Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Extração de Conceitos e Relações Taxonômicas usando Análise de Conceitos Formais e Agrupamento Fuzzy de Dados

Suzane Carol de Lima

Orientadora: Profa. Dra. Heloisa de Arruda Camargo

São Carlos, SP, Brasil

Fevereiro, 2017

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Extração de Conceitos e Relações Taxonômicas usando Análise de Conceitos Formais e Agrupamento Fuzzy de Dados

Suzane Carol de Lima

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Inteligência Artificial.
Orientadora: Profa. Dra. Heloisa de Arruda Camargo

São Carlos, SP, Brasil

Fevereiro, 2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Suzane Carol de Lima, realizada em 17/02/2017:

Profa. Dra. Heloisa de Arruda Camargo
UFSCar

Profa. Dra. Marilde Terezinha Prado Santos
UFSCar

Profa. Dra. Cristiane Akemi Yaguinuma
IFSP

Agradecimentos

Primeiramente gostaria de agradecer a Deus por me ter aberto essa oportunidade na minha vida. Por me dar forças nessa caminhada, coragem e fé. Agradeço a Nossa Senhora por me abraçar como filha nos momentos de aflição.

Aos meus guerreiros e amados pais Edna e Elio, que sempre estiveram ao meu lado, me apoiando, guiando e amando. Obrigada mãe pela paciência, amor, dedicação e por não me deixar desistir dos meu sonhos. Obrigada pai por me apoiar nos momentos difíceis e dar carinho em todos os momentos da minha vida.

Agradeço ao meu irmão Daniel por ter me ajudado muitas vezes, sendo meu confidente e me fazendo rir nas piores situações.

À minha avó Antônia por todas as vezes de aflições e desespero, por me transmitir paz e aconselhar. Sem contar as inúmeras vezes que rezou por essa conquista.

Agradeço a minha orientadora Prof. Dra. Heloisa de Arruda Camargo por ter aceitado embarcar nessa jornada comigo. Obrigada pela paciência, amizade e por me acolher na família CIG, no qual pude aprender muito durante esses anos.

Obrigada a todos os professores que compartilharam do seu conhecimento e me ensinaram durante esse período do mestrado. À Capes pelo apoio financeiro a pesquisa e a UFSCar, pela infraestrutura e recursos concedido.

Agradeço por fim, aos meus amigos de laboratório pelo companheirismo e todos os momentos juntos (alegres e tristes). Pelas discussões acadêmicas e sobre a vida. Obrigada por fazer parte, mesmo que pequena na vida de cada um!

Resumo

Algumas estruturas para representação do conhecimento se organizam a partir conceitos e relacionamentos entre conceitos, entre as quais podemos citar as redes semânticas e as ontologias. Uma importante ferramenta que auxilia no processo de criação dessas estruturas é a Análise de Conceitos Formais (ACF). ACF tem sido aplicada em diversos campos de pesquisa, tais como mineração de dados, aprendizado de máquina, inteligência artificial e engenharia de Software. A ACF pode ser considerado atualmente um formalismo importante para a representação do conhecimento, extração e análise com aplicações em diferentes áreas, sendo que é utilizado para a construção de ontologias, pois oferece uma base para desenvolvimento e implementação de métodos para extrair conceitos ontológicos, bem como a taxonomia ontológica envolvendo os conceitos extraídos. Na análise de conceitos formais, conceitos são conjuntos de objetos que compartilham dos mesmos atributos. Conceitos são extraídos de um conjunto de dados e organizados na forma de um reticulado de conceitos, definido pela relação de inclusão entre os conceitos. A estrutura do Reticulado de Conceitos pode se tormar grande em função do número elevado de conceitos e relações, tornando uma estrutura complexa, e muitas vezes, de dificil processamento computacional. O objetivo deste trabalho é reduzir o contexto formal de um domínio específico, utilizando dois algoritmos de agrupamento *fuzzy*, para que seja gerado um Reticulado de conceitos também reduzido. Os resultados mostraram que o algoritmo de agrupamento *Fuzzy C-Means* teve um desempenho superior que o algoritmo *Possibilistic Fuzzy C-Means*.

Palavras-chaves: Reticulado de conceitos; FCA; redução; agrupamento *fuzzy*.

