito mais informação sobre a instância que representa, mas pode frequentemente ser registrado incorretamente, e conter valores errôneos (GU et al., 2003).

Os estudos sobre Resolução de Entidades muitas vezes consideram todos os atributos comuns às duas instâncias envolvidas no processo. Outras vezes, utilizam a seleção de atributos de forma manual, como nos frameworks para Resolução de Entidades analisados e comparados pelo estudo de Kopcke e Rahm (2010), deixando a cargo do usuário analisar os dados e selecionar os atributos que considera relevantes para a etapa de comparação. Existem ainda trabalhos em que são selecionados alguns atributos para realizar o processo de Resolução de Entidades, mas não é explicado de que modo ou quais critérios foram utilizados para serem selecionados tais atributos.

No entanto, algumas estratégias de seleção de atributos para a Resolução de Entidades foram propostas (SU et al., 2010; FAN et al., 2009; CHEN et al., 2012). Por serem trabalhos relacionados a estratégia proposta neste trabalho, serão detalhados no próximo capítulo.

2.3 Considerações

Neste capítulo, foram apresentados os principais assuntos relacionados à estratégia proposta nessa dissertação. Inicialmente, abordamos o problema de Integração de Dados, descrevendo os principais conceitos, e detalhando as etapas do processo. Em seguida, focamos na etapa de Resolução de Entidades, onde a estratégia proposta neste trabalho está inserida. Detalhamos cada uma de suas fases e apresentamos as métricas utilizadas para avaliar a qualidade do processo de Resolução de Entidades.

Também neste capítulo, foram mostrados os principais conceitos e objetivos acerca do processo de seleção de atributos, tanto para uso na Mineração de Dados e Aprendizagem de Máquina quanto na Resolução de Entidades. Pode-se verificar que muitas técnicas para seleção de atributos foram propostas nas áreas de Mineração de Dados e Aprendizagem de Máquina. No entanto, na Resolução de Entidades, o problema de seleção de atributos não é muito abordado, tendo poucos trabalhos relacionados ao assunto.

No próximo capítulo, os trabalhos relacionados à estratégia proposta neste trabalho serão descritos.

3

Trabalhos Relacionados

Neste capítulo são apresentadas estratégias de seleção de atributos, com uma introdução às estratégias proposta na Mineração de Dados e Aprendizagem de Máquina (Seção 3.1.1), e foco nas estratégias de seleção de atributos para o processo de Resolução de Entidades (Seção 3.1.2). São apresentadas algumas discussões, comparando as estratégias mais similares à proposta nesta dissertação (Seção 3.1.3).

3.1 Estratégias de Seleção de Atributos

A Seleção de Atributos tem sido alvo de pesquisas em áreas da Ciência da Computação, tais como: Mineração de Dados e Aprendizagem de Máquinas (DY; BRODLEY, 2004; JOUVE; NICOLOYANNIS, 2005; LI; LU; WU, 2006; COVOES; HRUSCHKA, 2011) e Integração de Dados (CHEN et al., 2012; SU et al., 2010; FAN et al., 2009). Nas próximas seções serão detalhados os trabalhos mais relevantes em cada uma dessas áreas.

3.1.1 Estratégias de Seleção de Atributos na Mineração de Dados e Aprendizagem de Máquina

A Seleção de Atributos nas áreas de Mineração de Dados e Aprendizagem de Máquina, além de melhorar a eficiência do processo, tem como objetivo diminuir a dimensionalidade das amostras, minimizando o custo computacional e também o espaço de armazenamento dos dados. Várias técnicas foram propostas, as quais podem ser divididas em três abordagens: *wrappers*, filtros e híbridas. As abordagens baseadas em *wrapper* utilizam o resultado do algoritmo que realiza a tarefa desejada (por exemplo, uma tarefa de classificação ou agrupamento de um conjunto de dados) para selecionar um subconjunto de atributos ótimos (DY; BRODLEY, 2004). As abordagens baseadas em filtro fazem a seleção de atributos com base na avaliação individual de cada atributo, utilizando métricas como correlação, consistência ou ganho de informação, e são independentes do algoritmo utilizado (JOUVE; NICOLOYANNIS, 2005). As abordagens híbridas utilizam tanto informações dos dados, quanto o resultado do algoritmo para fazer a

seleção (LI; LU; WU, 2006).

O trabalho de Dy e Brodley (2004) propõe um método de seleção de atributos (FSSEM) baseado na abordagem de *wrapper*, que tem como objetivo identificar subconjuntos de atributos que melhor descobrem agrupamentos naturais nos dados. A busca no espaço de atributos para formar os subconjuntos é feita de forma incremental, começando com zero atributos, e adicionando um atributo de cada vez. A busca é interrompida quando, ao adicionar novos atributos, o critério de seleção não melhora.

Em Jouve e Nicoloyannis (2005) é proposta uma técnica de seleção de atributos para clusterização, baseada na abordagem de filtro e, portanto, independente do algoritmo de agrupamento. O método proposto baseia-se em dois índices para avaliar a adequação entre dois conjuntos de atributos (isto é para determinar se dois conjuntos de atributos contêm as mesmas informações).

O trabalho de Li, Lu e Wu (2006), se baseia na abordagem híbrida, e propõem um método de seleção de atributos utilizando aprendizagem não supervisionada. O método se baseia na classificação dos atributos de acordo com sua relevância para os *clusters*.

Em Covoes e Hruschka (2011), é proposto um algoritmo para seleção de atributos baseado na abordagem híbrida. Esse algoritmo particiona o conjunto original de atributos em grupos de atributos correlacionados e, posteriormente, são selecionados atributos de cada um desses grupos. Os atributos dentro do mesmo grupo são mais semelhantes entre si do que os atributos pertencentes a grupos distintos.

3.1.2 Estratégias de Seleção de Atributos na Integração de Dados

Na Integração de Dados, mais especificamente no processo de Resolução de Entidades, o objetivo da seleção de atributos é encontrar um subconjunto de atributos a partir do conjunto original, removendo aqueles atributos que não contribuem positivamente para a fase de comparação entre instâncias, visando encontrar o maior número possível de correspondências verdadeiras, com o menor número de correspondências erradas. Considerando este contexto, poucos são os trabalhos que tratam do problema de Seleção de Atributos. Dentre eles, destacam-se (CHEN et al., 2012), (SU et al., 2010), e (FAN et al., 2009), que serão mais detalhados a seguir.

3.1.2.1 Um método de Resolução de Entidades Baseado em Aprendizagem de Máquina

Chen et al. (2012) propõem um método para Resolução de Entidades baseado em Aprendizagem de Máquina. Além de encontrar o atributo, ou grupo de atributos mais relevantes para o processo de Resolução de Entidades, para cada atributo ou grupo de atributos é encontrada a função de similaridade e o limiar adequados.

Para ilustrar a motivação do trabalho, os autores utilizaram uma fração de dados da base de dados Cora [1], onde quatro citações são apresentadas e se referem à duas publicações distintas,

[1] http://www.cs.umass.edu/ mccallum/data/cora-refs.tar.gz

como podemos ver na Tabela 3.1.

Tabela 3.1: Um exemplo utilizando a base de dados Cora.

Id Registro	Id Entidade	Autor	Título	Endereço	Data
1	1	carla e. brodley and paul e, utgoff.	multivariate versus univariate decision trees.	amherst ma	1992.
2		c. e. brodley and p. e. utgoff.	multivariate decision trees.	amherst massachusetts	1992.
3	2	c. e. brodley and p. e. utgoff.	multivariate decision trees.	NULL	1995.
4		carla e. brodley and paul e. utgoff	multivariate decision trees.	NULL	1995.

Adaptado de: Chen et al. (2012)

Os autores justificam que não é adequado utilizar qualquer atributo na comparação, já que os valores de um atributo podem conter erros ou estar ausentes. Considerando as tuplas 1 e 2, ambas se referem à mesma entidade do mundo real, no entanto o atributo título possui valores diferentes. No caso das tuplas 2 e 3, que se referem à entidades distintas, os valores dos atributos autor e título, são iguais. Se esses atributos fossem utilizados na comparação entre instâncias, provavelmente ocasionariam correspondências erradas. Com esta motivação, o trabalho tem como objetivo encontrar o grupo de atributos, a função de similaridade e o limiar adequados para a fase de comparação de instâncias.

As funções de similaridade avaliadas foram cinco: *Cosine Distance*, *Jaccard Distance*, *Jaro Distance*, *Edit Distance* e *Q-gram Distance*. Para isso, eles utilizam como base o *F-measure*, ou seja, em cima do conjunto de treinamento são feitas combinações entre atributos, funções de similaridade e limiares, e a combinação que resultar no maior *F-measure* é dada como adequada para o processo de Resolução de Entidades. Depois, são avaliados grupos de atributos, onde os grupos com os maiores valores de *F-measure* são escolhidos como regra para a fase de comparação.

A solução proposta é dividida em duas fases: fase de treinamento, onde o objetivo é encontrar um grupo de atributos para ser utilizado no processo de Resolução de Entidades, e fase de teste, onde para os pares de instâncias, a similaridade é calculada e o par é classificado.

Encontrar o grupo de atributos apropriado para a Resolução de Entidades é um processo custoso se o número de atributos de uma entidade for grande. Para reduzir o custo de processamento foi desenvolvido um método heurístico, evitando testar todos os grupos de atributos possíveis. Dessa forma, só é necessário testar um pequeno número de grupos de atributos. Para isso, é proposto um algoritmo denominado *getCandidateGroups*, responsável por gerar os grupos de atributos candidatos.

Para encontrar o grupo de atributos mais relevante, é proposto um algoritmo denominado *getAttributeGroup*. Esse algoritmo recebe como parâmetros c, e E, onde c é o número máximo de atributos que cada grupo deve conter, e E é a entidade. Primeiro, todos os atributos de E são testados em triplas contendo o *F-measure* que o atributo alcança, o *threshold* adequado, e a função de similaridade adequada. Depois, o grupo de atributos candidato é gerado incrementalmente, por meio da chamada ao algoritmo *getCandidateGroups*. Se não houver mais grupos a serem gerados, a criação dos grupos é finalizada. Finalmente, o algoritmo retorna o grupo que alcançou o máximo *F-measure*, e esse grupo é dado como apropriado para ser utilizado no processo de

Resolução de Entidades.

Tabela 3.2: Resultados do experimento que avalia os atributos individualmente - base de dados Cora.

Atributo	MAXF	Limiar Adequado	Função de similaridade Adequada
Autor	0.60	0.32	q-gram distance
Título	**0.94**	**0.28**	**edit distance**
Endereço	0.13	0.45	q-gram distance
Data	0.33	0.50	Jaro distance
Páginas	0.65	0.15	q-gram distance
Volume	0.53	0.67	q-gram distance
Editora	0.11	0.56	q-gram distance
Editor	0.20	0.60	q-gram distance
Periódico	0.37	0.23	q-gram distance

Adaptado de: Chen et al. (2012)

Para avaliar a estratégia proposta, foram feitos experimentos com as bases de dados Cora e Restaurantes. Primeiro foi encontrado o valor de *F-measure* de cada atributo nas duas bases e a função de similaridade para cada um. Para exemplificar, a Tabela 3.1 mostra os resultados para a base de dados Cora, onde foram considerados apenas 9 atributos. Inicialmente, os atributos são avaliados individualmente. Como resultado inicial, o atributo título utilizando um limiar de 0,28 e a função de similaridade *edit distance* foram dados como parâmetros adequados para a comparação, atingindo um *F-measure* de 0,94.

Depois, foi realizado um experimento para avaliar o desempenho dos grupos de atributos obtidos a partir do conjunto de treinamento, para as duas bases. Como resultado para a base de dados Cora, podemos ver na Tabela 3.3 que os três primeiros grupos elencados proporcionaram um *F-measure* de 0,94 ao processo de Resolução de Entidades.

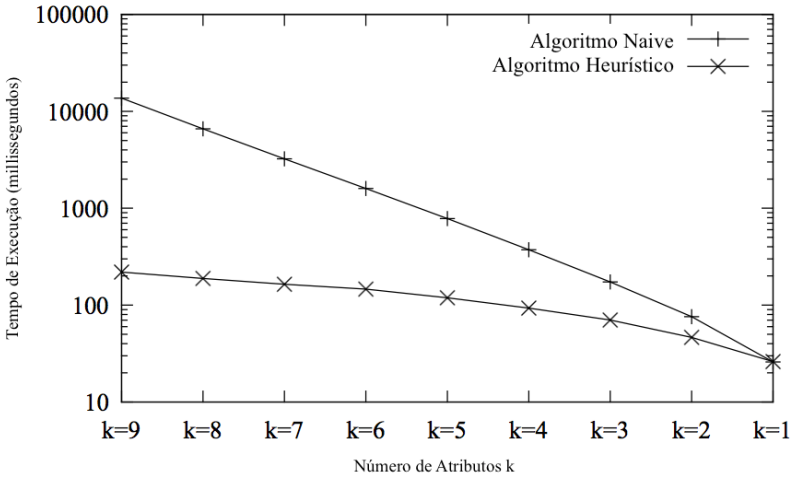
Tabela 3.3: Resultado do experimento que avalia grupos de atributos - base de dados Cora.

Classificação	Grupos de Atributos	MAXF Treinamento	MAXF Teste
1	Autor, Título, Endereço, Periódico	0.94	0.94
2	Título, Volume	0.94	0.95
3	Título	0.94	0.93
4	Autor, Endereço, Páginas, Volume, Editor	0.74	0.71
5	Autor, Endereço, Data, Páginas, Editora, Editor, Periódico	0.74	0.71

Adaptado de: Chen et al. (2012)

No último experimento, foi avaliado o desempenho do algoritmo heurístico responsável por formar os grupos de atributos candidatos. O algoritmo heurístico proposto no artigo foi comparado com o algoritmo *Naive*. Os experimentos mostram que o algorimo proposto supera o algoritmo Naive, como mostra o gráfico da Figura 3.1. A linha horizontal do gráfico representa a quantidade de atributos contidos nos grupos formados, e a linha vertical representa o tempo de execução.

Figura 3.1: Resultado do desempenho do algoritmo heurístico.



Adaptado de: Chen et al. (2012)

O método de Resolução de Entidades proposto foi comparado com outras duas técnicas existentes, *Op-Trees* e *SiFi-Hill*, e superou ambas, tanto em tempo de execução quanto no *F-measure* obtido. Como conclusão, os autores explicam que com os experimentos realizados é possível verificar que o método proposto é eficiente, eficaz e estável.

3.1.2.2 Resolução de Entidades Sobre Resultados de Consultas em Múltiplas Fontes de Dados Web

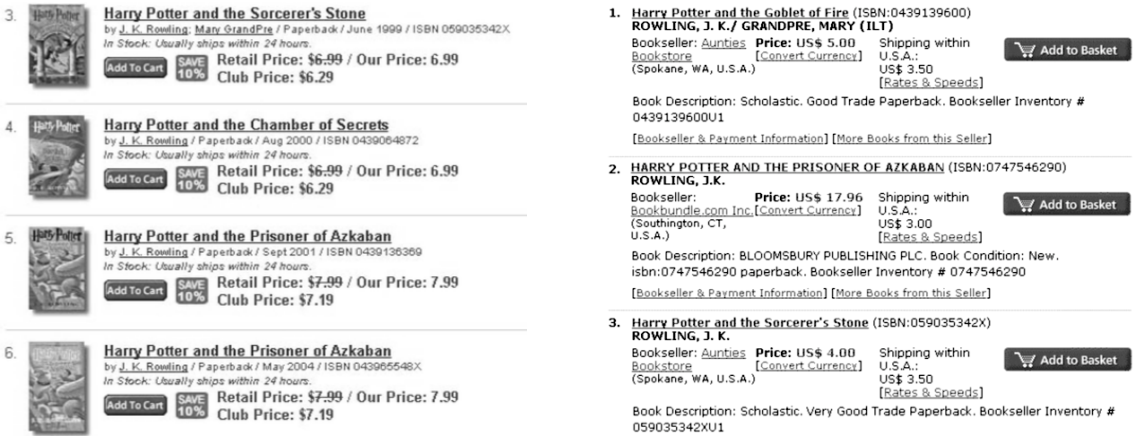
Em Su et al. (2010) é proposto um novo método de detecção de dados duplicados, especificamente para o cenário de identificação de duplicados entre instâncias contidas em resultados de consultas a múltiplas fontes de dados *Web* do mesmo domínio. O foco do trabalho é na técnica de ajustar os pesos dos atributos para o cálculo de similaridade entre duas instâncias. Os autores destacam que precisam ser atribuídos diferentes pesos aos atributos de acordo com sua importância, de maneira dinâmica e adaptativa.

Os autores explicam que no cenário da *Web*, foco deste trabalho, as instâncias para correspondência são altamente dependentes de consulta, uma vez que são obtidas por meio de consultas *online*. Além disso, elas são apenas uma pequena parte de todos os dados dos bancos de dados da *Web*. Deste modo, métodos de aprendizado *offline* não são apropriados. Primeiro porque o conjunto de dados completo não está disponível e, portanto, conseguir um bom conjunto de treinamento é uma tarefa difícil, e segundo porque mesmo se existir um bom conjunto de treinamento para o aprendizado, as regras aprendidas sobre esse conjunto de dados "completo"podem não ser boas, para o conjunto de dados resultante de uma consulta, já que o resultado de uma consulta é uma parte parcial e tendenciosa desse conjunto de dados.

Para ilustrar este cenário, os autores apresentaram o seguinte exemplo. Considere uma consulta por livros do autor *"J. K. Rowling"*, e tome como resultados a Figura 3.2.

Dependendo de como os bancos de dados na *Web* processam tal consulta, todos os

Figura 3.2: Um exemplo no cenário da Web.



(a) (b)

Fonte: Su et al. (2010)

resultados obtidos podem ter apenas *"J. K. Rowling"* como valor para o atributo autor. Neste caso, o atributo autor é incapaz de distinguir essas instâncias, que podem ser correspondentes ou não. Para reduzir a influência desse atributo na comparação, seu peso deve ser ajustado para ser menor do que os pesos dos outros atributos, ou até ser igual a 0. Além disso, para cada nova consulta, dependendo do resultado, os pesos dos atributos provavelmente devem mudar, o que torna os métodos supervisionados pouco aplicáveis.

Para superar esses problemas, os autores propõem um novo método para Resolução de Entidades, denominado *Unsupervised Duplicate Detection (UDD)*. A principal ideia do método é uma técnica para ajustar os pesos dos atributos no cálculo de similaridade entre duas instâncias. Como foi ilustrado no exemplo, a motivação principal é que diferentes atributos podem ter diferentes "importâncias"na comparação e, portanto, pesos diferentes devem ser atribuídos à eles.

O framework UDD, segundo os autores, é o primeiro que resolve o problema de detecção de duplicados *online* e o primeiro que consegue obter vantagens das dissimilaridades existentes sobre as instâncias de uma única base de dados *Web* para identificar correspondências, pois a maioria dos trabalhos existentes, utilizava conjuntos de treinamento positivo (conjunto de de dados contendo pares de instâncias correspondentes).

Neste trabalho, uma “amostra de dados universal” que consiste de pares de instâncias de diferentes fontes de dados é utilizada. Essa amostra é chamada de conjunto de treinamento negativo, já que a maioria dos pares de instâncias contidos no conjunto não são correspondentes. Os autores defendem que utilizar essa amostra facilita o processo de Resolução de Entidades já que, no cenário de consultas a bases de dados na *Web*, a porcentagem de instâncias duplicadas é bem menor do que a porcentagem de instâncias não duplicadas.