Abstract

Some structures for knowledge representation are organized from concepts and relationships between concepts, among which we can mention semantic networks and ontologies. An important tool that help in the creation process of these structures is the Formal Concept Analysis (FCA). FCA has been applied in several fields of research, such as data mining, machine learning, artificial intelligence and Software Engineering. The FCA can now be considered an important formalism for the representation of knowledge, extraction and analysis with applications in different areas, and is used for the construction of ontologies, since it provides a basis for the development and implementation of methods to extract ontological concepts as well as the ontological taxonomy involving the extracted concepts. In the Formal Concept Analysis, concepts are sets of objects that share the same attributes. Concepts are extracted from a set of data and organized in the form of a Concept Lattice, defined by the relation of inclusion between concepts. The structure of the Conceptual Framework can become large due to the high number of concepts and relations, making a complex structure, and often difficult computational process. The purpose of this work is to reduce the formal context of a specific domain by using two fuzzy clustering algorithms, so that a reduced Concept Lattice is generated.The results showed that the Fuzzy C-Means clustering algorithm performed better than Possibilistic Fuzzy C-Means algorithm.

Keywords: Concept Lattice; FCA; reduce; fuzzy clustering.

Lista de Ilustrações

Figura 2.1 Reticulado de conceitos 18
Figura 2.2 Função de pertinência Triangular 19
Figura 2.3 Função de pertinência Trapezoidal 20
Figura 2.4 Função de pertinência Gaussiana 20
Figura 2.5 Representação gráfica de função de Pertinência trapezoidal dos conceitos jovem, adulto e idoso (KLIR; YUAN, 1995) 21
Figura 2.6 Função de Pertinência trapezoidal dos conceitos jovem, adulto e idoso (KLIR; YUAN, 1995) 21
Figura 2.7 Objetos físicos:(a) Objetos antes do agrupamento.(b) Objetos após o agrupamento. Adaptado de (BEZDEK; PAL, 1992) 24

Figura 4.1 Ilustração das etapas do projeto proposto 36
Figura 4.2 Exemplo de base de dados de Documentos e Termos 37

Lista de Tabelas

Tabela 2.1 Exemplo de Contexto formal 16
Tabela 2.2 Exemplo de base de dados 22
Tabela 2.3 Exemplo de Partição *fuzzy* 23
Tabela 2.4 Partição *crisp* resultante do agrupamento 24
Tabela 2.5 Partição *fuzzy* resultante do agrupamento 24

Tabela 3.1 Exemplo de Contexto Formal 30
Tabela 3.2 Contexto Formal da matriz Termo-Documento antes da aplicação de SVD extraído de (CHEUNG; VOGEL, 2005) 31
Tabela 3.3 Contexto Formal da matriz Termo-Documento depois da aplicação de SVD extraído de (CHEUNG; VOGEL, 2005) 32
Tabela 3.4 Comparação dos resultados das técnicas de agrupamento *Fuzzy C-Means* e SVD (KUMAR; SRINIVAS, 2010) 32

Tabela 4.1 Trecho do contexto formal sobre documentos e termos 37

Tabela 5.1 Base de dados utilizadas nos experimentos 42
Tabela 5.2 Média dos índices de validação de FCA para a base de dados Iaarticle . 43
Tabela 5.3 Média dos índices de validação de PFCM para a base de dados Iaarticle 43
Tabela 5.4 Média dos índices de validação do FCM para a base de dados Opinosis 44
Tabela 5.5 Média dos índices de validação do PFCM para a base de dados Opinosis 44
Tabela 5.6 Média dos índices de validação do FCM para a base de dados Reuters . 44
Tabela 5.7 Média dos índices de validação do PFCM para a base de dados Reuters 45
Tabela 5.8 Média dos índices de validação do FCM para a base de dados Newyork-times 45
Tabela 5.9 Média dos índices de validação do PFCM para a base de dados Newyork-times 45
Tabela 5.10 Média dos índices de validação do FCM para a base de dados Newgroups 46
Tabela 5.11 Média dos índices de validação do PFCM para a base de dados Newgroups 46

Tabela 5.12 Número de conceitos e arestas do reticulado de gerados antes do agrupamento *fuzzy* 47
Tabela 5.13 Número de conceitos e arestas dos reticulados gerados após a aplicação do algoritmo FCM 47
Tabela 5.14 Número de conceitos e arestas dos reticulados gerados após a aplicação do algoritmo PFCM 47