O algoritmo de atribuição de pesos *Weighted Component Similarity Summing (WCSS)*, atribui um peso a um atributo para indicar sua importância na correspondência, e a soma de todos

os pesos dos atributos deve ser igual a 1. Para o vetor de dados duplicados são atribuídos pesos altos aos atributos com alta similaridade entre seus valores, e pesos baixos para atributos com baixa similaridade em seus valores, e o contrário é feito para o vetor de dados não duplicados. Dessa forma, o algoritmo é capaz de aprender a ajustar os pesos dos atributos das instâncias de acordo com os vetores de dados.

Os experimentos foram realizados em cinco bases de dados, dentre elas, algumas bases de dados da *Web* em três domínios: *Livros*, dividido em duas bases, sendo *Livro-Completo* a base contendo os atributos *título, autor, editora,* e *ISBN*, e *Livro-titau*, contendo apenas os atributos *título* e *autor*, *Hotel* e *Filme*. Para avaliar o método proposto, foram utilizadas as métricas de precisão, revocação e *F-measure*.

Tabela 3.4: Resultado dos experimentos nas bases de dados da Web.

	Precisão	**Revocação**	**F-measure**	**Média Tempo de Execução (seg)**
Livro-Completo	0.954	0.925	0.939	0.85
Livro-Titau	0.947	0.952	0.950	0.36
Hotel	0.961	0.952	0.955	0.74
Filme	0.932	0.928	0.930	0.21

Fonte: Su et al. (2010)

A Tabela 3.4 mostra o resultado dos experimentos nas bases de dados da *Web*. Pode ser visto que o UDD é eficiente para identificar duplicados entre instâncias de múltipas fontes de dados, com alta precisão e revocação e uma boa média de tempo de execução.

Por fim, o UDD foi comparado com outros quatro métodos existentes, sendo eles: *SVM*, *OSVM*, *PEBL* e *Christen*. A Tabela 3.5 mostra a precisão, revocação, *F-measure* e tempo médio de execução do UDD em comparação com os outros métodos. Os resultados dos experimentos mostraram que a abordagem proposta funciona bem para o cenário de bases de dados *Web*, onde métodos supervisionados não são aplicáveis.

Tabela 3.5: Resultado da Comparação do UDD com outros métodos.

	Precisão	**Revocação**	**F-measure**	**Média Tempo de Execução (seg)**
UDD	0.924	0.915	0.919	0.85
SVM	0.926	0.933	0.929	0.36
OSVM	0.580	0.460	0.513	0.42
PEBL	0.902	0.803	0.851	1.42
Christen	0.886	0.867	0.876	1.64

Fonte: Su et al. (2010)

3.1.2.3 Inferência de Regras para Correspondência entre Instâncias

O trabalho de Fan et al. (2009) está inserido no contexto de Integração de Dados, na etapa de Resolução de Entidades, mais especificamente no problema de comparação de instâncias. Nas

abordagens tradicionais, o processo de comparação de instâncias é realizado apenas de forma semântica. Os autores justificam que nem sempre esse tipo de abordagem é suficiente, uma vez que os dados utilizados podem não ser de fontes confiáveis e possuírem erros. Nesse panorama, os autores propõem o uso de *Matching Dependencies (MD)* (dependências de matching).

A ideia de MD foi inspirada no conceito de *Dependence Functional* (dependência funcional), em outras palavras, a MD avalia o quão dependente um atributo é do outro. Os autores justificam que, por meio de um conjunto de MD, é possível extrair *Relative Candidate Keys (RCK)* (chave candidata relativa) que irão auxiliar no processo para determinar quais atributos e como comparar na etapa de comparação de instâncias. Os conceitos de MD e RCK foram propostos pelos autores, e a seguir iremos detalha-los.

A necessidade de utilizar dependências na Resolução de Entidades não é tema apenas deste trabalho, mas os autores afirmam que nenhum trabalho ainda tinha investigado como utilizar dependências para Resolução de Entidades em fontes de dados não confiáveis.

Por meio da especificação de uma MD é possível identificar correlações entre os atributos. Dessa forma, uma MD pode ser criada por meio da seguinte estrutura: Duas relações de esquema $R1$ e $R2$, onde cada relação possui uma lista de atributos $Y1$ e $Y2$, respectivamente, e um conjunto de operadores de similaridades. Nesse sentido, os operadores de similaridades podem ser de igualdade ($=$), uma função de similaridade ($\approx$), que pode ser utilizando a métrica de similaridade *q-grams*, *Jaro distance* ou *edit distance*, e um operador de *match* ($\rightleftharpoons$).

Para ilustrar, os autores utilizaram um exemplo de uma aplicação que precisa integrar dados para detecção de fraude no pagamento de cartão. A Figura 3.3 apresenta a estrutura das fontes de dados e a Tabela 3.6 mostra um conjunto fracionado de tuplas (instâncias) das fontes de dados.

Figura 3.3: Estrutura das fontes de dados.

credit (c#, SSN, FN, LN, addr, tel, email, gender, type),
billing (c#, FN, LN, post, phn, email, gender, item, price).

Fonte: Fan et al. (2009)

Tabela 3.6: Conjunto de Instâncias.

	c#	SSN	FN	LN	addr	tel	email	gender	type
t1:	111	079172485	Mark	Clifford	10 Oak Street, MH, NJ 07974	908-1111111	mc@gm.com	M	master
t2:	222	191843658	David	Smith	620 Elm Street, MH, NJ 07976	908-222222	dsmith@hm.com	M	visa

	c#	FN	LN	post	phn	email	gender	item	price
t3:	111	Marx	Clifford	10 Oak Street, MH, NJ 07974	908	mc	null	iPod	169.99
t4:	111	Marx	Clifford	NJ	908-1111111	mc	null	book	19.99
t5:	111	M.	Clivord	10 Oak Street, MH, NJ 07974	1111111	mc@gm.com	null	PSP	269.99
t6:	111	M.	Clivord	NJ	908-222222	mc@gm.com	null	CD	14.99

Adaptado de: Fan et al. (2009)

Para ilustrar o conceito de uma MD, os autores apresentaram o seguinte conjunto de MDs (Figura 3.4).

Antes de explicar as MDs, é importante destacar o operador de *match*. Ele é usado para indicar que quaisquer valores x e y, $(x \rightleftharpoons y)$, são identificados como correspondentes. Por

Figura 3.4: Conjunto de Dependências de Matching.

φ_1: credit[LN] = billing[LN] ∧ credit[addr] = billing[post] ∧
credit[FN] $\approx_d$ billing[FN] → credit[Y_c] ⇌ billing[Y_b]
φ_2: credit[tel] = billing[phn] → credit[addr] ⇌ billing[post]
φ_3: credit[email] = billing[email] →
credit[FN, LN] ⇌ billing[FN, LN]

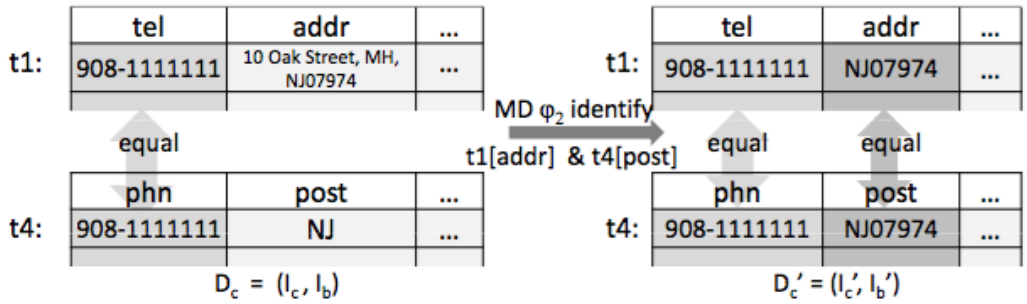
Fonte: Fan et al. (2009)

exemplo, considere a MD φ 2 na Figura 3.4, onde diz que se o valor do atributo **tel** de uma instância de *credit* for igual ao valor do atributo **phn** de uma instância de *billing*, os valores para endereço devem ser identificados como correspondentes. Para ilustrar essa situação, considere o exemplo a seguir.

Observe as instâncias t_1 e t_4 (Figura **??**). Se fossem comparadas apenas semanticamente, utilizando funções de similaridade, poderíamos afirmar que os valores para **tel** são similares. No entanto, a similaridade dos valores para **addr** seria baixa. Dessa forma, possivelmente, as instâncias não seriam dadas como correspondentes. Utilizando o conceito de dependência de *matching* proposto pelo no, é possível identificar que o atributo **tel** exerce uma dependência para com o atributo **addr**. Sendo assim, independentemente do valor de similaridade do atributo **addr**, se os valores de **tel** forem iguais, as instâncias serão dadas como correspondentes. Esse conceito é chamado de *Dynamic semantics*, ou semântica dinâmica.

Segundo a proposta do artigo, aplicando esse conceito, quando uma situação como a descrita acima ocorre, os valores dos atributos são modificados, como pode ser visto na Figura 3.5.

Figura 3.5: Exemplo - Conceito de Semântica Dinâmica.



Fonte: Fan et al. (2009)

Como contribuição, os autores propuseram um algoritmo chamado *MDClosure*. Esse algoritmo permite que múltiplas MDs possam ser inferidas com base em um conjunto de MDs previamente definidas.

Uma vez que as dependências de *matching* foram criadas e/ou inferidas, são geradas as chaves candidatas relativas, ou RCK. Uma RCK é definida com base em uma MD. Nela são especificados quais atributos devem ser considerados e como eles podem ser comparados, ou

seja, quais operadores de similaridade utilizar. A Figura 3.6 apresenta o exemplo de um conjunto de RCKs.

Figura 3.6: Exemplo - Conjunto de RCKs.

rck_1: ([LN, addr, FN], [LN, post, FN] $\|$ $[=, =, \approx_d]$)
rck_2: ([LN, tel, FN], [LN, phn, FN] $\|$ $[=, =, \approx_d]$)
rck_3: ([email, addr], [email, post] $\|$ $[=, =]$)
rck_4: ([email, tel], [email, phn] $\|$ $[=, =]$)

Fonte: Fan et al. (2009)

Por exemplo, se fosse utilizada a rck_2 os atributos a serem comparados seriam: LN/LN, utilizando o operador de igualdade, tel/phn, utilizando operador de igualdade e FN/FN, utilizando uma função de similaridade.

Para gerar as RCKs foi proposto um algoritmo chamado *findRCKs*. É importante salientar que dependendo do tamanho do conjunto de MDs utilizados o número de RCKs pode ser alto, então além de gerar as RCKs o algoritmo faz um filtro pelas melhores RCKs. Dessa forma, além de um conjunto de MDs, o algoritmo recebe como parâmetro um tamanho *m* de RCKs que deverão ser retornados. Também é importante destacar que o filtro considera alguns critérios como custo, tamanho, entre outros.

Para validar a abordagem proposta, os autores realizaram 4 experimentos. O primeiro tinha como objetivo avaliar a escalabilidade dos algoritmos propostos. Para isso, foram montados 10 cenários. O primeiro cenário considerou 2.000 MDs, o segundo 4.000 MDs, até chegar no último com 20.000. As MDs foram geradas considerando 4 diferentes conjuntos de atributos, onde o tamanho do conjunto de atributos era incrementado em 2, começando com 6 até chegar em 12.

Para avaliar o algoritmo *findRCKs* também foram utilizados 10 cenários. Em cada cenário foi considerado um tamanho *m* de RCKs geradas. O primeiro cenário iniciou com 5 RKCs e foi incrementando em 5, até chegar no último com 50. As RCKs também foram geradas considerando 4 diferentes conjuntos de atributos, seguindo a mesma configuração dita anteriormente.

Os experimentos 2 e 3 foram semelhantes. O primeiro considerando o método de *Fellegi-Sunter (FS)* e o segundo considerando o método *Sorted Neighborhood (SN)* para a Resolução de Entidades. Ambos tinham como objetivo comparar os resultados da Resolução de Entidades, considerando apenas a utilização de semântica, ou seja, sem as RCKs, e depois considerando o uso das RCKs. Para avaliar a qualidade do processo utilizaram as métricas de Precisão e Revocação e compararam também o tempo de execução.

Por fim, no experimento 4 foram realizados testes utilizando uma técnica de blocagem, antes da comparação de instâncias. O objetivo do experimento era verificar o quanto de tempo era possível reduzir aplicando a técnica. Foram comparadas as duas situações: com e sem RCKs.

É importante destacar que os experimentos 2, 3 e 4 foram realizados em 8 cenários, onde

em cada cenário variava o total de instâncias consideradas. O primeiro cenário utilizou 10.000 instâncias e foi incrementando com 10.000 até chegar no último cenário com 80.000. Em cada cenário a taxa de duplicação foi de 80%.

Como resultado, os autores comprovaram que os algoritmos *findRCKs* e *MDClosure* são escaláveis e eficientes, e destacaram que aplicando a abordagem proposta foi possível alcançar um ganho médio de 20% em precisão e revocação e 30% em desempenho, utilizando os métodos FS e SN. Os autores acreditam a proposta pode gerar uma ferramenta promissora para melhorar a qualidade e eficiência do processo de Resolução de Entidades.

3.1.3 Discussões

Dada a visão geral dos trabalhos relacionados, é possível realizar uma comparação e elencar as principais diferenças entre as estratégias propostas. As principais características dos trabalhos relacionados são apresentadas na Tabela 3.7.

No trabalho de (CHEN et al., 2012), o objetivo geral é a proposta de um método para Resolução de Entidades automático, onde é possível encontrar os atributos que serão utilizados na comparação, a função de similaridade ideal, e o limiar adequado. Sua proposta utiliza o resultado obtido por meio da Resolução de Entidades para avaliar os atributos, as funções e os limiares.

A proposta é baseada em aprendizagem de máquina, e para isso utiliza um conjunto de treinamento. Obter um conjunto de treinamento bom para a Resolução de Entidades em diferentes domínios, se torna uma tarefa difícil. Outro ponto importante a destacar, é que no cenário de grandes volumes de dados em que nos encontramos, a Resolução de Entidades tende a deixar de ser um processo *offline*, e se torna muitas vezes um processo que deve ser executado em tempo real. Com isso, o tempo de processamento precisa ser baixo, o que pode não ocorrer quando se utilizam algoritmos de aprendizagem sobre conjuntos de treinamento.

Já em (SU et al., 2010), é proposta uma técnica para ajustar os pesos dos atributos no cálculo de similaridade da fase de comparação, de acordo com a importância relacionada a esse atributo. Desse modo, diferentemente de (CHEN et al., 2012), o objetivo não é selecionar alguns atributos, e sim, ponderar todos os atributos disponíveis. Para isso, um algoritmo de aprendizado é proposto. No entanto, os autores destacam que o trabalho é não supervisionado, já que a amostra de dados utilizada é universal, contendo instâncias de diferentes fontes de dados. Para avaliar o peso dos atributos, são utilizados vetores de dados duplicados e não duplicados, e com base neles os pesos dos atributos são atribuídos.

Diferentemente dos dois trabalhos anteriores, em (FAN et al., 2009) é proposta uma abordagem para realizar o processo de Resolução de Entidades considerando dependências de atributos. Os autores afirmam que a comparação semântica, utilizada de maneira exclusiva, nem sempre é suficiente para identificar a correspondência entre pares de instâncias, sendo essa a maior justificativa para a proposta do artigo. Os autores propuseram dois algoritmos, o primeiro

Tabela 3.7: Resumo das principais características das Estratégias de Seleção de atributos.

Trabalhos	**Objetivos**	**Grandes volumes de dados**	**Base/ Conjunto de Dados**	**Aprendiza-gem de Máquina**	**Método de avaliação da Relevância dos Atributos**
[Chen et al.]	Método para Resolução de Entidades, que seleciona o grupo de atributos mais relevante, a função de similaridade e o limiar adequados.	Não	Base	Supervisio-nado	Resultado do processo de Resolução de Entidades - F-measure
[Su et al.]	Técnica pra ajustar os pesos dos atributos para o cálculo de similaridade entre duas instâncias.	Sim	Conjunto de Dados	Não supervisi-onado	Vetores de dados duplicados e não duplicados - similaridade dos valores dos atributos
[Fan et al.]	Selecionar atributos para criar regras a serem utilizadas na comparação utilizando o conceito de dependências.	Não	Base	-	Semântica dos dados/Conceito de dependências

para inferir dependência de *matching* entre os atributos e o segundo para identificar chaves candidatas relativas para o processo de Resolução de Entidades.

A maior limitação da abordagem proposta é que é necessário fornecer como entrada para o primeiro algoritmo um conjunto de MDs definidas previamente, o que não é uma tarefa trivial, considerando que para a descrição de uma MD é necessário ter um conhecimento do domínio e comportamento dos dados bem como da relação que um atributo possui com outro. Nem sempre isso é possível, uma vez que em muitas aplicações as fontes de dados são desconhecidas, no que se refere ao seu comportamento e estrutura. Além disso, a abordagem proposta é muito complexa, podendo ser difícil de ser implementada e replicada em diferentes domínios de aplicações.

3.2 Considerações

Este capítulo apresentou uma introdução a algumas estratégias de seleção de atributos propostas nas áreas de Mineração de Dados e Aprendizagem de Máquina. Posteriormente, de forma detalhada, foram apresentados trabalhos que tratam da seleção de atributos para o processo de Resolução de Entidades.

Uma análise comparativa entre os trabalhos apresentados foi realizada, enfatizando as principais características e diferenças dos mesmos.

O próximo capítulo descreve detalhadamente a estratégia de seleção de atributos para o processo de Resolução de Entidades proposta neste trabalho.

4

Uma Estratégia para Seleção de Atributos Baseada em Critérios de Avaliação da Relevância

A seleção de atributos vem demandando muitas pesquisas em áreas da Ciência da Computação, como Integração de Dados, Mineração de Dados e Inteligência Artificial. Quando se trata de Integração de Dados, mais especificamente na etapa de Resolução de Entidades, selecionar os melhores atributos para a comparação de pares de instâncias se torna um desafio.

No processo de Resolução de Entidades, a similaridade calculada entre as instâncias depende diretamente dos atributos que serão considerados para a comparação. Sendo assim, a qualidade do resultado do processo de Resolução de Entidades é diretamente afetada pela escolha dos atributos para a etapa de comparação de instâncias.

Considere como exemplo a base de dados Cora[1], que contém instâncias provenientes de múltiplas fontes de dados. As instâncias contidas no Cora são descritas pelos seguintes atributos: *id, author, title, journal, volume, pages, year, publisher, address, note, venue, editor, type, institution, month* e *class*. As perguntas que podemos fazer são: *"Todos esses atributos são relevantes na comparação de instâncias do processo de Resolução de Entidades? Se não, quais iremos utilizar na comparação?"*

Sabendo que o resultado do processo de Resolução de Entidades depende diretamente dos atributos utilizados na comparação, torna-se crucial o uso de estratégias capazes de selecionar os melhores atributos para serem considerados na fase de comparação.

Neste capítulo propomos uma estratégia de seleção de atributos relevantes para o processo de Resolução de Entidades. A seleção de atributos proposta considera critérios relacionados aos dados, e às fontes, para avaliar a relevância dos atributos.

Este capítulo está estruturado da seguinte maneira: A Seção 4.1 apresenta algumas definições preliminares, acerca dos principais conceitos relacionados à estratégia de seleção de atributos proposta. A Seção 4.2 apresenta uma visão geral da estratégia proposta neste

[1] Originalmente de https://people.cs.umass.edu/ mccallum/code-data.html

trabalho. Na Seção 4.3 são apresentados os critérios de avaliação da relevância dos atributos, propostos nesta dissertação. A Seção 4.4 apresenta as etapas da estratégia proposta, detalhando a avaliação das Relevâncias Individual e Global dos atributos, e apresentando também os algoritmos propostos para esta avaliação. Na Seção 4.5 apresentamos um exemplo para facilitar o entendimento da estratégia proposta. Por fim, as conclusões do capítulo são apresentadas na Seção 4.7.

4.1 Definições Preliminares

A estratégia proposta neste trabalho está inserida em um cenário de integração de dados distribuídos em múltiplas fontes de dados, cujo processo de Integração de Dados se divide em três etapas: Correspondência de Esquemas, Resolução de Entidades e Fusão de Dados, sendo o nosso foco a etapa de Resolução de Entidades.

Consideramos também que neste ambiente de Integração de Dados existe um catálogo (Oliveira et al., 2015) onde as fontes de dados disponíveis são registradas. São armazenadas informações como: nome, url, esquema (contendo entidades e atributos), e metadados de qualidade, sendo um deles a confiabilidade da fonte. Como parte do processo de especificação da estratégia de seleção de atributos, propomos as definições apresentadas a seguir. Estas definições dizem respeito aos principais conceitos empregados no processo de seleção de atributos.

Definição 4.1. Conjunto de Fontes de Dados. Seja $F = \{f_1, f_2, ...f_n\}$ um conjunto de fontes de dados, tal que cada f_i oferece dados sobre um ou mais conceitos do mundo real.

Definição 4.2. Conceito. Um conceito C em uma fonte de dados f_i, denotado por $f_i.C$, está associado a um conjunto de entidades $f_i.C = \{e_1, e_2, ...e_m\}$, tal que cada e_j representa uma instância de uma entidade do mundo real, e é descrito por um conjunto de atributos $C.A_i = \{a_{i1}, a_{i2}, ...a_{ik}\}$.

Definição 4.3. Entidade. Uma entidade e_j de um conceito C na fonte f_i, denotada por $f_i.C.e_j$, é definida por um conjunto de pares $\{(a_{i1}, v_1), (a_{i2}, v_2), ..., (a_{ik}, v_k)\}$, tal que $a_{ik} \in C.A_i$, e v_i é o valor de a_{ik} para a entidade $f_i.C.e_j$.

Definição 4.4. Conjunto de Dados. Seja $D = \{e_1, e_2, ...e_n\}$ um conjunto de entidades oriundo do conjunto de fontes de dados F, onde cada e_j representa uma entidade do mesmo conceito C.

Definição 4.5. Conjunto de Atributos Comuns. Dado o conjunto de dados D, o conjunto de atributos comuns às entidades de D é denotado por $A_{int} = \{f_1.C.A_1 \cap f_2.C.A_2 \cap ... \cap f_n.C.A_n\}$.

Definição 4.6. Relevância Individual de um atributo. Seja a_{ij} um atributo pertencente ao conjunto de atributos comuns A_{int}, a Relevância Individual de um atributo $R_{ind}(a_{ij})$ é o valor obtido utilizando a equação para o cálculo de Relevância Individual, que leva em consideração os critérios de avaliação da relevância dos atributos.

Definição 4.7. Relevância Global de um atributo. Seja a_{ij} um atributo pertencente ao conjunto de atributos comuns A_{int}, a Relevância Global de um atributo $R_{glob}(a_{ij})$ é o valor obtido utilizando a equação para o cálculo de relevância global, que leva em consideração $R_{ind}(a_{ij})$ e metadados de qualidade relacionados à fonte.

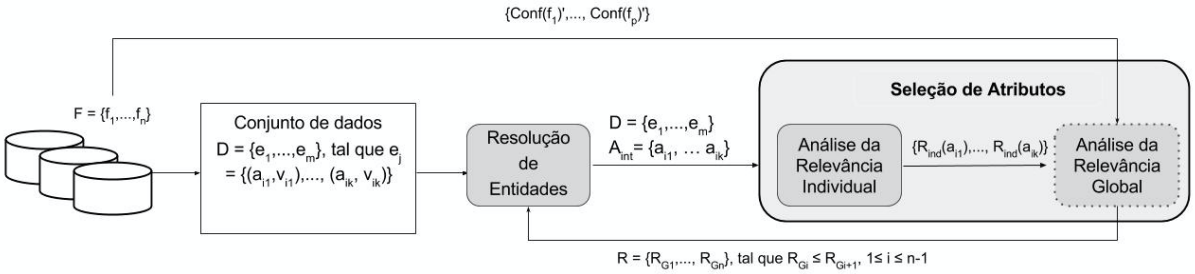
Considerando que, no processo de Resolução de Entidades, a avaliação da correspondência entre duas instâncias é feita a partir da combinação dos valores de similaridade entre os atributos que as descrevem, é importante utilizar apenas atributos que contribuam positivamente para a identificação de correspondências verdadeiras. Em outras palavras, atributos que contribuem para a identificação de falsos positivos ou falsos negativos não devem ser considerados na comparação.

Dado um conjunto de entidades que representam um conceito C, que esse conceito é descrito por um conjunto de atributos A, faz-se necessário selecionar os atributos mais relevantes do conjunto A, que descrevem esse conceito e que serão considerados na fase de comparação do processo de Resolução de Entidades.

4.2 Visão Geral da Estratégia de Seleção de Atributos

A estratégia de seleção de atributos proposta neste trabalho consiste de duas etapas: **(i)** Análise da Relevância Individual dos atributos; **(ii)** Análise da Relevância Global dos atributos, conforme mostra a Figura 4.1, e descrito a seguir.

Figura 4.1: Visão geral da Estratégia de Seleção de Atributos.



Como entrada da seleção de atributos tem-se um conjunto de entidades D referentes a um mesmo conceito. Cada conceito é descrito por um conjunto de atributos A_i, e contém um conjunto de entidades $f_i.C$ em uma dada fonte de dados f_i.

O conjunto de dados D, devido a ser oriundo de múltiplas fontes contidas no conjunto de fontes de dados F, possivelmente contém dados duplicados e precisa ser integrado. Logo, neste sentido, faz-se necessário realizar o processo de Resolução de Entidades.

Para selecionar os atributos relevantes, primeiramente é necessário identificar os atributos comuns a todas as entidades envolvidas na Resolução de Entidades (GU et al., 2003). Portanto, dos conceitos representados pelas entidades contidas em D encontra-se A_{int}, conjunto de atributos comuns a todas os conceitos das entidades do conjunto de dados D.

Para analisar a Relevância Individual (R_{Ind}) de cada atributo, propomos avaliar características relacionadas aos valores de dados. Para analisar a Relevância Global (R_{Glob}), propomos avaliar características relacionadas às fontes de dados que oferecem entidades para o conjunto de dados avaliado. Para atribuir os valores de relevância individual e global dos atributos, elencamos critérios para avaliar as características citadas. Os critérios elencados serão detalhados na próxima seção.

Como saída do processo de seleção de atributos, tem-se uma classificação ordenada $R = \{RG_1, ..., RG_n\}$. Cada RG_i contém um par (a_{ij}, r_j), em que a_{ij} é o atributo, e r_j é o valor de Relevância Global do atributo $a_{ij} \in A_{int}$, onde r_j de $RG_i \geq r_j$ de RG_{i+1}, tal que $1 \leq i \leq n-1$.

4.3 Critérios de Avaliação

O conceito de atributo relevante pode ter diversas interpretações. Para identificar se um atributo é relevante, primeiramente é necessário se estabelecer o conceito de atributo relevante para a tarefa que será realizada, para depois definir os critérios que o tornam relevante.

As características dos atributos que afetam a decisão de seleção incluem o nível de erros nos valores dos atributos e o número (e distribuição) dos valores dos atributos, ou seja, o conteúdo informativo do atributo. Por exemplo, um campo como o sexo só tem dois possíveis valores e, consequentemente, não poderia dar informações suficientes para identificar a similaridade entre instâncias. Por outro lado, um atributo como o sobrenome contém muito mais informação, mas pode ser frequentemente registrado incorretamente (GU et al., 2003).

A relevância de um atributo é proporcional a sua contribuição em discriminar instâncias que pertencem a entidades diferentes e em não discriminar instâncias que pertencem à mesma entidade.

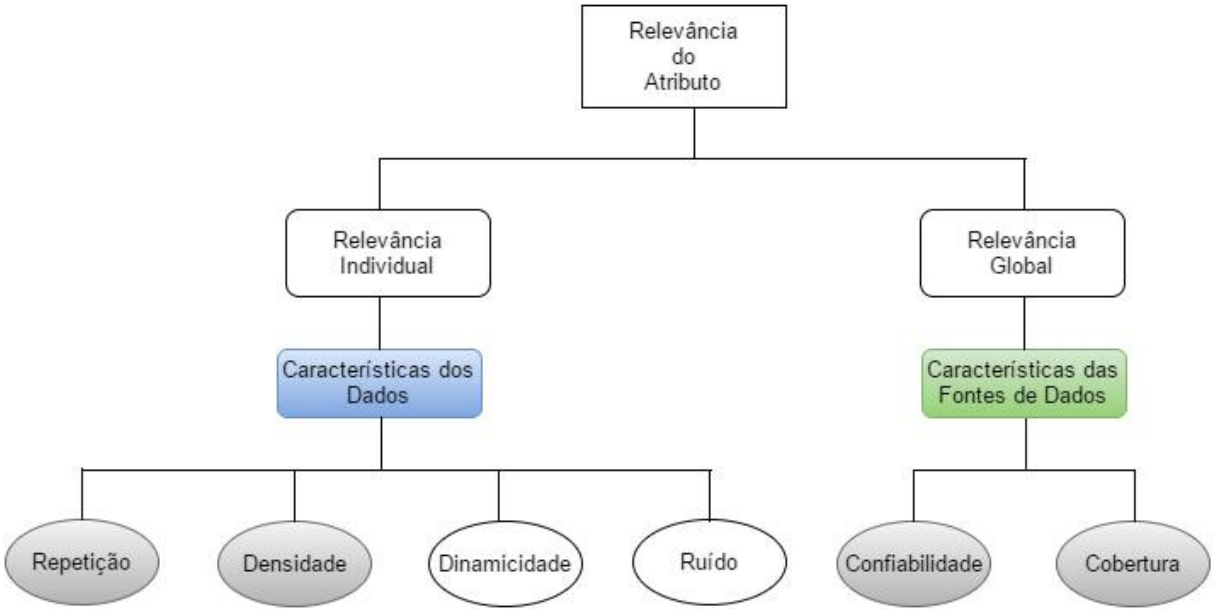
No entanto, acreditamos que a qualidade das fontes que oferecem os dados para o conjunto de dados avaliado pode exercer influência direta na relevância de um atributo. Por exemplo, uma fonte de dados é passível de fornecer dados que não sejam verdadeiros. Nesse sentido, ponderar a relevância de um atributo com características inerentes à qualidade das fontes de dados, pode ajudar a verificar se o resultado da seleção de atributos é realmente confiável.

Desse modo, como uma das contribuições desta pesquisa, de maneira empírica, elencamos seis critérios (Figura 4.2), sendo quatro deles relacionados a características dos dados, com o objetivo de calcular a Relevância Individual de um atributo, e dois relacionados às fontes, com o objetivo de calcular a Relevância Global. A seguir, apresentamos uma visão geral de cada critério.

Critérios de avaliação relacionados a características dos dados

- **Repetição -** Este critério tem como objetivo avaliar a quantidade de valores repetidos nos dados. Acreditamos que um atributo com alto valor de repetição não contribui na discriminação de instâncias, podendo ocasionar em falsos positivos. (Mais detalhes serão apresentados na Seção 4.3.1.1)

Figura 4.2: Critérios de Avaliação.



- **Densidade -** Este critério tem como objetivo avaliar a quantidade de valores nulos nos dados. Quanto mais denso um atributo for, ou seja, quanto menos valores nulos ele tiver, maior a probabilidade de contribuir para a identificação de correspondências corretas. (Mais detalhes serão apresentados na Seção 4.3.1.2)

- **Dinamicidade -** Este critério tem como objetivo avaliar o grau de dinamicidade dos atributos, ou seja, seu nível de alteração ao longo do tempo. A dinamicidade de um atributo é sua tendência a sofrer alterações ao longo do tempo. Por exemplo, se duas instâncias provenientes de duas fontes de dados distintas, mas referentes à mesma entidade do conceito *Pessoa*, forem comparadas, não é difícil encontrar casos em que uma mesma pessoa tenha diferentes endereços (já que as pessoas podem mudar de endereços algumas vezes na vida), ou mesmo nomes diferentes (em virtude de casamentos ou divórcios).

  Se um atributo dinâmico for utilizado na comparação entre pares de instâncias, comparar um valor antigo do atributo com um valor novo, para duas instâncias que se referem à mesma entidade, pode não gerar um resultado correto. Isso ocorre devido a um atributo dinâmico contribuir para a conclusão de que duas instâncias dizem respeito a entidades diferentes quando na realidade dizem respeito a uma mesma entidade, fazendo com que na etapa de comparação não seja possível identificá-las corretamente, gerando um falso negativo.

  Tome como exemplo duas entidades que se referem a mesma pessoa. Ambas se referem a “Maria de Souza Silva”, como mostra a Tabela 4.1.

  Se os atributos *idade* ou *telefone*, fossem utilizados na comparação entre as instâncias, possivelmente teríamos como resultado que a instância 1 não corresponde à

Tabela 4.1: Conjunto de dados contendo entidades referentes ao conceito pessoa.

Id	Nome	Idade	Cpf	Telefone	Sexo
1	Maria de Souza Silva	29	201.202	9333-0022	Feminino
2	Maria Sousa	30	201.202	8722-0234	Feminino

instância 2, sendo que as duas instâncias se referem a mesma pessoa. Isso ocasionaria uma correspondência errada, falso negativo, afetando negativamente o resultado do processo de Resolução de Entidades. Diferentemente do atributo *Cpf*, que nunca irá sofrer alterações, sendo único para cada pessoa.

- **Ruído -** Este critério tem como objetivo analisar o quanto um atributo é suscetível a erros. O ruído é um erro aleatório ou variabilidade presente nos valores de entrada de um atributo. Os erros nos valores podem ocorrer devido a vários motivos, por exemplo, problemas na entrada de dados, inconsistência na convenção de nomes, ou simplesmente erros de digitação.

  Considerar o ruído também é importante quando se analisa a relevância de um atributo, já que atributos muito suscetíveis a erros de digitação podem fazer com que o resultado de uma comparação entre duas instâncias seja incorreto.

  Para exemplificar, considere novamente a Tabela 4.1. Supondo que o valor correto para o atributo *nome* seja "Maria de Souza Silva", podemos verificar que o valor de entrada para a instância com *Id* 2, foi inserido incorretamente.

  Se utilizássemos o atributo *nome*, que contém um valor errôneo na comparação, esse atributo não identificaria corretamente as duas instâncias, retornando uma correspondência errada, ou seja, um falso negativo. Devido a isso, utilizar atributos muito suscetíveis a ruído pode gerar correspondências incorretas afetando negativamente no processo de Resolução de Entidades.

Critérios de avaliação relacionados a características das Fontes de Dados

- **Confiabilidade -** Este critério tem como objetivo verificar o grau em que os dados fornecidos por uma fonte de dados são verídicos e confiáveis. (Mais detalhes serão apresentados na Seção 4.3.2.1)

- **Cobertura -** Este critério tem como objetivo identificar o percentual de instâncias que uma fonte oferece ao conjunto de dados avaliado. (Mais detalhes serão apresentados na Seção 4.3.2.2)

Especificamente, neste trabalho, iremos considerar apenas os critérios Densidade e Repetição, para avaliar características dos dados, e Confiabilidade e Cobertura, para avaliar características das fontes. Os critérios foram escolhidos por serem adaptáveis a qualquer cenário

de aplicação, ou seja, fazem uma avaliação mais geral do conjunto de dados considerado. Os demais critérios, apesar de serem importantes, são aplicáveis em alguns cenários específicos.

Na próxima seção, os critérios utilizados na estratégia de seleção de atributos proposta neste trabalho serão detalhados. Também serão apresentadas as métricas utilizadas para calcular cada critério.

4.3.1 Critérios de Avaliação da Relevância Individual

Nas seções a seguir, considere um conjunto de dados D que contém entidades de um conceito C, e sendo A_{int} o conjunto de atributos comuns das entidades de D.

4.3.1.1 Repetição

A Repetição de um atributo a_{ij} é dada pela quantidade de vezes que um mesmo valor para o atributo aparece no conjunto de dados. Um atributo possui uma alta repetição quando seu valor aparece muitas vezes no conjunto de dados, e baixa repetição caso contrário.

A escolha do critério de Repetição para avaliar a relevância individual de um atributo na seleção de atributos foi motivada pelo fato de que utilizar atributos com alta repetição de valores para a comparação entre pares de instâncias pode contribuir para a geração de falsos positivos, ou seja, instâncias distintas classificadas como similares.

Tabela 4.2: Conjunto de dados contendo entidades referentes ao conceito Pessoa.

ID	Nome	Idade	Telefone	Estado Civil
1	Maria Suzana	41	(81) 9893-0029	Casado(a)
2	Suilan Maria	42	(81) 9832-3311	Casado(a)
3	Suilan Maria	41	NULL	Casado(a)

Tome como exemplo o conjunto de dados apresentado na Tabela 4.2. Considerando que o atributo *Estado Civil* do conceito *Pessoa* pode conter os seguintes valores: solteiro(a), casado(a), divorciado(a) ou viúvo(a). Duas entidades que correspondem a pessoas distintas mas que contêm os mesmos valores para *Estado Civil* poderiam ser consideradas entidades similares pela comparação, gerando assim uma correspondência incorreta. Neste exemplo, se o subconjunto de atributos *Nome*, *Idade* e *Estado Civil* fosse utilizado na comparação, possivelmente as instâncias 1 e 3 seriam consideradas correspondentes, quando na verdade são distintas. Portanto, o atributo *Estado civil* por possuir muita repetição, não contribuiria de forma significativa para a comparação entre as instâncias.

Neste trabalho, o valor para o critério de repetição para cada $a_{ij} \in A_{int}$ é calculado conforme a Equação 4.1.

$$Rep(a_{ij}) = 1 - \left(\frac{\tau}{\eta}\right) \tag{4.1}$$

Onde τ é o total de valores distintos de a_{ij}, e η é o número total de valores de a_{ij}. O cálculo do total de valores distintos é implementado utilizando uma função de similaridade. Existem inúmeras funções propostas na literatura (JACCARD, 1901; LEVENSHTEIN, 1966; JARO, 1989).

Neste trabalho, foi adotada a função de similaridade de *Levenshtein*, ou *edit distance*, por ser uma das funções mais utilizadas quando se deseja comparar *strings* relativamente pequenas e que não precisam necessariamente ter o mesmo tamanho. Utilizando a *edit distance*, a similaridade entre duas *strings* é dada pelo menor número de operações necessárias para transformar uma *string* em outra, onde uma operação pode ser uma inserção, deleção ou substituição de um caractere. A função de cálculo da similaridade é definida pela Equação 4.2.

$$edSim\left(S_i, S_j\right) = \frac{ed\left(S_i, S_j\right)}{min\left(\left|S_i\right|, \left|S_j\right|\right)} \tag{4.2}$$

Onde $edSim\left(S_i, S_j\right)$ é a função de cálculo da distância de edição, e $min\left(\left|S_i\right|, \left|S_j\right|\right)$ retorna o tamanho mínimo das cadeias de caracteres.