Lista de Algoritmos

Algoritmo 1 Fuzzy C-Means (FCM) 26
Algoritmo 2 Possibilistic Fuzzy C-Means (PFCM) 28

Sumário

1 Introdução 12
1.1 Contextualização 12
1.2 Motivação e Objetivo 13
1.3 Estrutura da dissertação 14

2 Conceitos Básicos 15
2.1 Análise de Conceitos Formais 15
2.1.1 Contexto formal 16
2.1.2 Conceito formal 16
2.2 Reticulado de conceitos 17
2.3 Agrupamento Fuzzy 18
2.3.1 Conjuntos *Fuzzy* 18
2.3.2 Algoritmos de Agrupamento Fuzzy 22
2.3.3 *Fuzzy C-Means* (FCM) 25
2.3.4 **Possibilistic Fuzzy C-Means** (PFCM) 26

3 Redução de Reticulados de Conceitos por Agrupamento de Dados 29
3.1 Técnicas de redução de reticulados de conceitos 29
3.1.1 Remoção de informação redundante 29
3.1.2 Simplificação 31
3.1.3 Seleção 33
3.1.4 Análise das técnicas de redução 34

4 Redução de Reticulado de Conceitos usando Agrupamento Fuzzy 35
4.1 Proposta 35
4.1.1 Transformação do conjunto de dados em um Contexto Formal 36
4.1.2 Agrupamento do contexto formal 37
4.1.3 Cálculo de Decomposição de matriz utilizando matriz de centróides . 39
4.1.4 Transformação para valores binários 40
4.1.5 Aplicação da ACF no contexto formal reduzido 40

5 Experimentos e Análise dos Resultados 41
5.1 Experimentos 41

5.2 Avaliação dos agrupamentos 42
5.3 Geração do reticulado de conceitos 46

6 Conclusão 49
6.1 Considerações finais 49
6.2 Trabalhos Futuros 50

Referências 51

1

Introdução

Neste capítulo são apresentados o contexto, os pontos motivacionais para o desenvolvimento da proposta, assim como o objetivo. Por fim, no final deste capítulo é apresentada a organização da dissertação.

1.1 Contextualização

O campo de pesquisa da Inteligência Artificial é um campo extenso, sendo que muitas de suas subáreas requerem o desenvolvimento de um sistema inteligente capaz de adquirir, representar e manipular o conhecimento. Toda representação de conhecimento deve possibilitar a representação de objetos, seus atributos e o relacionamento entre os mesmos, tendo como características principais (REZENDE, 2003):

- Transparência: permitir o entendimento do que está sendo dito;
- Rapidez: possibilitar o armazenamento e a recuperação de informações em tempo curto;
- Computabilidade: possibilitar a sua criação, utilizando um procedimento computacional existente.

Com a representação do conhecimento é possível inferir novos conhecimentos através de relações, fatos e conceitos, sendo possível resolver problemas complexos.

Existem várias formas de representar e organizar conhecimento. Uma dessas formas de organização se dá a partir de termos: glossários e dicionários. As estruturas que são organizadas à partir de classificação e categorias: cabeçalhos e taxonomias. E por fim, existem as estruturas que se organizam à partir de conceitos e relacionamentos: redes semânticas e as ontologias. Existem algumas ferramentas que ajudam no processo de criação dessas estruturas e uma delas é a Análise de Conceitos Formais (ACF) (GANTER; WILLE, 1999).

ACF tem sido aplicada em diversos campos de pesquisa, tais como mineração de dados, aprendizagem de máquinas, inteligência artificial e Engenharia de Software. A ACF pode ser considerado atualmente um formalismo importante para a representação, extração e análise do conhecimento com aplicações em diferentes áreas. A ACF pode ser utilizada para a construção de ontologias.