4.3.1.2 Densidade

A Densidade de um atributo a_{ij} é dada pelo percentual de valores não nulos contidos no conjunto de valores que descreve esse atributo (NAUMANN; FREYTAG; LESER, 2000). A ausência de valor em um atributo pode fazer com que, duas instâncias que são correspondentes sejam dadas como distintas, uma vez que, em alguns casos quando se compara atributos com valores ausentes, a similaridade é igual a 0. Dessa forma, a ausência de valores pode contribuir para a identificação de falsos negativos, ou seja, instâncias similares classificadas como distintas.

Para ilustrar o critério de Densidade, considere o conjunto de dados da Tabela 4.2. Se o subconjunto de atributos contendo *Nome*, *Idade* e *Telefone* fosse utilizado na comparação entre instâncias, possivelmente as instâncias 2 e 3 seriam dadas como não correspondentes, mesmo sendo duplicadas. Dessa forma, pode-se notar que um atributo que contém valores nulos ou ausentes não traz ganhos ao processo de Resolução de Entidades. Sendo assim, a densidade nos dados também é um aspecto a ser analisado ao se avaliar a relevância de um atributo.

O valor de Densidade para cada $a_{ij} \in A_{int}$ é calculado por meio da Equação 4.3.

$$Den(a_{ij}) = \frac{\alpha}{\beta} \tag{4.3}$$

$$Den(A_i) = \frac{nao - nulos}{total} \tag{4.4}$$

Onde α é referente ao total de valores não nulos de a_{ij}, e β é referente ao total de valores de a_{ij}.

4.3.2 Critérios de Avaliação da Relevância Global

Nas seções a seguir, considere que um conjunto de fontes de dados F oferece instâncias para o conjunto de dados S. Para cada fonte contida em F, os critérios Confiabilidade e Cobertura serão obtidos, como descrito a seguir.

4.3.2.1 Confiabilidade

A Confiabilidade de uma fonte diz respeito ao grau em que os dados fornecidos por ela são verídicos e confiáveis (WANG; STRONG, 1996). Não faz parte deste trabalho calcular valores de confiabilidade das fontes. De forma semelhante ao trabalho de Mihaila, Raschid e Vidal (2000), assumimos que, quando uma fonte de dados é catalogada, informações de qualidade associadas a elas estão presentes. Nesse sentido, o valor de confiabilidade é extraído desses metadados.

O valor de confiabilidade, é denotado por $Conf(f_k)$, tal que, $0 \leq Conf(f_k) \leq 1$, e pode ser obtido por meio dos metadados de qualidade associados as fontes.

4.3.2.2 Cobertura

A Cobertura de uma fonte de dados é definida pelo percentual de instâncias que ela fornece para o conjunto de dados avaliado. Para isso, neste trabalho utiliza-se a métrica de avaliação proposta em (NAUMANN; FREYTAG; LESER, 2000) (Equação 4.8).

Para o cálculo da Cobertura de uma fonte de dados f_k, denotada por $Cob(f_k)$, é necessário dividir o total de instâncias que uma fonte de dados fornece para D, denotado por π, pelo total de instâncias contidas em D, denotado por $|D|$.

$$Cob(f_k) = \frac{\pi}{|D|} \tag{4.5}$$

$$Cob(f_k) = \frac{instancias\_fornecidas}{total\_instancias} \tag{4.6}$$

$$Conf(f_k) = metadados\_qualidade \tag{4.7}$$

$$Den(A_i) = \frac{nao\_nulos}{total} \tag{4.8}$$

4.4 Etapas do Processo de Seleção de Atributos

Calcular a relevância de um atributo é necessário para selecionar quais os atributos mais relevantes para serem considerados na fase de comparação do processo de Resolução de

Entidades. Para esse fim, propomos calcular a Relevância Individual e a Relevância Global de cada atributo.

A seguir, detalharemos as etapas do processo de seleção de atributos proposto neste trabalho.

4.4.1 Análise de Relevância Individual

A Relevância Individual R_{ind} de um atributo a_{ij} é calculada com base nos critérios de avaliação de relevância descritos na seção anterior. Neste trabalho, apenas os critérios de Repetição e Densidade foram considerados. Deste modo, a relevância individual de um atributo é calculada de acordo com a Equação 4.9.

$$R_{ind}(a_{ij}) = Den(a_{ij}) * p_d + (1 - Rep(a_{ij})) * p_r \tag{4.9}$$

Onde $Den(a_{ij})$ é o valor do critério de Densidade para o atributo a_{ij}, $Rep(a_{ij})$ é o valor do critério de Repetição para o atributo a_{ij}, p_d é o peso para o critério de Densidade, e p_r é o peso para o critério de Repetição. O critério de repetição, é um critério de custo, quanto mais próximo de 0 melhor, e o critério de densidade, é um critério de qualidade, ou seja, quanto mais próximo de 1 melhor. Dessa forma, para uniformizar a fórmula, se faz necessário o $1-$ antes do critério de repetição.

A atribuição de um valor p para o peso de um critério deve ser feita conforme o grau de importância de tal critério para o cálculo da relevância individual, de tal forma que a soma dos pesos deve ser igual a 1. Portanto, seu valor deve ser distribuído entre os critérios utilizados no cálculo da relevância individual, podendo ser ajustado de acordo com a necessidade do usuário.

Para o cálculo da Relevância Individual dos atributos, propomos o Algoritmo 1, que será detalhado a seguir.

- A entrada do algoritmo é um conjunto de dados D, e o conjunto de atributos comuns A_{int} a todas as instâncias, tal que $a_{ij} \in A_{int}$, onde a corresponde a um atributo. Como entrada também temos os valores dos pesos correspondentes a cada critério. A saída do algoritmo é um vetor contendo os valores de relevância individual para cada atributo do conjunto A_{int}.
- Rep recebe os valores de repetição de cada atributo contido em A_{int} (linha 5).
- Den recebe os valores de densidade de cada atributo contido em A_{int} (linha 6).
- Para cada entidade $e_j \in D$ e para cada atributo $a_{ij} \in A_{int}$ o laço é executado e a relevância individual é calculada utilizando a equação mostrada anteriormente. (linhas 7-14)
- Por fim, é retornado um vetor contendo os valores de relevância individual de cada atributo de A_{int}. (linha 15)

Algoritmo 1: Algoritmo CalculaRelevanciaIndividual

Entrada: D: Conjunto de dados; A_{int} conjunto de atributos comuns das entidades de D, tal que $a \in A_{int}$, onde a corresponde a um atributo,
$Peso_R$: Valor do peso para o critério de Repetição,
$Peso_D$: Valor do peso para o critério de Densidade
Saída: R_{ind}: Vetor com os valores da Relevância Individual de cada atributo do conjunto de atributos A_{int}

1 $Rep \leftarrow \emptyset$;
2 $Den \leftarrow \emptyset$;
3 $R_{ind} \leftarrow \emptyset$;
4 **início**
5 Rep ← CalculaRepeticao(D) ; `// recebe o vetor com os valores de repetição dos atributos`
6 Den ← CalculaDensidade(D) ; `// recebe o vetor com os valores de incompletude dos atributos`
7 **para cada** $e_j \in D$ **faça**
`// executa o laço para cada instância de entidade contida no conjunto de dados`
8 **para cada** $a_{ij} \in A_{int}$ **faça**
`// executa o laço para cada atributo contido no conjunto de atributos comuns`
9 $R_{ind}[i] \leftarrow Den[a_{ij}]) \times Peso_{Den} + (1 - Rep[a_{ij}]) \times Peso_{Rep}$;
10 **fim**
11 $i \leftarrow i + 1$;
12 **fim**
13 **retorna** R_{ind}
14 **fim**

4.4.2 Análise de Relevância Global

A Relevância Global de um atributo $a_{ij} \in A_{int}$ é denotada por $R_{glob}(a_{ij})$. O objetivo da $R_{glob}(a_{ij})$ é ponderar a relevância de um atributo, uma vez que no cálculo da $R_{ind}(a_{ij})$ apenas os valores dos atributos são levados em consideração. Muitas vezes isso pode não ser suficiente, visto que fontes de dados com baixa qualidade podem estar contidas no conjunto F que fornece dados para D. Sendo assim, para calcular a $R_{glob}(a_{ij})$, além do valor de $R_{ind}(a_{ij})$, consideramos também informações de qualidade de cada fonte de dados $f_k \in F$.

Para calcular a Relevância Global dos atributos, é necessário conhecer o valor de qualidade do conjunto de fontes de dados F, denotado por $Q(F)$. Esse valor é calculado com base nos valores dos critérios de Confiabilidade ($Conf(f_k)$) e Cobertura ($Cob(f_k)$) de cada fonte de dados pertencente ao conjunto F. O valor para $Q(F)$, pode ser dado pela Equação 4.10.

$$Q(F) = \sum_{k=1}^{|F|} Conf(f_k) * Cob(f_k) \tag{4.10}$$

Após calcular o valor de qualidade do conjunto de fontes de dados F, a $R_{glob}(a_{ij})$ pode

ser calculada conforme a Equação 4.12, onde $R_{ind}(a_{ij})$ é o valor da Relevância Individual do atributo a_{ij}.

$$R_{glob}\left(a_{ij}\right) = R_{ind}\left(a_{ij}\right) * Q(F) \tag{4.11}$$

$$R_{glob}\left(A_i\right) = R_{ind}\left(A_i\right) * Q(F) \tag{4.12}$$

Para exemplificar a relevância global, considere um conjunto de dados D contendo 20 instâncias, provenientes de duas fontes de dados distintas f_1 e f_2 com seus valores de confiabilidade de 0.5 e 0.8 respectivamente, onde f_1 fornece 12 instâncias, e f_2 fornece 8 instâncias. Aplicando a métrica dada pela equação 4.8, obtemos os seguintes resultados $Cob(f_1) = 0.6$ e $Cob_(f_2) = 0.4$. Sabendo os valores de confiabilidade das fontes obtidos por meio dos metadados, a qualidade do conjunto de fontes é dada como na equação 4.10, tendo um resultado de $Q(F) = 0.62$. Suponha que, depois de realizar o cálculo de relevância individual para cada $a_{ij} \in A_{int}$, o atributo A_1 obteve um valor de $R_{ind} = 0.9$. Isso ocorreu porque na maioria das instâncias o atributo a_1 não continha valores nulos ou repetidos. No entanto, a fonte que fornece a maioria das instâncias para o conjunto de dados D possui uma confiabilidade baixa. Sendo assim, a relevância dos atributos pode ser questionada. No cálculo de relevância global, o valor de R_{ind} dos atributos é ponderado, conforme a equação 4.12. Para este exemplo, $R_{glob}(a_1) = 0.9 * 0.62 = 0.55$.

Para o cálculo da Relevância Global, propomos o Algoritmo 2, e seus passos são detalhados a seguir.

Algoritmo 2: Algoritmo CalculaRelevanciaGlobal

Entrada: D: Conjunto de dados do mesmo conceito; A_{int} conjunto de atributos comuns de D, tal que $a \in A_{int}$, onde a_{ij} corresponde a um atributo,
$Q(F)$: Valor de qualidade de F; `// valor calculado utilizando a cobertura e a confiabilidade de cada` $f \in F$.
Saída: R_{glob}: Vetor com os valores da Relevância Global do conjunto de atributos A_{int}

```
1  R_glob ← ∅;
2  R_ind ← ∅;
3  início
4      R_ind ← CalculaRelevanciaIndividual(D); // vetor que recebe o
           valor de relevância individual de cada a_ij.
5      para cada a_ij ∈ A_int faça
6          // para cada atributo do conjunto de atributos
               comuns executa o laço
7          R_glob[i] ← (R_ind[i] × Q(F);
8      fim
9      i ← i + 1;
10     retorna R_glob
11 fim
```

- A entrada do algoritmo é um conjunto de dados D, e o valor de qualidade do conjunto de fontes das quais as instâncias resultaram. A saída desse algoritmo é um vetor contendo os valores de relevância global de cada atributo do conjunto A_{int}.
- R_{ind} recebe os valores de relevância individual dos atributos contidos em A_{int}, por meio da execução do algoritmo 1.
- Para cada atributo $a_{ij} \in A_{int}$, o laço é executado e a relevância global é calculada através da equação proposta anteriormente (linhas 5-7).
- Finalmente, um vetor contendo os valores da relevância global de cada atributo do conjunto A_{int} é retornado.

Dessa maneira, tem-se a Relevância Global para cada atributo do conjunto A_{int}, possibilitando analisar se os atributos são realmente confiáveis para serem utilizados na Resolução de Entidades.

4.5 Exemplo

Para facilitar o entendimento, nesta seção será apresentado um cenário de aplicação da estratégia proposta. Neste exemplo, a relevância será calculada utilizando apenas os critérios de Repetição e Densidade.

Considere um conjunto de dados contendo instâncias de publicações oriundas de duas fontes de dados, f_1 e f_2, que compõem o conjunto de fontes de dados F. Esse conjunto de dados pode ser visto na Tabela 4.3.

Tabela 4.3: Conjunto de dados contendo instâncias de publicações.

Fonte	Author	Title	Year	Address	Institution	...
Fonte 1	B. Buth et al.	Provablly Correct Compiler Implementation	1992			
	Aha, D. W, Kibler. & Albert, M	Instance-based learning algorithms	1991			
	M. Passani and D. Kibler	The utility of knowledge in inductive learning	1992		University of Massachusetts	
Fonte 2	Aha, D. W, Kibler. & Albert, M	Instance-based learning algorithms	1992			
	Paul E. Utgoff and Carla E. Brodley	Linear machine decision trees	1991	University of Massachusets		
	Dennis Kibler		1993	Amherest		
	Carla E. Brodley and Paul Utgoff	Multivariate versus univariate decisions trees				
	Fawcett, T. E & Utgof, P. E	A Hybrid Method for Feature Generation	1991			
		Planning for conjunctive goals				
	Carla E. Brodley and Paul Utgoff	Multivariate versus univariate decisions trees		University of Massachusets	Departament od Computer Science	

Supondo que é necessário integrar esse conjunto de dados, a Resolução de Entidades precisa ser realizada e, para tanto, precisamos selecionar os atributos que serão utilizados para comparar as instâncias. Utilizando a estratégia de seleção de atributos proposta nesse trabalho, inicialmente calculamos para cada atributo do conjunto de dados, os valores dos critérios de Repetição e Densidade, e atribuímos pesos iguais, de 0,5 para cada um deles.

Depois, sabendo os valores dos critérios, a relevância individual é calculada para cada atributo, conforme a Tabela 4.4.

Tomando o atributo *title* como exemplo, para exemplificar o cálculo da relevância individual, têm-se:

$$Den(title) = \frac{9}{10} = 0,9$$

$$Rep(title) = 0,19$$

$$R_{ind}(title) = ((0.9 * 0.5) + (1 - 0,19) * 0.5) = 0,85$$

Tabela 4.4: Valores de Repetição, Densidade e Relevância Individual dos atributos.

Atributos	Repetição	Densidade	Relevância Individual
Title	0,19	0,9	0,85
Author	0,19	0,9	0,85
Year	0,5	0,7	0,60
Address	0,7	0,3	0,30
Institution	0,7	0,2	0,25

Continuando com o exemplo do atributo *title*, para calcular a relevância global verificamos as informações de qualidade das fontes. Obtemos por meio dos metadados os valores de confiabilidade para cada fonte, sendo a $Conf(f_1) = 90\%$, e a $Conf(f_2) = 65\%$. Para cada fonte de dados, é calculada a cobertura com relação ao conjunto de dados, conforme a Equação 4.8, tendo o resultado abaixo.

$$Cob(f_1) = 0,3$$

$$Cob(f_2) = 0,7$$

Depois, calculamos a qualidade do conjunto de fontes de dados F, composto por $F1$ e $F2$, conforme a Equação 4.10, tendo o resultado abaixo.

$$Q(F) = 0,72$$

Com isso, para cada atributo, o valor de relevância individual é ponderado utilizando o valor de $Q(F)$, conforme o cálculo abaixo, e os resultados são mostrados na Tabela 4.5.

Continuando o exemplo anterior com o cálculo de relevância global do atributo *title*, seria calculada como abaixo.

$$R_{glob}(title) = 0,85 * 0,72 = 0,61$$

Tabela 4.5: Valor de relevância global dos atributos.

Atributos	Relevância Global
Title	0,61
Author	0,61
Year	0,43
Address	0,21
Institution	0,18

Dessa forma, os atributos são classificados de acordo com a relevância individual, e depois, para confirmar se os atributos realmente possuem uma relevância confiável, analisamos a relevância global, e os ranqueamos novamente.

4.6 Análise Comparativa

Nesta seção, iremos destacar resumidamente os diferenciais da estratégia proposta neste capítulo, em relação aos trabalhos apresentados no Capítulo 3. Os principais diferenciais da nossa estratégia podem ser vistos na Tabela 4.6

- **Objetivo -** Diferentemente do trabalho de Chen et al. (2012), não é o objetivo deste trabalho propor um método para a Resolução de Entidades. O foco desta dissertação é a proposta de uma estratégia para selecionar os atributos mais relevantes para a fase de comparação do processo de Resolução de Entidades. Para isso, a estratégia proposta classifica os atributos de acordo com o valor de relevância calculado com base nos critérios propostos. Os atributos mais bem classificados são os mais relevantes para serem utilizados na fase de comparação de instâncias.

- **Aprendizagem de Máquina-** Nossa estratégia não utiliza aprendizagem de máquina, diferentemente dos trabalhos de Chen et al. (2012) e Fan et al. (2009). A vantagem disso é que nossa estratégia é facilmente aplicável a múltiplos domínios. Além do mais, um conjunto de treinamento torna o processo custoso, o que para Resolução de Entidades sobre resultados de consultas a bases de dados da *Web*, motivação do nosso trabalho, seria inviável.

- **Conjunto de Dados/Base de Dados -** Outro diferencial da nossa estratégia é que a seleção de atributos é realizada sobre um conjunto de dados (resultado de consulta), semelhante ao trabalho de Su et al. (2010), e não para a base de dados completa. No entanto, diferente de Su et al. (2010), nós analisamos características dos dados e das fontes de dados para analisar a relevância dos atributos.

- **Método de Avaliação da Relevância dos Atributos -** A estratégia de seleção de atributos proposta neste trabalho propõe critérios de avaliação para calcular a rele-

vância dos atributos. Para o cálculo de Relevância Individual, são utilizados critérios relacionados aos dados, e para o cálculo da Relevância Global, critérios relacionados às fontes.

4.7 Considerações

Neste capítulo apresentamos a estratégia de seleção de atributos proposta neste trabalho. Foram apresentadas algumas definições preliminares acerca da estratégia proposta, com o intuito de facilitar o entendimento do processo de seleção de atributos. O problema de seleção de atributos para o processo de Resolução de Entidades foi formalizado. Também foi apresentada uma visão geral do processo de seleção de atributos, detalhando as etapas da estratégia proposta.

Os critérios de qualidade que determinam a relevância individual dos atributos foram detalhados, juntamente com a maneira que são calculados. Apresentamos também as equações propostas para o cálculo de relevância individual e relevância global dos atributos, e os algoritmos que as implementam.

Ainda neste capítulo, exemplificamos a estratégia proposta, calculando a relevância individual e global dos atributos considerando um conjunto de dados exemplo, possibilitando maior entendimento da estratégia de seleção de atributos proposta.

No próximo capítulo apresentaremos um estudo de caso com a implementação da estratégia de seleção de atributos proposta neste trabalho. Serão realizados experimentos com o intuito de avaliar o comportamento da estratégia em diferentes cenários de duplicação de dados. O processo de Resolução de Entidades será realizado utilizando os atributos apontados como mais relevantes pela nossa estratégia, afim de avaliar a qualidade do resultado obtido. Por fim, apresentaremos os resultados obtidos, e discussões acerca dos mesmos.

Tabela 4.6: Resumo das principais características das Estratégias de Seleção de atributos.

Trabalhos	**Objetivos**	**Grandes Volumes de Dados**	**Base/Conjunto de Dados**	**Aprendizagem de Máquina**	**Método de avaliação de Relevância dos atributos**
[Chen et al.]	Método para Resolução de Entidades, que seleciona o grupo de atributos mais relevante, a função de similaridade e o limiar adequados.	Não	Base	Supervisionado	Resultado do processo de Resolução de Entidades - F-measure
[Su et al.]	Técnica pra ajustar os pesos dos atributos para o cálculo de similaridade entre duas instâncias.	Sim	Conjunto de Dados	Não supervisionado	Vetores de dados duplicados e não duplicados - similaridade dos valores dos atributos
[Fan et al.]	Selecionar atributos para criar regras a serem utilizadas na comparação utilizando o conceito de dependências.	Não	Base	-	Semântica dos dados / Conceito de dependências
Este trabalho	**Uma estratégia de seleção de atributos que tem como objetivo principal classificar os atributos de acordo com sua relevância para o processo de Resolução de Entidades.**	**Sim**	**Conjunto de Dados**	**-**	**Critérios de avaliação relacionados aos dados e às fontes de dados.**

5

Implementação e Experimentos

Neste capítulo, são apresentados aspectos de implementação relacionados à estratégia de seleção de atributos proposta neste trabalho, e os experimentos realizados afim de avaliar a estratégia proposta. O capítulo está estruturado da seguinte maneira: A Seção 5.1 descreve a arquitetura do protótipo desenvolvido com o objetivo de auxiliar na validação da estratégia de seleção de atributos. Na Seção 5.2 são descritas as funcionalidades do protótipo. A Seção 5.3 apresenta a avaliação experimental realizada, e os resultados obtidos, além de uma discussão sobre os mesmos. E por fim, na Seção 5.4 apresentamos as conclusões.

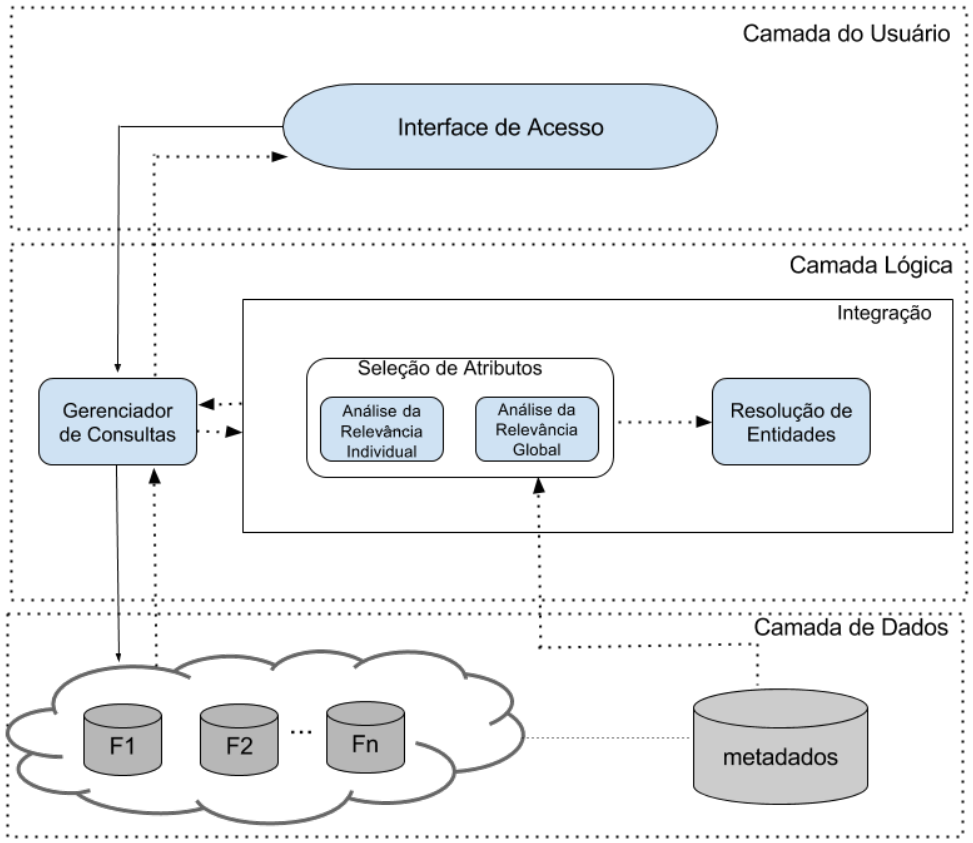
5.1 Arquitetura do Protótipo

Com o objetivo de avaliar a estratégia de seleção de atributos proposta neste trabalho, foi elaborado um protótipo, dividido em três camadas, como apresentado na Figura 5.1. A camada do usuário, é utilizada para o usuário interagir com o sistema, a camada lógica é responsável pelas funcionalidades do protótipo, e a camada de dados gerencia as bases utilizadas no processo. A seguir explicaremos detalhadamente cada componente da arquitetura do protótipo.

- **Interface de Acesso:** Módulo responsável pela iteração do usuário com a camada lógica. Ao usuário é permitido acesso a algumas funcionalidades, como: (i) execução de consultas sobre as bases; (ii) visualização dos resultados obtidos; (iii) cálculo de relevância para seleção dos atributos; (iv) visualização da classificação dos atributos e dos grupos de atributos com os valores de relevância; (v) possibilidade de adicionar algum atributo que julgar relevante.

- **Gerenciador de Consultas:** Módulo que permite acesso às fontes de dados, no qual é possível realizar consultas sobre as mesmas e obter os resultados destas consultas. Este módulo interage com o módulo de integração, composto pelo processo de seleção de atributos e pelo processo de Resolução de Entidades. Na fase de integração, as instâncias resultantes da consulta são resolvidas, retornadas ao gerenciador de consultas e devolvidas ao usuário.

Figura 5.1: Visão geral da arquitetura do protótipo.



- **Módulo de Seleção de Atributos:** É responsável por executar os algoritmos de cálculo de relevância individual e global, definidos no Capítulo 4. Ele retorna a classificação ordenada dos atributos de acordo com a relevância calculada. Os primeiros atributos do *ranking* são os mais relevantes para serem utilizados na Resolução de Entidades.

- **Módulo de Resolução de Entidades:** Utilizando os atributos mais bem classificados pelo módulo de seleção de atributos, o processo de Resolução de Entidades é realizado. Ao fim do processo, a qualidade do resultado obtido por meio da Resolução de Entidades é dada utilizando métricas de avaliação de qualidade implementadas. Essas métricas utilizam as correspondências e não correspondências encontradas pelo algoritmo de Resolução, para avaliar a qualidade do processo de Resolução de Entidades.

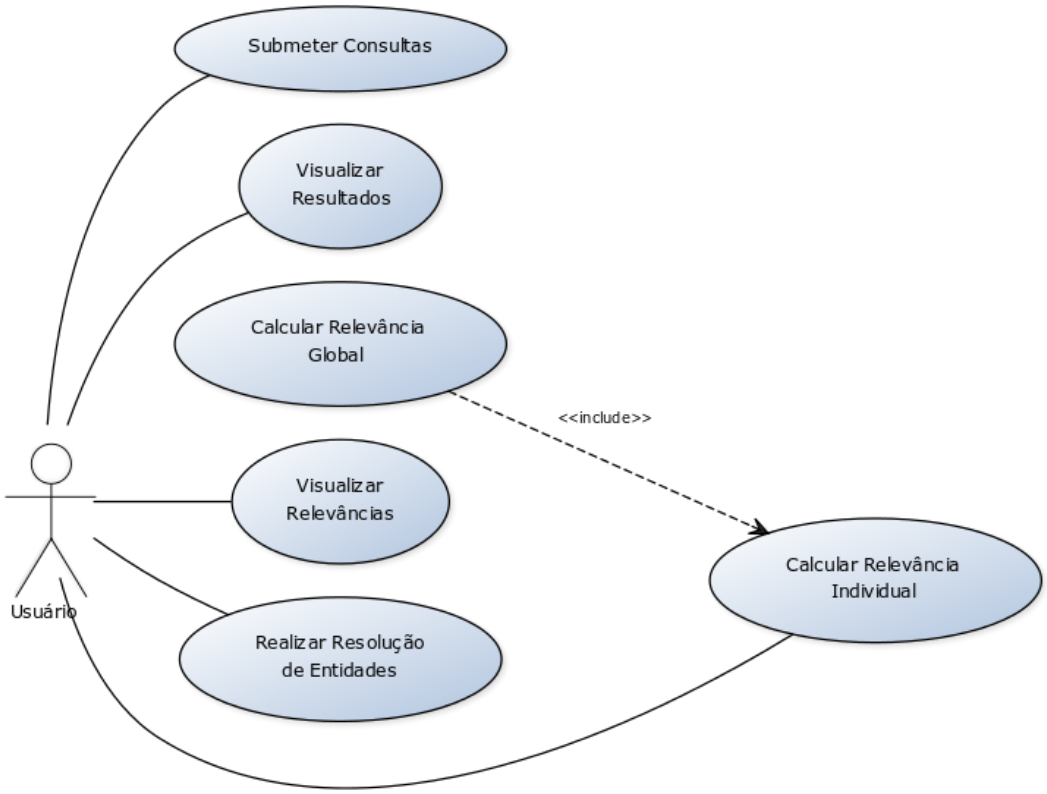
- **Fontes de Dados e Metadados:** Na camada de dados se encontram as fontes de dados que participam do processo. Essas fontes de dados possuem metadados de qualidade relacionados à elas.

5.2 Funcionalidades do Protótipo

O diagrama de casos de uso do protótipo, que corresponde às principais funcionalidades que foram desenvolvidas, pode ser visto na Figura 5.2.

A seguir, as tarefas que podem ser executadas por meio do protótipo são detalhadas.

Figura 5.2: Diagrama de Casos de Uso do Protótipo.



A seguir, as tarefas que podem ser realizadas pelo usuário por meio do protótipo são detalhadas.

- **Submeter Consultas:** o usuário pode submeter consultas sobre as fontes de dados.
- **Visualizar Resultados:** o usuário pode visualizar os resultados obtido por meio da consulta, contendo as instâncias resultantes e os atributos que as descrevem.
- **Calcular Relevância Individual dos Atributos:** Após o usuário visualizar o resultado obtido, ele pode calcular a relevância de cada atributo.
- **Calcular Relevância Global dos Atributos:** Após calcular a relevância individual, o usuário pode utilizar o metadado de confiabilidade das fontes para calcular a relevância global dos atributos.
- **Visualizar Relevâncias:** o usuário pode visualizar os valores de relevância individual e global de cada atributo. Visualizando o *ranking* ordenado dos atributos de acordo

com as relevâncias calculadas, o usuário pode selecionar os atributos para realizar o processo de Resolução de Entidades e, caso julgar necessário, pode selecionar algum atributo que não ficou bem classificado no *ranking*.

- **Realizar Resolução de Entidades:** Por fim, o usuário pode realizar a Resolução de Entidades utilizando o conjunto de atributos selecionados, e analisar a qualidade do processo por meio das métricas de qualidade implementadas, como *recall*, *precision* e *F-measure*.

Como pode ser visto na Figura 5.2, no protótipo elaborado para validar nossa estratégia deixamos como funcionalidades o cálculo da relevância individual e global dos atributos, e fica a cargo do usuário escolher se deseja utilizar apenas a relevância individual, ou se deseja utilizá-la juntamente com a relevância global. Após a seleção dos atributos, um *ranking* é retornado ao usuário, e também é de responsabilidade dele escolher quais atributos serão utilizados na Resolução de Entidades. Depois, o usuário pode realizar a Resolução de Entidades com os atributos escolhidos, e verificar por meio do resultado obtido, se os atributos selecionados realmente proporcionaram um bom resultado.

5.3 Avaliação Experimental

A avaliação experimental apresentada nesta seção, tem como objetivo avaliar a estratégia de seleção de atributos proposta neste trabalho. Para este fim, primeiramente, apresentamos o cenário de avaliação. Em seguida, é apresentada uma discussão a respeito do critério de avaliação utilizado para validar os grupos de atributos elencados pela nossa estratégia. Por fim, são mostrados os resultados obtidos, juntamente com uma discussão acerca dos mesmos.

5.3.1 Cenário

O cenário escolhido para realização desta avaliação experimental foi o domínio de referências bibliográficas de Ciência da Computação. A base de dados Cora foi selecionada, por ser uma base de dados muito utilizada em estudos de Resolução de Entidades (BILENKO; MOONEY, 2003; DONG; HALEVY; MADHAVAN, 2005; SINGLA; DOMINGOS, 2006).

A Cora contém 1.879 instâncias de diferentes fontes de dados, descritas por 15 atributos: *id*, *author*, *title*, *journal*, *volume*, *pages*, *year*, *publisher*, *address*, *note*, *venue*, *editor*, *type*, *institution* e *month*. Ainda é disponibilizado o atributo *class*, que é um identificador de instâncias duplicadas, onde todas as instâncias correspondentes contêm o mesmo valor para o *class*. Devido a ser um atributo ótimo, considerado um atributo espelho, ele foi desconsiderado neste trabalho.

Por ser uma base que contém muitos erros nos valores dos dados (também chamados de *outliers*), houve a necessidade de se realizar um pré-processamento, afim de remover essas discrepâncias. Por exemplo, o atributo *year* possuia caracteres não condizentes com o formato atribuído à ele (ex: "(1989)"ao invés de "1989").

A Cora possui também uma alta porcentagem de dados duplicados (±90%). Intuitivamente, nós acreditamos que a estratégia proposta poderia não ser eficiente em cenários com grande quantidade de duplicação nos dados. Devido a isso, foram criados conjuntos de dados com diferentes porcentagens de duplicação, com o intuito de avaliar como a estratégia proposta se comporta em cada cenário. Os cenários criados estão descritos na Tabela 5.1.

Tabela 5.1: Cenários com as porcentagens de dados duplicados.

Cenários	Porcentagem de Dados Duplicados
Cenário 1	5% - 10%
Cenário 2	15% - 30%
Cenário 3	35% - 50%
Cenário 4	55% - 70%
Cenário 5	>75%

5.3.2 Critério de Avaliação

Para avaliar a eficiência da estratégia de seleção de atributos proposta neste trabalho, foi necessário realizar a Resolução de Entidades utilizando os conjuntos de atributos ranqueados pela nossa estratégia, afim de validar o resultado obtido pela seleção de atributos. Para isso, o processo de Resolução de Entidades é realizado utilizando conjuntos de atributos formados de acordo com a classificação gerada pela estratégia de seleção de atributos. Por fim, é realizada uma avaliação de qualidade no resultado do processo, para validar se os atributos mais bem classificados no *ranking* são realmente os que fornecem o melhor resultado ao processo de Resolução de Entidades.

A base de dados Cora tem disponível o seu *gold standard* (G) que contém os pares de instâncias correspondentes combinados por meio de seus identificadores. O algoritmo de Resolução de Entidades compara os pares de instâncias e os classifica em CR (pares de instâncias consideradas correspondentes pelo algoritmo) e NCR (pares de instâncias considerados não correspondentes). Dada a classificação de CR e NCR, e considerando G, para cada comparação uma atribuição deve ser feita em uma das quatro categorias ilustradas na matriz de confusão da Figura 5.3.

Na Figura 5.3, os pares de instâncias de $CR \subset G$, são chamados de verdadeiros positivos (**TP**), os pares de *CR* que não estão contidos em *G*, são chamados de falsos positivos (**FP**), os pares de instâncias de $NCR \subset G$, são chamados de falsos negativos (**FN**), e os pares de instâncias de *NCR* que não estão contidas em *G*, são chamados de verdadeiros negativos (**TN**).

Com base no número de TP, TN, FP e FN identificados por meio da etapa de Resolução de Entidades, diferentes medidas de qualidade podem ser calculadas (CHRISTEN, 2012). Neste trabalho, a qualidade do resultado da etapa de Resolução de Entidades é dada por meio da métrica de *F-measure*, recomendada como melhor métrica para avaliar a qualidade do processo de Resolução de Entidades (CHRISTEN; GOISER, 2007), e explicada anteriormente no Capítulo

Figura 5.3: Matriz de Confusão dos tipos de correspondências encontradas na Resolução de Entidades.

	Classificação	
	Correspondente (⊂CR)	**Não Correspondente (⊂NCR)**
Correspondente (⊂G)	Correspondência Verdadeira Verdadeiro Positivo (TP)	Não Correspondência Falsa Falso Negativo (FN)
Não Correspondente (⊄G)	Correspondência Falsa Falso Positivo (FP)	Não Correspondência Verdadeira Verdadeiro Negativo (TN)

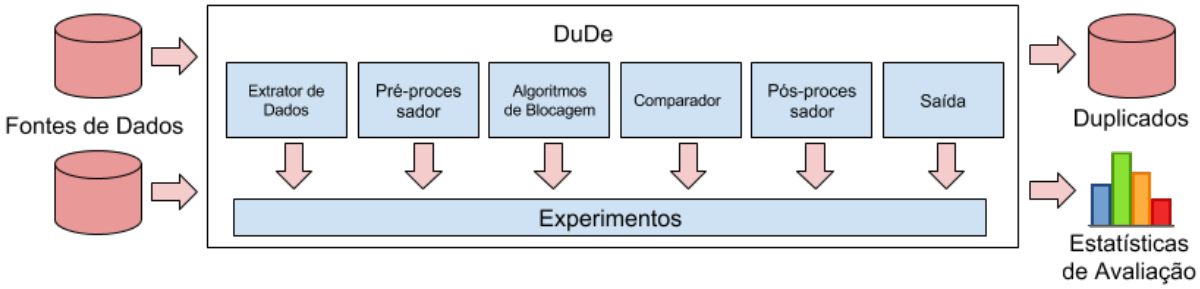
Fonte: Christen e Goiser (2007)

2. O valor de *F-measure* obtido com base no resultado da Resolução de Entidades é utilizado para validar a classificação dos atributos feita pela estratégia de seleção de atributos proposta neste trabalho.

5.3.3 Ferramenta para Resolução de Entidades

Para realizar a Resolução de Entidades em nosso experimento, escolhemos utilizar a ferramenta DuDe - *Duplicate Detection* [1]. O DuDe é um kit de ferramentas implementado na linguagem Java e desenvolvido na Universidade de Potsdam (UP) (DRAISBACH; NAUMANN, 2010), e livre, sob licença *General public license (GNU)*. A estrutura modular do DuDe permite que os usuários desenvolvam seu próprio código para substituir ou estender funcionalidades fornecidas pela ferramenta.

Figura 5.4: Arquitetura do DuDe.



Adaptado de: Draisbach e Naumann (2010)

O DuDe contém vários módulos, como pode ser visto na Figura 5.4 cada um com uma interface bem definida. O módulo **Extrator de Dados** fornece métodos para acessar arquivos *Comma-separated values (CSV)*, documentos *XML*, arquivos *JavaScript Object Notation (JSON)*, banco de dados relacionais e bibliografias no formato Bibtex. O módulo **Pré-processador** é opcional, sendo utilizado para colher dados estatísticos sobre o conjunto de dados de entrada. O módulo **Algoritimos de Blocagem** possui uma série de técnicas de blocagem que geram pares de instâncias candidatas do conjunto de dados de entrada. O módulo **Comparador** possui várias

[1] http://hpi.de/naumann/projects/data-quality-and-cleansing/dude-duplicate-detection.html#c115302

funções de cálculo de similaridade, e é onde os pares de instâncias candidatas são comparadas. O módulo **Pó-processador** recebe os pares de instâncias com os valores de similaridade calculados, e permite o cálculo de dados estatísticos como: número de pares duplicados, tempo de processamento, número de verdadeiros positivos, número de falsos positivos, entre outros. Por último, o módulo **Saída** fornece funções para escrever o resultado da Resolução de Entidades em diferentes formatos de saída, incluindo arquivos CSV e JSON.