Ontologias (GUARINO; OBERLE; STAAB, 2009) são utilizadas para a representação de um conjunto de conceitos e seus relacionamentos de um determinado domínio. A ACF oferece uma base para desenvolvimento e implementação de métodos para extrair conceitos ontológicos, bem como a taxonomia ontológica envolvendo os conceitos extraídos. A ideia principal do funcionamento da ACF é transformar o conjunto de dados em uma matriz relacional composta por relações entre objetos e atributos denominada de contexto formal, seguido da aplicação de operadores nessa matriz relacional, gerando conceitos formais e posteriormente organizar o conjunto de conceitos em uma estrutura taxonômica, o reticulado de conceitos (HAAV, 2004), (OBITKO et al., 2004),(BAIN, 2003),(TOUZI; MASSOUD; AYADI, 2013). Maio et al. (2009) utilizou uma extensão da ACF, a Análise de Conceitos Formais *Fuzzy* que encorpora a teoria dos conjuntos *fuzzy* para a geração automática de ontologias *fuzzy*. Os autores realizam um mapeamento simplificado dos elementos do reticulado de conceitos *fuzzy* para uma ontologia *fuzzy*, não levando em consideração o número elevado de conceitos gerados.

1.2 Motivação e Objetivo

Na Análise de Conceitos Formais, a estrutura do reticulado de conceitos pode se tornar grande em função do número elevado de conceitos e relações, tornando uma estrutura complexa, e muitas vezes, de difícil processamento computacional. Dias (2016), expõe a importância de manter todos os relacionamentos entre os conceitos do contexto formal no reticulado de conceitos ao se tratar de completude, porém , em sua maioria, o volume de relacionamentos sobrecarrega o reticulado.

Observa-se que mesmo com um conjunto de dados pequeno, o número de conceitos gerado é alto e o reticulado de conceitos pode atingir um alto nível de complexidade no relacionamento entre esses conceitos, dificultando a análise dos dados e consequentemente, causar um alto custo computacional.

Apesar de o pior caso ser raramente encontrado em casos práticos (GODIN; SAUNDERS; GECSEI, 1986), o custo computacional pode ser um grande obstáculo para muitas aplicações, dificultando a análise do reticulado conceitual ou até mesmo impossibilitando a geração do reticulado de conceitos.

Para aplicações que dependem da Análise Conceitos Formais também podem ser prejudicados com esse alto nível de complexidade, como é o caso de construções automá-

ticas de ontologias. A dificuldade nesse caso também se dá na realização no mapeamento dos elementos ontológicos encontrados com a ACF para a ontologia. Quanto maior a estrutura do reticulado de conceitos, maior a ontologia gerada, resultando assim, em um custo computacional tanto na realização do mapeamento quanto no processo de recuperação de informação ou qualquer operação necessária na ontologia.

O objetivo da dissertação apresentado neste documento é propor uma metodologia para extrair conceitos e relacionamentos taxonômicos utilizando Análise de Conceitos Formais, tendo em vista a redução do Contexto formal de um domínio específico, utilizando dois algoritmos de agrupamento *fuzzy*, visando gerar um reticulado de conceitos menor que o original.

1.3 Estrutura da dissertação

O capítulo 1 introduz e fornece informações básicas relacionadas a motivação do problema a ser tratado, assim como o objetivo proposto desta dissertação. A fundamentação teórica utilizada para o desenvolvimento deste trabalho é detalhada no capítulo 2, explicando conceitos básicos fundamentais para o este trabalho. O capítulo 3 aborda as técnicas de redução de reticulado de conceitos encontradas na literatura. O capítulo 4 apresenta a proposta de trabalho bem como as etapas do desenvolvimento da proposta. No capítulo 5, são descritos os resultados do trabalho e por fim, no capítulo 6 é apresentada a conclusão e trabalhos futuros.

2

Conceitos Básicos

Neste capítulo são apresentados os conceitos básicos relativos a Análise de Conceitos Formais e Agrupamento de Dados, necessários para compreensão desta proposta. O papel das técnicas derivadas desses conceitos no trabalho desenvolvido será apresentado com mais detalhes no capítulo 4.

2.1 Análise de Conceitos Formais

A técnica de Análise de Conceitos Formais (AFC) foi introduzido na década de 1980 pelo matemático alemão Rudolf Wille como um método para realização de análise de dados, sendo baseada na teoria dos reticulados (DAVEY B. A.;PRIESTLEY, 1990), permitindo assim a visualização, investigação e interpretação dos dados e suas estruturas, assim como implicações e dependências (WILLE., 1997). É um método que pode ser utilizado em diversas áreas do conhecimento, desde a biologia até a engenharia industrial (WOLF, 1991). Na ciência da computação, a ACF é utilizada para construir automaticamente estruturas conceituais, provendo organização para os dados. Essa técnica pode ser interpretada como uma técnica de agrupamento conceitual (CIMIANO, 2006).