O DuDe não possui uma interface gráfica, mas possui uma extensa documentação onde se encontram os primeiros passos para desenvolver um programa em Java utilizando as funcionalidades contidas nos módulos do *toolkit*. Além disso, na documentação se encontram programas exemplo, facilitando o entendimento do usuário.

Para nosso experimento, o algoritmo de Resolução de Entidades utilizado foi o *Naive Duplicate Detection*, com a função de Levenshtein para o cálculo da similaridade na comparação de instâncias.

5.3.4 Experimentos e Resultados

A avaliação experimental foi desenvolvida em um MacBook Air com Sistema Operacional OS X Yosemite versão 10.10.5, processador Intel Core i5 – 1,4 GHz, 4 GB de memória. Além disso, utilizamos MySQL 5.6 para armazenar os dados relacionados às bases.

Com os experimentos realizados, buscamos validar as seguintes hipóteses:

H1 - Considerar todos os atributos na fase de comparação da etapa de Resolução de Entidades ocasiona em um resultado contendo um alto número de correspondências erradas (Falsos positivos e Falsos Negativos), e um baixo número de correspondências corretas (Verdadeiros positivos e Verdadeiros negativos).

H2 - Considerar os atributos mais relevantes de acordo com a classificação realizada pela estratégia de seleção de atributos proposta, faz com que o resultado obtido contenha um maior número de correspondências corretas, com um menor número de correspondências erradas.

H3 - À medida que atributos menos relevantes são adicionados ao grupo de atributos selecionados para a fase de comparação de instâncias da etapa de Resolução de Entidades, o número de correspondências erradas aumenta, diminuindo o *F-measure* do resultado.

Para o cálculo de relevância individual dos atributos, é necessário definir o peso para os critérios de avaliação. Para isso, realizamos vários experimentos utilizando diferentes pesos para os critérios, e a partir desses experimentos, chegamos a conclusão de que o peso de 0,5 para cada critério era o mais adequado.

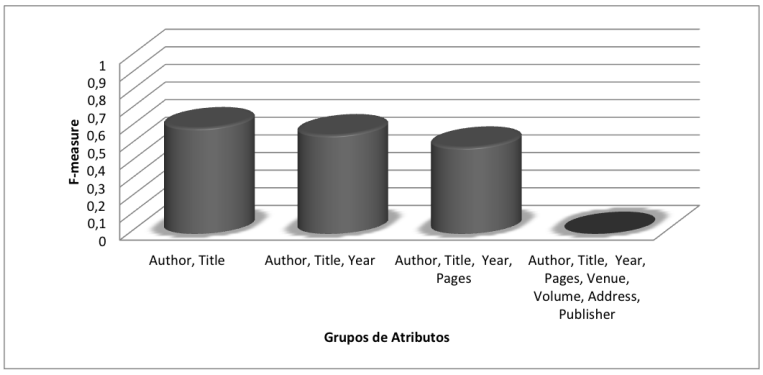
É importante destacar que, na etapa de Resolução de Entidades é necessário utilizar múltiplos atributos para comparar os pares de instâncias, ao invés de utilizar um único atributo (ou primeiro atributo do *ranking*) (CHEN et al., 2012). Por este motivo, foram elencados grupos de atributos a serem considerados na fase de comparação da etapa de Resolução de Entidades, de acordo com a ordenação do *ranking*. O **grupo 1** contendo os dois atributos mais relevantes, o

grupo 2 contendo os três atributos mais relevantes, o **grupo 3** contendo os quatro atributos mais relevantes, e por fim o **grupo 4**, contendo os 8 atributos mais relevantes, para representar todos os atributos que descrevem as instâncias, ou uma grande quantidade de atributos. Com os grupos elencados, realizamos o processo de Resolução de Entidades. Os atributos são identificados por **A**, **D** são os valores obtidos para o critério de Densidade, **R** são os valores obtidos para o critério de Repetição, e **RI** são os valores de relevância individual.

Primeiramente, realizamos a seleção de atributos para a base toda, e depois, para cada cenário criado, ranqueando os atributos de forma descendente. Na Figura 5.5, podemos visualizar os resultados do experimento utilizando a base completa. A Figura 5.5(a) mostra a classificação dos atributos de acordo com a relevância individual. O gráfico da Figura 5.5(b) mostra o *F-measure* do resultado do processo de Resolução de Entidades obtido para cada grupo de atributos elencado.

A	D	R	RI
Author	1,00	0,87	0,56
Title	0,98	0,90	0,54
Year	0,93	0,97	0,48
Pages	0,66	0,91	0,37
Venue	0,48	0,94	0,27
Volume	0,45	0,95	0,25
Address	0,40	0,95	0,22
Publisher	0,38	0,98	0,20
Journal	0,34	0,98	0,18
Editor	0,24	0,95	0,15
Type	0,14	0,98	0,08
Institution	0,13	0,98	0,08
Note	0,04	0,98	0,03
Month	0,02	0,99	0,01

(a) Ranking de atributos



(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.5: Experimento 1 realizado com a base Cora completa.

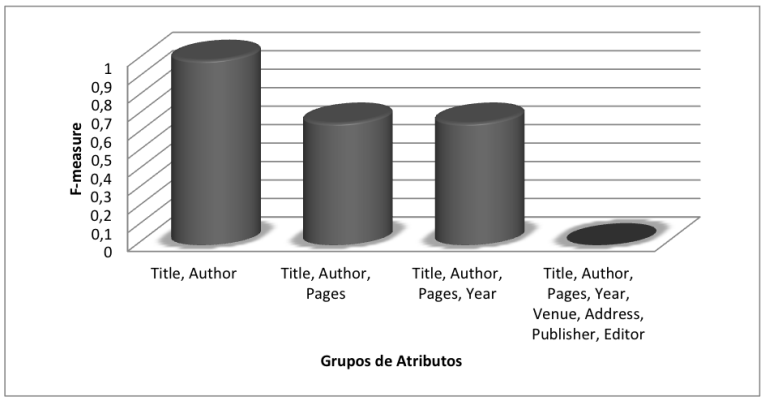
Como podemos verificar, o grupo mais relevante para a Resolução de Entidades, neste caso, é composto por *title*, com uma relevância individual de 0,56, e *author*, com uma relevância individual de 0,54. O grupo contendo *title* e *author*, proporcionou um *F-measure* de 0,6 ao processo de Resolução de Entidades. O resultado obtido contém um grande número de correspondências erradas, podendo não ser um resultado satisfatório. O atributo mais bem classificado pela estratégia, teve uma relevância de 0,56%. O restante dos atributos tiveram valores de relevância abaixo disso. Isso ocorre devido à base ter uma alta porcentagem de dados duplicados.

Podemos concluir com esse primeiro experimento, que realizar a Resolução de Entidades para a base toda não é a melhor opção, já que a qualidade do resultado do processo é baixa. Consequentemente, selecionar os atributos para a base toda, também não é vantajoso, uma vez que o grupo de atributos selecionado como mais relevante pode variar de acordo com o conjunto de dados analisado.

No cenário 1, podemos visualizar a classificação dos atributos por meio da Figura 5.6(a), onde pode-se verificar que os atributos quatro mais bem classificados foram *title*, *author*,

A	D	R	RI
Title	1,00	0,05	0,98
Author	1,00	0,08	0,96
Pages	0,68	0,30	0,69
Year	0,85	0,55	0,65
Venue	0,63	0,48	0,58
Address	0,35	0,68	0,34
Publisher	0,45	0,78	0,34
Editor	0,30	0,75	0,28
Volume	0,23	0,75	0,24
Journal	0,15	0,83	0,16
Institution	0,15	0,85	0,15
Type	0,13	0,88	0,13
Note	0,10	0,88	0,11
Month	0,00	1,00	0,00

(a) Ranking de atributos.



(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.6: Resultados Experimento 1 - Cenário 1.

pages e *year*. O gráfico da Figura 5.6(b) mostra o resultado do processo de Resolução de Entidades utilizando os grupos de atributos. Os dois primeiros atributos do *ranking*, *title* e *author*, obtiveram um valor de relevância de 0,98 e 0,96 respectivamente. Dessa forma, o grupo de atributos selecionado como mais relevante, contendo os dois primeiros atributos do *ranking*, proporcionou ao processo de Resolução de Entidades um *F-measure* de 1, ou seja, o resultado ideal. O terceiro atributo do *ranking* (*pages*), obteve uma relevância de 0,68, e adicionado ao grupo de atributos considerados no processo de Resolução de Entidades, o *F-measure* obtido foi de 0,66. Em seguida, o atributo *year* teve 0,65 de relevância individual, e o grupo de quatro atributos, incluindo o *year*, proporcionou um *F-measure* de 0,66, igualmente ao anterior. O último grupo, contendo 8 atributos, proporcionou um *F-measure* de 0.

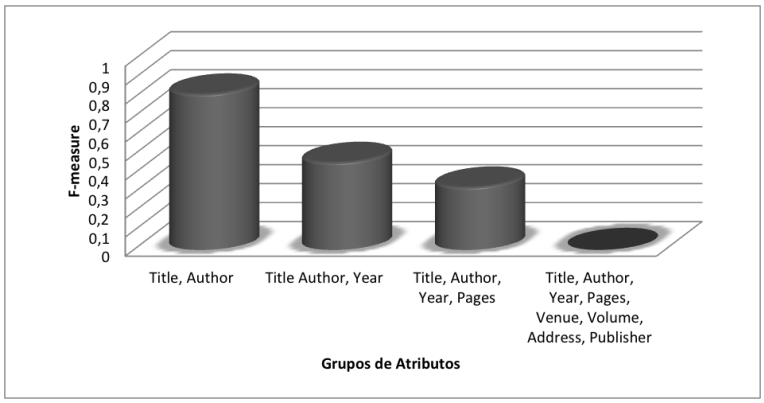
No cenário 2, conforme mostra a Figura 5.7(a) o grupo de atributos selecionado como mais relevante foi *title* e *author*, com valores de relevância de 0,90 e 0,79. O atributo *year* teve uma relevância de 0,65, e *pages* 0,58. Realizando o processo de Resolução de Entidades com os dois atributos mais bem classificados no *ranking*, foi obtido um *F-measure* de 0,82. Adicionando o atributo *year*, o *F-measure* foi de 0,46, e utilizando também o atributo *pages*, o *F-measure* caiu para 0,33. O último grupo, contendo 8 atributos, proporcionou um *F-measure* de 0.

No cenário 3, o maior *F-measure* obtido foi de 0,45. Mesmo com qualquer outra combinação de atributos, o *F-measure* não ultrapassou esse valor. Como pode ser visto na Figura 5.8(a), os atributos melhores classificados pela estratégia de seleção de atributos foram *title*, *author*, *year*, e *venue*, nesta ordem. O grupo composto por *title* e *author*, proporcionou um *F-measure* de 0,45 ao processo de Resolução de Entidades. Ao adicionar o atributo *year*, o *F-measure* caiu para 0,40, e considerando também o atributo *venue*, caiu para 0,36. O último grupo, proporcionou um *F-measure* de 0. No entanto, mesmo não sendo um alto *F-measure*, a estratégia selecionou os atributos corretamente, já que o grupo de atributos melhores classificados foi o que proporcionou o *F-measure* mais alto.

No cenário 4, a ordem de classificação foi *author*, *title*, *year*, e *pages*. O grupo com

A	D	R	RI
Title	0,97	0,17	0,90
Author	0,77	0,20	0,79
Year	0,90	0,60	0,65
Pages	0,60	0,43	0,58
Venue	0,47	0,60	0,43
Volume	0,43	0,63	0,40
Address	0,3	0,67	0,32
Publisher	0,33	0,77	0,28
Journal	0,33	0,83	0,25
Editor	0,20	0,77	0,22
Institution	0,20	0,80	0,20
Note	0,13	0,83	0,15
Type	0,17	0,87	0,15
Month	0,00	1,00	0,00

(a) Ranking de atributos.

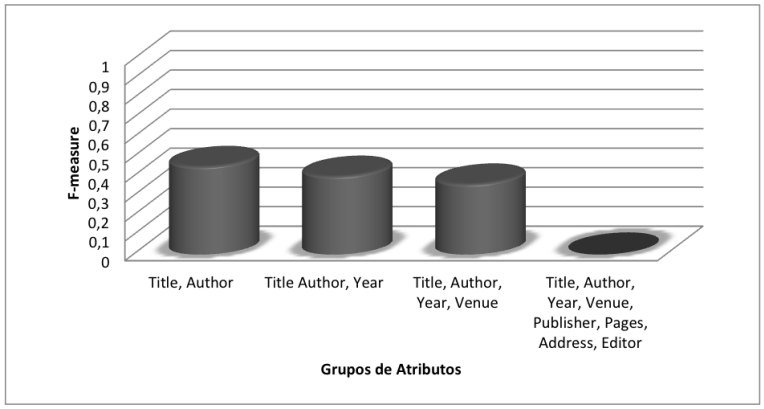


(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.7: Resultados Experimento 1 - Cenário 2.

A	D	R	RI
Title	1,00	0,34	0,83
Author	0,98	0,44	0,77
Year	0,90	0,62	0,64
Venue	0,68	0,60	0,54
Publisher	0,66	0,70	0,48
Pages	0,46	0,62	0,42
Address	0,46	0,74	0,36
Editor	0,38	0,76	0,31
Volume	0,34	0,78	0,28
Journal	0,20	0,80	0,20
Note	0,06	0,94	0,06
Institution	0,04	0,94	0,05
Type	0,04	0,94	0,05
Month	0	1	0

(a) Ranking de atributos.



(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

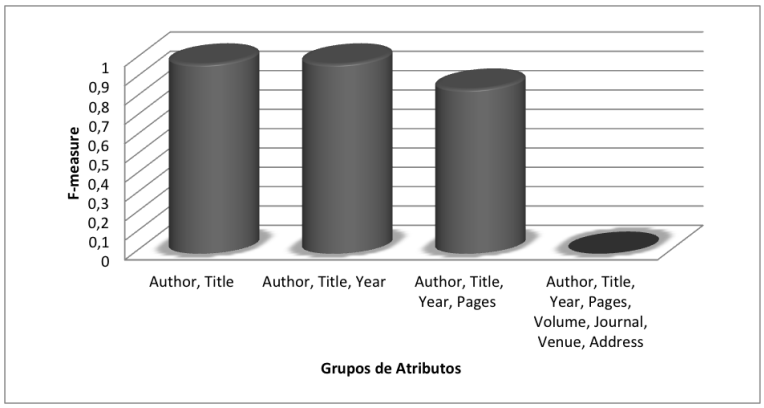
Figura 5.8: Resultados Experimento 1 - Cenário 3.

os dois primeiros atributos do *ranking*, proporcionou um *F-measure* de 0,98 ao processo de Resolução de Entidades. O segundo grupo, contendo os três atributos mais relevantes, também proporcionou um *F-measure* de 0,98. Ao adicionar o atributo *pages*, o *F-measure* caiu para 0,85. E o grupo contendo 8 atributos, proporcionou um *F-measure* de 0. A classificação dos atributos no cenário 4 pode ser vista na Figura 5.9(a), e na Figura 5.9(b) é mostrado o gráfico com os resultados do processo de Resolução de Entidades considerando os grupos de atributos do *ranking*.

Na Figura 5.10, o experimento do cenário 5 é apresentado. Os atributos classificados de acordo com sua relevância individual são mostrados na Figura 5.10(a), onde os atributos mais bem classificados no *ranking* foram *author*, *title*, *year* e *venue*. O resultado do processo de Resolução de Entidades com os dois primeiros atributos do *ranking* obteve um *F-measure* de 1, caso ideal. Com os três primeiros atributos do *ranking*, o *F-measure* se manteve 1. Ao adicionar o atributo *venue*, quarto classificado no *ranking*, o *F-measure* do processo de Resolução de Entidades caiu para 0,66, e com os 8 primeiros atributos do *ranking*, o *F-measure* foi 0.

A	D	R	RI
Author	1,00	0,65	0,68
Title	0,90	0,63	0,64
Year	1,00	0,78	0,61
Pages	0,68	0,75	0,46
Volume	0,48	0,75	0,36
Journal	0,55	0,85	0,35
Venue	0,37	0,80	0,29
Address	0,35	0,80	0,28
Publisher	0,33	0,88	0,23
Editor	0,23	0,85	0,19
Note	0,10	0,88	0,11
Type	0,08	0,93	0,08
Institution	0,03	0,95	0,04
Month	0,00	1,00	0,00

(a) Ranking de atributos.

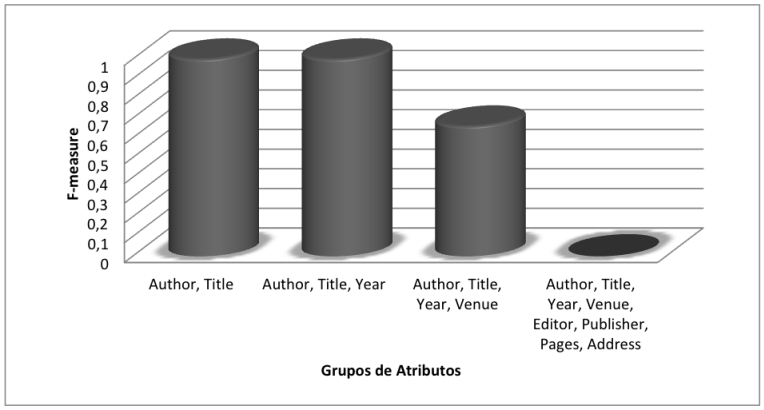


(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.9: Resultados Experimento 1 - Cenário 4.

A	D	R	RI
Author	1,00	0,74	0,63
Title	1,00	0,76	0,62
Year	0,84	0,80	0,52
Venue	0,85	0,82	0,52
Editor	0,54	0,78	0,38
Publisher	0,62	0,88	0,37
Pages	0,56	0,86	0,35
Address	0,44	0,82	0,31
Institution	0,10	0,88	0,11
Type	0,08	0,92	0,08
Volume	0,04	0,96	0,04
Journal	0,04	0,96	0,04
Note	0,02	0,96	0,03
Month	0,00	1,00	0,00

(a) Ranking de atributos.



(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.10: Resultados Experimento 1 - Cenário 4.

Para avaliar como a análise da relevância global se comporta, utilizamos o cenário 1, que possui um conjunto de dados com 40 instâncias de publicações. No entanto, como todas as instâncias são oriundas da mesma fonte de dados Cora - que mesmo contendo instâncias de múltiplas fontes de dados, não disponibiliza a origem individual de cada instância - precisamos supor duas fontes de dados (F1 e F2), para realizar esse experimento.

Logo, supondo saber a proveniência dessas instâncias, em que 25 instâncias são oriundas da fonte de dados 1 (F1), que possui uma confiabilidade de 70%, e 15 instâncias são oriundas da fonte de dados 2 (F2), que possui uma confiabilidade de 90%, para cada f_k, calculamos a cobertura com relação ao conjunto de dados, conforme a Equação 4.8, tendo como resultado $Cob(F1) = 0,625, Cob(F2) = 0,375$. Depois, calculamos a qualidade do conjunto de fontes de dados F, composto por $F1$ e $F2$, conforme a Equação 4.10, tendo $Q(F) = 0,76$. Com isso, para cada A_i, o valor de relevância individual é ponderado utilizando o valor de $Q(F)$, e os resultados podem ser vistos na Tabela 5.2.

Um segundo experimento foi realizado, utilizando uma base de dados sintética com dados

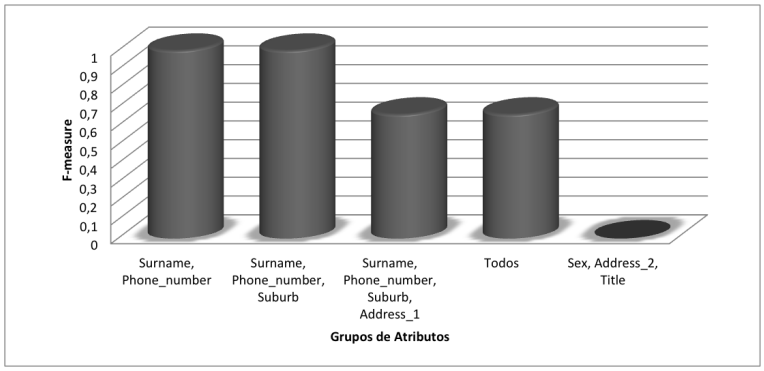
Tabela 5.2: Resultado da análise de relevância global dos atributos - Cenário 1.

A	RI	RG
Title	0,98	0,74
Author	0,96	0,72
Pages	0,69	0,52
Year	0,65	0,50
Venue	0,58	0,44
Address	0,34	0,25
Publisher	0,34	0,25
Editor	0,28	0,21
Volume	0,24	0,18
Journal	0,16	0,12
Institution	0,15	0,11
Type	0,13	0,09
Note	0,11	0,08
Month	0,00	0,00

pessoais, gerada pelo FEBRL[2]. A base de dados contém 130 mil instâncias, e uma porcentagem de mais ou menos 23% de dados duplicados. Particionamos as bases em cenários conforme mostrados na Tabela 5.1. No entanto, devido a pequena porcentagem de duplicação, só foi possível obter os 4 primeiros cenários a partir dessa base, o último cenário, contendo mais que 75% de dados duplicados, não foi utilizado. Além disso, criamos mais um grupo de atributos, onde o **grupo 5** é composto pelos 3 atributos mais irrelevantes, de acordo com a classificação dada pela estratégia proposta.

A	D	B	RI
Surname	1,00	0,06	0,97
Phone_number	1,00	0,06	0,97
Suburb	1,00	0,09	0,96
Address_1	0,97	0,09	0,94
Postcode	0,97	0,13	0,92
Given_name	0,97	0,16	0,91
Street_number	0,97	0,19	0,89
Date_of_birth	0,84	0,16	0,84
Culture	0,97	0,50	0,74
Age	0,87	0,43	0,72
State	0,86	0,73	0,57
Sex	0,80	0,90	0,45
Address_2	0,40	0,66	0,37
Title	0,50	0,83	0,34

(a) Ranking de atributos.



(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.11: Resultados Experimento 2 - Cenário 1.

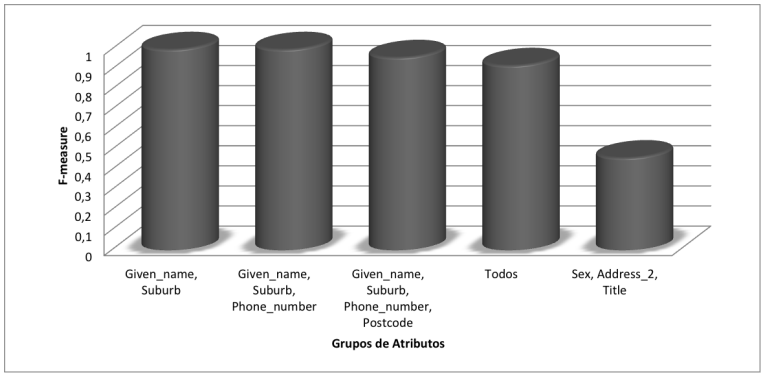
No cenário 1, podemos visualizar a classificação de relevância dos atributos na Figura 5.11(a). O grupo de atributos selecionado como mais relevante foi *Surname* e *Phone_number*, com valores de relevância de 0,97% para ambos. Em seguida, o atributo *Suburb*, teve 0,96% de relevância, e o atributo *Address_1* teve 0,94%. Realizando a Resolução de Entidades, o

[2]https://sourceforge.net/projects/febrl/

primeiro grupo, contendo os dois primeiros atributos do *ranking*, obteve um *F-measure* de 1, ou seja, resultado ideal para o processo. O segundo grupo também obteve um *F-measure* de 1. Adicionando o atributo *Address_1*, o *F-measure* caiu para 0,66. O grupo de todos os atributos também obteve um *F-measure* de 0,66, e o último grupo, contendo 3 atributos irrelevantes classificados pela estratégia, obteve um *F-measure* de 0.

A	D	R	RI
Given_name	1,00	0,18	0,91
Suburb	1,00	0,18	0,91
Phone_number	1,00	0,18	0,91
Postcode	1,00	0,20	0,90
Date_of_birth	0,90	0,25	0,83
Address_1	0,88	0,23	0,83
Surname	0,88	0,25	0,81
Street_number	1,00	0,38	0,81
Culture	0,95	0,70	0,63
State	0,98	0,79	0,59
Age	0,73	0,63	0,55
Sex	0,78	0,93	0,43
Address_2	0,38	0,70	0,34
Title	0,48	0,85	0,31

(a) Ranking de atributos.



(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.12: Resultados Experimento 2 - Cenário 2.

No cenário 2, a Figura 5.12(a) mostra a classificação dos atributos por ordem de relevância, dada pela nossa estratégi, e a Figura 5.12(b), mostra os resultados obtidos no processo de Resolução de Entidades, utilizando os grupos de atributos elencados. O primeiro grupo, contendo os atributos *Given_name* e *Suburb*, com relevâncias de 0,91% para ambos, obteve um *F-measure* de 1, alcançando novamente o resultado ideal. O grupo contendo os três atributos mais relevântes, com *Phone_number* que foi classificado com uma relevância de 0,91, também atingiu o *F-measure* de 1. O terceiro grupo, adicionando o atributo *Postcode* com uma relevância de 0,90, obteve um *F-measure* de 0,96. O grupo de todos os atributos obteve um *F-measure* de 0,92. O grupo de atributos irrelevantes, segundo a classificação dada pela estratégia, obteve um *F-measure* de 0,46.

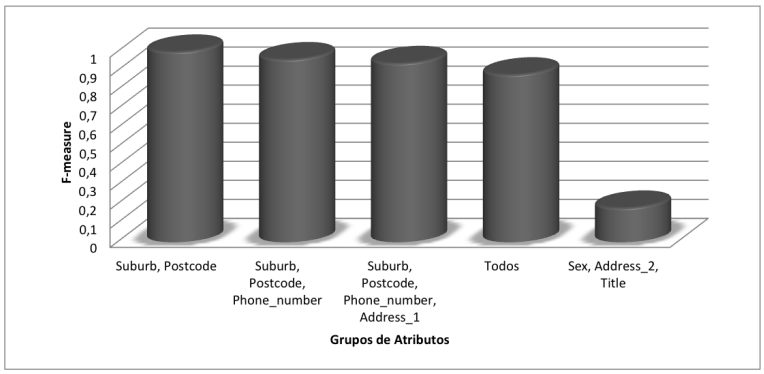
No cenário 3, os atributos melhores classificados pela estratégia proposta foram *Suburb*, *Postcode*, com relevância de 0,81% para ambos, *Phone_number* com relevância de 0,79, e *Address_1* com 0,77. O primeiro grupo obteve um *F-measure* de 1, resultado ideal. O segundo grupo obteve um *F-measure* de 0,96. O terceiro grupo, obteve um *F-measure* de 0,94, e o grupo de todos os atributos obteve um *F-measure* de 0,88. O grupo de atributos irrelevantes, obteve um *F-measure* de 0,18.

No cenário 4, os atributos mais bem classificados pela nossa estratégia foram *Surname*, com 0,72% de relevância, *Given_name*, com o,71%, *Postcode*, com 0,71% e *Suburb*, com 0,70%. O grupo 1 e o grupo 2 obteviveram um *F-measure* de 0,97, muito próximo do resultado ideal. O grupo 3, obteve um *F-measure* de 0,90. O grupo contendo todos os atributos, obteve um *F-measure* de 0,80, e o grupo contendo os atributos irrelevantes, obteve um *F-measure* de 0,23.

Como podemos verificar, os resultados obtidos com o segundo experimento foram

A	D	R	RI
Suburb	1,00	0,39	0,81
Postcode	1,00	0,39	0,81
Phone_number	0,96	0,39	0,79
Address_1	0,96	0,41	0,77
Street_number	1,00	0,48	0,76
Surname	0,95	0,44	0,75
Given_name	0,97	0,52	0,73
Date_of_birth	0,88	0,44	0,72
Culture	1,00	0,85	0,57
State	0,95	0,85	0,55
Age	0,73	0,71	0,51
Sex	0,79	0,96	0,41
Address_2	0,34	0,79	0,28
Title	0,44	0,89	0,27

(a) Ranking de atributos.

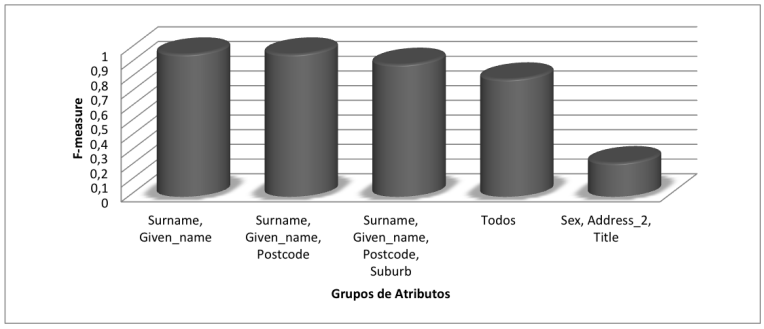


(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.13: Resultados Experimento 2 - Cenário 3.

A	D	R	RI
Surname	1,00	0,56	0,72
Given_name	1,00	0,58	0,71
Postcode	1,00	0,58	0,71
Suburb	1,00	0,60	0,70
Street_number	1,00	0,60	0,70
Address_1	1,00	0,60	0,70
Date_of_birth	0,94	0,60	0,67
Phone_number	0,94	0,60	0,67
Culture	0,94	0,78	0,58
State	0,92	0,84	0,54
Age	0,72	0,80	0,46
Sex	0,76	0,94	0,41
Address_2	0,52	0,78	0,37
Title	0,50	0,90	0,30

(a) Ranking de atributos.



(b) F-measure do processo de Resolução de Entidades com os grupos de atributos.

Figura 5.14: Resultados Experimento 2 - Cenário 4.

parecidos com os resultados obtidos no primeiro. A estratégia elencou corretamente os atributos, onde os primeiros do *ranking* foram os que proporcionaram os melhores resultados ao processo de Resolução de Entidades.

Um aspecto interessante a ser notado nesse segundo experimento, é que o grupo de atributos relevantes variou bastante dependendo nos cenários de dados avaliados. Isso comprova que, de acordo com o conjunto de dados avaliado, os atributos relevantes podem mudar.

Devido a base não conter muitos valores nulos e repetidos, podemos verificar que grande parte dos atributos teve uma classificação com alta relevância, e por isso, o grupo contendo todos os atributos não obteve um resultado com muitas correspondências erradas. Deste modo, criamos o grupo contendo apenas os atributos menos relevantes para mostrar que, se atributos irrelevantes são utilizados na comparação, o resultado do processo de Resolução de Entidades contém um alto número de correspondências erradas e um baixo número de correspondências corretas.

5.3.5 Discussões

Realizando uma análise dos resultados obtidos podemos concluir que a nossa estratégia de seleção de atributos se mostrou eficiente em todos os cenários, em ambos experimentos, e

as hipóteses elencadas foram validadas. Nos cenários com grande porcentagem de duplicação (Experimento 1, cenário 4 e cenário 5; Experimento 2, cenário 4), é possível observar que o valor de relevância dos atributos não é tão alto, quando comparado com os demais cenários. Isso se deve ao fato de que nos cenários com grande porcentagem de duplicação o valor do critério de repetição é sempre muito alto. No entanto, nossa intuição inicial de que um alto percentual de dados duplicados influenciaria negativamente na eficiência da estratégia proposta não foi confirmada, uma vez que mesmo os valores de relevância dos atributos sendo baixos, nossa estratégia conseguiu elencar corretamente os melhores atributos em todos os cenários, inclusive os cenários com alto percentual de duplicação.

No primeiro Experimento, os dois atributos mais bem classificados pela nossa estratégia de seleção de atributos em todos os cenários foram *title* e *author*. O terceiro atributo mais bem classificado variou entre *pages*, e *year*, e o quarto classificado variou entre *venue*, *pages* e *year*.

No Experimento 1, em todos os cenários, podemos verificar que na maioria dos casos em que atributos irrelevantes foram adicionados ao grupo de atributos considerados na Resolução de Entidades, o *F-measure* do processo diminuiu. No cenário 1, ao adicionar *pages* ao grupo, o *F-measure* caiu de 1 para 0,66, um percentual de 34% de diminuição. Podemos ver também que em 4 cenários o resultado obtido no processo de Resolução de Entidades considerando o grupo de atributos mais relevantes proporcionou um *F-measure* muito alto. Tanto no cenário 1 como no cenário 5, o *F-measure* obtido com o grupo de atributos mais relevantes foi de 1, alcançando o resultado ideal. Nos cenários 2 e 4, o *F-measure* foi maior que 0,8, muito próximo do ideal. No cenário 3, o maior *F-measure* obtido foi de 0,45, não sendo um bom resultado. Acreditamos que isso ocorre devido à qualidade da amostra de dados utilizada para compor o cenário 3.

O grupo de 8 atributos do experimento 1, foi utilizado com o intuito de demonstrar que quando se considera muitos atributos irrelevantes para a comparação de instâncias do processo de Resolução de Entidades, o resultado tende a ser muito ruim. Com 8 atributos o *F-measure* do processo de Resolução de Entidades em todos os cenários testados foi igual a 0. A base Cora possui 14 atributos, deste modo, considerar todos os atributos não é uma boa opção. Portanto, confirmamos que utilizar uma grande quantidade de atributos no processo de Resolução de Entidades não é viável, como mostram os resultados obtidos utilizando o conjunto de atributos contendo 8 atributos. No entanto, no experimento 2 o grupo contendo todos os atributos, em alguns casos, proporcionou um resultado contendo poucas correspondências erradas. Isso se deve ao fato de a base conter poucos valos nulos e repetidos, e a maioria dos atributos foram classificados com uma relevância alta. No entando, o grupo criado contendo 3 atributos menos relevantes, nunca proporcionou um resultado bom ao processo, validando a hipótese de que atributos irrelevantes quando utilizados na comparação ocasionam em muitas correspondências erradas.

Foi possível observar também que o maior *F-measure* do processo de Resolução de Entidades foi obtido utilizando o grupo 1, contendo os dois atributos mais relevantes elencados pela nossa estratégia, tanto no experimento 1, como no experimento 2. No experimento 2, o

grupo 2, contendo os três atributos mais relevantes, também atingiu um *F-measure* de 1, nos cenários 1 e 2. Sendo assim, percebemos também que à medida que atributos com menor valor de relevância são considerados na comparação, o *F-measure* diminui. Assim sendo, os resultados obtidos por meio dos experimentos validaram nossas hipóteses.

Sobre a análise de relevância global, realizada no experimento 1, podemos concluir que se uma fonte contribui com a maioria das instâncias para um conjunto de dados, mas essa fonte tem uma confiabilidade baixa, provavelmente a relevância dos atributos possa ser contestada. Deste modo, para avaliar se o resultado da seleção de atributos é realmente confiável, utilizamos a relevância global. Com os resultados obtidos nos experimentos, acreditamos que a relevância global é útil no que se propõe a fazer.

5.4 Considerações

Neste capítulo discutimos a implementação da estratégia proposta e os experimentos realizados para sua validação. Foi apresentada a arquitetura do protótipo implementado para auxiliar na realização dos experimentos, bem como suas funcionalidades. Apresentamos os resultados obtidos por meio dos experimentos realizados e algumas discussões acerca dos mesmos.

Um experimento inicial foi realizado para classificar os atributos ordenadamente de acordo com sua relevância individual, levando em consideração a base Cora completa. Depois elencamos vários cenários contendo diferentes porcentagens de duplicação nos dados, com o intuito de avaliar o comportamento da estratégia proposta neste trabalho. Utilizando os atributos melhores classificados no *ranking*, agrupando-os conforme a ordenação resultante da seleção de atributos, a Resolução de Entidades foi realizada em cada cenário, considerando cada grupo de atributos elencado. Por meio do resultado da Resolução de Entidades, utilizando a métrica de *F-measure* para mensurar a qualidade do processo, validamos os atributos selecionados pela estratégia proposta.

Com esse experimento, podemos concluir que a estratégia de seleção de atributos proposta nesse trabalho foi eficiente em todos os cenários elencados. Os atributos foram ordenados corretamente, de acordo com sua relevância para o processo de Resolução de Entidades, ou seja, os primeiros atributos do *ranking* são realmente os que proporcionam o melhor resultado ao processo de Resolução de Entidades.

Depois, para avaliar a análise de relevância global dos atributos, utilizamos o cenário 1. Como a maioria das instâncias contidas no cenário 1 são provenientes da fonte de dados F1, que possui uma confiabilidade de 70%, os valores de relevância individual dos atributos diminuíram, já que esse valor de confiabilidade não é tão alto. Podemos constatar que a relevância global pode ser utilizada para determinar se o resultado do processo de seleção de atributos é realmente confiável. Por fim, fizemos uma discussão acerca dos resultados obtidos por meio da avaliação experimental realizada.

6

Conclusão

Neste trabalho foi proposta uma estratégia para seleção de atributos relevantes no processo de Resolução de Entidades. O objetivo desta estratégia é selecionar os melhores atributos para serem utilizados na fase de comparação entre pares de instâncias, afim de proporcionar o melhor resultado possível ao processo de Resolução de Entidades. A estratégia proposta se divide em duas etapas: Analise da Relevância Individual, e Análise da Relevância Global.

Na etapa de Análise da Relevância Individual, propomos dois critérios de avaliação de relevância, sendo eles Densidade e Repetição. Propomos também uma fórmula para o cálculo de Relevância Individual. Na etapa de Análise da Relevância Global, são considerados metadados de qualidade das fontes de dados envolvidas no processo. Neste trabalho utilizamos o metadado de confiabilidade, e propomos uma fórmula para o cálculo da Relevância Global dos atributos.

Uma visão geral sobre a estratégia de seleção de atributos proposta neste trabalho foi apresentada. Além disso, foram definidas as métricas e os algoritmos implementados neste trabalho. O primeiro com o objetivo de avaliar a Relevância Individual dos atributos, por meio dos critérios de avaliação da relevância, e o segundo, com o intuito de avaliar a Relevância Global dos atributos, com base nos metadados de confiabilidade das fontes, e considerando a cobertura de cada fonte no conjunto de dados. O resultado final dos algoritmos é um *ranking* ordenado dos atributos e o valor de relevância individual e global para cada um.

6.1 Principais Contribuições

Dentre os principais diferenciais desta abordagem, destaca-se o fato da relevância dos atributos ser calculada utilizando critérios de avaliação de relevância e metadados de qualidade das fontes. Desde modo, a relevância é calculada avaliando características dos dados e das fontes de dados. Sendo assim, a estratégia proposta não faz uso algoritmos de Aprendizagem de Máquina, sabendo dos pontos negativos da necessidade de um conjunto de treinamento.

Dentre as principais contribuições deste trabalho, destacam-se:

- **Definição dos critérios para avaliação de relevância dos atributos.** Foram definidos os critérios de Densidade e Repetição, com as medidas de avaliação para ambos.

Com base nesses critérios a relevância individual dos atributos é determinada.

- **Definição das métricas para cálculo da relevância individual e global dos atributos.** Propomos as métricas para o cálculo de relevância individual e relevância global dos atributos. Por meio dessas métricas são definidos os atributos a serem considerados no processo de Resolução de Entidades.

- **Desenvolvimento de um protótipo.** Com o objetivo de validar a estratégia proposta, os algoritmos de cálculo de relevância individual e global foram implementados em um protótipo.

É importante destacar que, a partir da avaliação experimental realizada com o intuito de validar nossa estratégia, podemos concluir que a estratégia se mostrou eficiente em todos os cenários. Os experimentos comprovaram que os grupos de atributos selecionados pela nossa estratégia são os que proporcionam o melhor resultado para a Resolução de Entidades, e as hipóteses elencadas foram validadas. Com isso, podemos concluir que o presente trabalho conseguiu atingir seus objetivos, disponibilizando uma solução eficiente para a seleção de atributos no processo de Resolução de Entidades.

Até a presente data, as seguintes publicações foram geradas a partir dos resultados obtidos durante o desenvolvimento desta dissertação:

- Canalle, G. K.; Lóscio, B. F.; Salgado, A. C. (2015) **Uma estratégia para Seleção de Atributos Relevantes no Processo de Resolução de Entidades.** *In Proceedings* of WTDBD 2015, Petrópolis - RJ, Brasil, 2015.

- Canalle, G. K.; Lóscio, B. F.; Salgado, A. C. (2016) **Uma estratégia para Seleção de Atributos Relevantes no Processo de Resolução de Entidades.** 31 Simpósio Brasileiro de Banco de Dados (SBBD), Salvador - BA, Brasil, 2016. (aceito para apresentação em Outubro/2016)

6.2 Trabalhos Futuros

A seguir, apresentamos algumas direções que podem ser exploradas em trabalhos futuros:

- **Analisar outros critérios para avaliação da relevância dos atributos**
  Nossa estratégia considera dois critérios para avaliar a relevância individual de um atributo: Densidade e Repetição. Como trabalho futuro é interessante analisar outros critérios a serem incluídos na avaliação da relevância individual dos atributos. Como exemplo de critérios a serem analisados, citamos a a tendência de um atributo a conter erros (ex: sobrenome), e a dinamicidade de um atributo, ou seja, um valor que pode sofrer muitas alterações ao longo do tempo (ex: valor do produto). Acreditamos

que tais características também podem ser úteis para ajudar a avaliar a relevância de um atributo para o processo de Resolução de Entidades, incrementando a estratégia proposta.

- **Avaliar a relevância de outros medadados de qualidade para análise de relevância global dos atributos**
  Nossa estratégia considera metadados de qualidade das fontes de dados para avaliar a relevância global de um atributo. Atualmente, apenas a confiabilidade das fontes de dados envolvidas no processo é utilizada. Como um trabalho futuro deve-se investigar outros metadados de qualidade que sejam possíveis de serem incluídos na análise de relevância global, e impactem positivamente no processo de seleção de atributos.

- **Realização de melhorias na estratégia de seleção de atributos proposta**
  A estratégia de seleção de atributos proposta, faz a seleção dos atributos para um conjunto de dados que está associado a apenas um conceito. Como trabalho futuro, estudos devem ser realizados para investigar maneiras de realizar a seleção de atributos para conjuntos de dados associado a mais de um conceito.

- **Avaliar a aplicabilidade da estratégia de seleção em outras fases do processo de Resolução de Entidades**
  O foco da estratégia proposta neste trabalho é selecionar os atributos relevantes para a fase de comparação de pares de instâncias do processo de Resolução de Entidades. Como trabalho futuro, é relevante avaliar uma expansão do foco da estratégia proposta, estudando sua aplicabilidade em outras fases do processo de Resolução de Entidades como, por exemplo, na fase de Blocagem.

Referências

ABITEBOUL, S. et al. The lowell database research self-assessment. *Commun. ACM*, v. 48, n. 5, p. 111–118, Junho 2003. Disponível em: <http://dblp.uni-trier.de/db/journals/cacm/cacm48.html#AbiteboulABCCCDFGGGHHHIKPLMNSSSSSUWWZ05>.

ASLAM, J. A.; PELEKHOV, E.; RUS, D. The star clustering algorithm for static and dynamic information organization. *J. Graph Algorithms Appl.*, v. 8, p. 95–129, 2004. Disponível em: <http://dblp.uni-trier.de/db/journals/jgaa/jgaa8.html#AslamPR04>.

BANSAL, N.; BLUM, A.; CHAWLA, S. Correlation clustering. *Machine Learning*, Pittsburgh, PA, v. 56, n. 1-3, p. 89–113, 2004.

BELLAHSENE, Z.; BONIFATI, A.; RAHM, E. (Ed.). *Schema Matching and Mapping*. Springer, 2011. (Data-Centric Systems and Applications). ISBN 978-3-642-16518-4. Disponível em: <http://dblp.uni-trier.de/db/books/collections/Bellahsene2011.html>.

BERNSTEIN, P. A.; MADHAVAN, J.; RAHM, E. Generic schema matching, ten years later. *PVLDB*, Seattle, Washington, v. 4, n. 11, p. 695–701, 2011. Disponível em: <http://dblp.uni-trier.de/db/journals/pvldb/pvldb4.html#BernsteinMR11>.

BHATTACHARYA, I.; GETOOR, L. Query-time entity resolution. *J. Artif. Intell. Res. (JAIR)*, v. 30, p. 621–657, 2007. Disponível em: <http://dblp.uni-trier.de/db/journals/jair/jair30.html#BhattacharyaG07>.

BILENKO, M.; KAMATH, B.; MOONEY, R. J. Adaptive blocking: Learning to scale up record linkage. In: *ICDM*. IEEE Computer Society, 2006. p. 87–96. ISBN 0-7695-2701-9. Disponível em: <http://dblp.uni-trier.de/db/conf/icdm/icdm2006.html#BilenkoKM06>.

BILENKO, M.; MOONEY, R. J. Adaptive duplicate detection using learnable string similarity measures. In: *Proceedings of the Ninth ACM SIGKDD International*. Washington, DC: ACM, 2003.

BORGES, H. B. *Redução de dimensionalidade em bases de dados de expressão gênica*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Paraná, Curitiba - PR, Junho 2011.

CHEN, J. et al. A learning method for entity matching. In: *In Proceedings of 10th International Workshop on Quality in Databases*. Istanbul, Turkey: QDB, 2012.

CHRISTEN, P. *Data Matching*. Heidelberg: Springer, 2012. ISSN 2197-9723. ISBN 978-3-642-31163-5.

CHRISTEN, P.; GOISER, K. Quality and complexity measures for data linkage and deduplication. In: GUILLET, F.; HAMILTON, H. J. (Ed.). *Quality Measures in Data Mining*. Springer, 2007, (Studies in Computational Intelligence, v. 43). p. 127–151. ISBN 978-3-540-44911-9. Disponível em: <http://cs.anu.edu.au/~Peter.Christen/publications/qmdm-linkage.pdf>.

COHEN, W. W.; RAVIKUMAR, P.; FIENBERG, S. A comparison of string metrics for matching names and records. In: *Proceedings of the Workshop on Data Cleaning and Object Consolidation*. Washington, DC: KDD, 2003.

COVOES, T. F.; HRUSCHKA, E. R. Towards improving cluster-based feature selection with a simplified silhouette filter. *Inf. Sci.*, v. 181, n. 18, p. 3766–3782, 2011. Disponível em: <http://dblp.uni-trier.de/db/journals/isci/isci181.html#CovoesH11>.

DASH, M. et al. Feature selection for clustering - a filter solution. In: *ICDM*. IEEE Computer Society, 2002. p. 115–122. ISBN 0-7695-1754-4. Disponível em: <http://dblp.uni-trier.de/db/conf/icdm/icdm2002.html#DashCSL02>.

DASH, M.; LIU, H. Feature selection for classification. *Intelligent Data Analysis*, v. 1, n. 3, p. 131–156, 1997. Disponível em: <http://dx.doi.org/10.1016/S1088-467X(97)00008-5>.

DASH, M.; LIU, H. Feature selection for clustering. In: TERANO, T.; LIU, H.; CHEN, A. L. P. (Ed.). *PAKDD*. Springer, 2000. (Lecture Notes in Computer Science, v. 1805), p. 110–121. ISBN 3-540-67382-2. Disponível em: <http://dblp.uni-trier.de/db/conf/pakdd/pakdd2000.html#DashL00>.

DONG, X.; HALEVY, A.; MADHAVAN, J. *Reference reconciliation in complex information spaces.* 2005. In Proceedings of the 2005 ACM SIGMOD international Conference on Management of Data (Baltimore, Maryland, June 14 - 16, 2005). SIGMOD '05. ACM Press, New York, NY, 85-96.

DONG, X. L.; NAUMANN, F. Data fusion - resolving data conflicts for integration. *PVLDB*, Lyon, France, v. 2, n. 2, p. 1654–1655, 2009. Disponível em: <http://dblp.uni-trier.de/db/journals/pvldb/pvldb2.html#DongN09>.

DONG, X. L.; SAHA, B.; SRIVASTAVA, D. Less is more: Selecting sources wisely for integration. *Proc. VLDB Endow.*, VLDB Endowment, v. 6, n. 2, p. 37–48, dez. 2012. ISSN 2150-8097. Disponível em: <http://dx.doi.org/10.14778/2535568.2448938>.

DONG, X. L.; SRIVASTAVA, D. *Big Data Integration*. Morgan & Claypool Publishers, 2015. 1-198 p. (Synthesis Lectures on Data Management). Disponível em: <http://dx.doi.org/10.2200/S00578ED1V01Y201404DTM040>.

DRAISBACH, U.; NAUMANN, F. Dude: The duplicate detection toolkit. In: *In Proceedings of the International Workshop on Quality in Databases (QDB)*. Singapore: VLDB Endowment, 2010.

DRAPER, D.; HALEVY, A. Y.; WELD, D. S. The nimble integration engine. In: MEHROTRA, S.; SELLIS, T. K. (Ed.). *SIGMOD Conference*. ACM, 2001. p. 567–568. ISBN 1-58113-332-4. Disponível em: <http://dblp.uni-trier.de/db/conf/sigmod/sigmod2001.html#DraperHW01>.

DY, J. G.; BRODLEY, C. E. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, v. 5, p. 845–889, 2004. Disponível em: <http://dblp.uni-trier.de/db/journals/jmlr/jmlr5.html#DyB04>.

EVANGELISTA, L. O. et al. Adaptive and flexible blocking for record linkage tasks. *JIDM*, v. 1, n. 2, p. 167–182, June 2010. Disponível em: <http://dblp.uni-trier.de/db/journals/jidm/jidm1.html#EvangelistaCSM10>.

FAGIN, R. et al. Clio: Schema mapping creation and data exchange. In: *Conceptual Modeling: Foundations and Applications - Essays in Honor of John Mylopoulos*. [S.l.]: Springer-Verlag Berlin Heidelberg, 2009. p. 198–236.

FAN, W. et al. Reasoning about record matching rules. *PVLDB*, Lyon, France, v. 2, n. 1, p. 407–418, 2009. Disponível em: <http://dblp.uni-trier.de/db/journals/pvldb/pvldb2.html#FanJLM09>.

GU, L. et al. *Record linkage: Current practice and future directions*. Camberra, Austrália, 2003. Disponível em: <http://festivalofdoubt.uq.edu.au/papers/record_linkage.pdf>.

HALEVY, A. Y. *Data Integration: A Status Report*. Seattle, Washington, USA, 2003.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780123814791. Disponível em: <http://books.google.at/books?id=pQws07tdpjoC>.

HASSANZADEH, O. et al. Framework for evaluating clustering algorithms in duplicate detection. *PVLDB*, Lyon, France, v. 2, n. 1, p. 1282–1293, 2009.

HERNÁNDEZ, M. A.; STOLFO, S. J. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, Boston, v. 2, n. 1, p. 9–37, January 1998.

IVES, Z. G. et al. An adaptive query execution system for data integration. In: *Proceedings of the 1999 ACM SIGMOD international conference on Management of data*. Philadelphia, Pennsylvania, United States: ACM, 1999. p. 299–310. ISBN 1-58113-084-8. Disponível em: <http://portal.acm.org/citation.cfm?id=304182.304209&type=series>.

JACCARD, P. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, Zürich, Schweiz, v. 37, p. 547–579, 1901.

JARO, M. A. *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida.* 1989. Journal of the American Statistical Association 84:414–420.

JASPER, E. et al. View generation and optimisation in the automed data integration framework. In: EDER, J.; WELZER, T. (Ed.). *CAiSE Short Paper Proceedings*. CEUR-WS.org, 2003. (CEUR Workshop Proceedings, v. 74). ISBN 86-435-0549-8. Disponível em: <http://dblp.uni-trier.de/db/conf/caise/caisefo2003.html#JasperTMP03>.

JOUVE, P.-E.; NICOLOYANNIS, N. A filter feature selection method for clustering. In: HACID, M.-S. et al. (Ed.). *ISMIS*. Springer, 2005. (Lecture Notes in Computer Science, v. 3488), p. 583–593. ISBN 3-540-25878-7. Disponível em: <http://dblp.uni-trier.de/db/conf/ismis/ismis2005.html#JouveN05>.

KOPCKE, H.; RAHM, E. Frameworks for entity matching: A comparison. *Data Knowl. Eng.*, v. 69, n. 2, p. 197–210, 2010. Disponível em: <http://dblp.uni-trier.de/db/journals/dke/dke69.html#KopckeR10>.

KöPCKE, H.; RAHM, E. Training selection for tuning entity matching. In: MISSIER, P. et al. (Ed.). *QDB/MUD*. [s.n.], 2008. p. 3–12. Disponível em: <http://dblp.uni-trier.de/db/conf/iqis/qdbmud2008.html#KopckeR08>.

LEE, H. D. *Seleção de atributos importantes para extração de conhecimento de bases de dados*. Tese (Doutorado) — IC C-USP, São Carlos - SP, Dezembro 2005.

LENZERINI, M. Data integration: A theoretical perspective. In: *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. New York, NY, USA: ACM, 2002. (PODS '02), p. 233–246. ISBN 1-58113-507-6. Disponível em: <http://doi.acm.org/10.1145/543613.543644>.

LEVENSHTEIN, V. Binary Codes Capable of Correcting Deletions and Insertions and Reversals. *Soviet Physics Doklady*, v. 10, n. 8, p. 707–710, 1966.

LEVY, A. Combining artificial intelligence and databases for data integration. In: WOOLDRIDGE, M.; VELOSO, M. (Ed.). *Artificial Intelligence Today: Recent Trends and Developments*. [S.l.]: Springer-Verlag, Heidelberg, Germany, 1999. (Lecture Notes in Computer Science, 1600), p. 249–268.

LI, Y.; LU, B.-L.; WU, Z.-F. A hybrid method of unsupervised feature selection based on ranking. In: *ICPR (2)*. IEEE Computer Society, 2006. p. 687–690. ISBN 0-7695-2521-0. Disponível em: <http://dblp.uni-trier.de/db/conf/icpr/icpr2006-2.html#LiLW06>.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.*, v. 17, n. 4, p. 491–502, 2005. Disponível em: <http://dblp.uni-trier.de/db/journals/tkde/tkde17.html#LiuY05>.

MACEDO, D. C. d. *Comparação da redução de dimensionalidade de dados usando seleção de atributos e conceito de framework: um experimento no domínio de clientes*. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, Ponta Grossa - PR, Março 2012.

MCCALLUM, A.; NIGAM, K.; UNGAR, L. H. Efficient clustering of high-dimensional data sets with application to reference matching. In: *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM Press, 2000. p. 169–178. ISBN 1-58113-233-6.

MIHAILA, G. A.; RASCHID, L.; VIDAL, M.-E. Using quality of data metadata for source selection and ranking. In: *WebDB (Informal Proceedings)*. Dalas, TX: ACM, 2000. p. 93–98. Disponível em: <http://dblp.uni-trier.de/db/conf/webdb/webdb2000.html#MihailaRV00>.

MONGE, A.; ELKAN, C. *The field-matching problem: algorithm and applications.* 1996. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.

NAUMANN, F.; FREYTAG, J. C.; LESER, U. Completeness of information sources. *Information Systems*, Heidelberg, v. 29, n. 7, p. 583–615, 2000. Elsevier 2004.

PATRAO, D. F. C. et al. Ontocloud - a clinical information ontology based data integration system. In: BAX, M. P.; ALMEIDA, M. B.; WASSERMANN, R. (Ed.). *ONTOBRAS*. CEUR-WS.org, 2013. (CEUR Workshop Proceedings, v. 1041), p. 118–129. Disponível em: <http://dblp.uni-trier.de/db/conf/ontobras/ontobras2013.html#PatraoBFW13>.

RAHM, E.; BERNSTEIN, P. A. A survey of approaches to automatic schema matching. *The VLDB Journal*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, v. 10, n. 4, p. 334–350, 2001. ISSN 1066-8888. Disponível em: <http://dx.doi.org/10.1007/s007780100057>.

RIBEIRO-NETO, B.; BAEZA-YATES, R. *Modern Information Retrieval*. [S.l.]: ACM Press/Addison-Wesley, 1999.

SHETH, A. P.; LARSON, J. A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 22, n. 3, p. 183–236, 1990. ISSN 0360-0300. Disponível em: <http://delivery.acm.org/10.1145/100000/96604/p183-sheth.pdf?key1=96604&key2=8693487821&coll=GUIDE&dl=ACM&CFID=107002774&CFTOKEN=13258373>.

SINGLA, P.; DOMINGOS, P. Entity resolution with markov logic. In: *ICDM*. IEEE Computer Society, 2006. p. 572–582. Disponível em: <http://dblp.uni-trier.de/db/conf/icdm/icdm2006.html#SinglaD06>.

SU, W. et al. Record Matching over Query Results from Multiple Web Databases. *IEEE Transactions on Knowledge and Data Engineering*, v. 22, n. 4, p. 578–589, 2010.

WANG, J. et al. Entity matching: How similar is similar. *PVLDB*, Seattle, Washington, v. 4, n. 10, p. 622–633, 2011. Disponível em: <http://dblp.uni-trier.de/db/journals/pvldb/pvldb4.html#WangLYF11>.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, M. E. Sharpe, Inc., Armonk, NY, USA, v. 12, n. 4, p. 5–33, mar. 1996. ISSN 0742-1222. Disponível em: <http://dx.doi.org/10.1080/07421222.1996.11518099>.

ZHAO, Z.; LIU, H. Spectral feature selection for supervised and unsupervised learning. In: GHAHRAMANI, Z. (Ed.). *ICML*. ACM, 2007. (ACM International Conference Proceeding Series, v. 227), p. 1151–1157. ISBN 978-1-59593-793-3. Disponível em: <http://dblp.uni-trier.de/db/conf/icml/icml2007.html#ZhaoL07>.