ACF estrutura os dados em unidades que são chamadas de conceitos formais. Esses conceitos formais são organizados na forma de um reticulado de conceitos, sendo estas duas nomenclaturas abstrações matemáticas para os conceitos e a hierarquia de conceitos, sendo estes vindos da filosofia (WILLE, 2005). Segundo Wille (2005), os conceitos podem ser entendidos, do ponto de vista filosófico, como unidades básicas do pensamento dos seres humanos, sendo a formação destas unidades o resultado de processos dinâmicos em ambientes culturais e sociais. Na mesma linha de raciocínio, conceitos têm como característica suas extensões (objetos que pertencem a um conceito) e intensões (atributos que descrevem as propriedades e os significados de todos os objetos presentes nas extensões). Conceitos se relacionam de forma hierárquica, sendo assim, existem relacionamentos do tipo subconceito-superconceito, indicando que a extensão do subconceito está contida na extensão do superconceito e a intensão do subconceito contém a intensão do superconceito.

Inicialmente, antes de descrever matematicamente as extensões e intensões, é necessário introduzir a noção de contexto formal.

2.1.1 Contexto formal

Contextos formais são definidos pela tripla (G, M, I), onde:

- G é o conjunto formado pelas entidades do domínio (objetos formais);
- M é constituído pelas características dessas entidades (atributos formais);
- I é uma relação binária sobre $G \times M$, chamada de relação de incidência ou relação binária, que associa um objeto formal ao seu respectivo atributo. Assim a relação gIm é lida como: "o objeto g tem o atributo m";

A tabela 2.1 apresenta uma definição de contexto formal, no qual os objetos são representados pelo conjunto $G = \{x_1, x_2, x_3, x_4, x_5\}$,o conjunto de atributos $M = \{y_1, y_2, y_3, y_4\}$ e O conjunto I corresponde às relações entre G e M.

Tabela 2.1 – Exemplo de Contexto formal

I	y1	y2	y3	y4
x1	1	1	1	1
x2	1	0	1	1
x3	0	1	1	1
x4	0	1	1	1
x5	1	0	0	0

2.1.2 Conceito formal

A definição de conceitos formais a partir de contextos formais se utiliza de operadores de derivação. Utilizando dois conjuntos arbitrários de objetos e atributos representados respectivamente por A e B, seus operadores de derivação $A^{\uparrow}$ e $B^{\downarrow}$ são definidos como (WILLE, 2005):

$$A^{\uparrow} = \{m \in M / \text{gIm para todo } g \in A\}$$

$$B^{\downarrow} = \{g \in G / \text{gIm para todo } m \in B\}$$

Considerando que $A^{\uparrow}$ determina todos os atributos em M compartilhados pelos objetos em A e $B^{\downarrow}$ determina todos os objetos em G que compartilham atributos em B, um conceito formal em (G, M, I) é definido pelo par (A, B) se e somente se $A \subseteq G$, $B \subseteq M$, tal que $A^{\uparrow} = B$ e $B^{\downarrow} = A$ (PRISS, 2006; WILLE, 2005). Assim, os operadores $A^{\uparrow}$ e $B^{\downarrow}$ expressam a conexão Galois (MORAES, 2012), formando conceitos do

tipo $(extensão, intensão)$. Aplicando-se os operadores $\uparrow\downarrow$ na tabela 2.1, obtemos alguns resultados:

- $\{x_2\}^{\uparrow} = \{y_1, y_3, y_4\}$
- $\{x_2, x_3\}^{\uparrow} = \{y_3, y_4\}$
- $\{y_1\}^{\downarrow} = \{x_1, x_2, x_5\}$
- $\{y_2, y_3\}^{\downarrow} = \{x_1\}$

Após a aplicação dos operadores $\uparrow\downarrow$ em todas as combinações possíveis para o conjunto G e M respectivamente, são realizadas as comparações e assim, gerados os conceitos formais, como pode ser observado a seguir:

$$(A, B) = (\{x_1, x_2, x_3, x_4\}\,\{y_3, y_4\})$$

, pois

- $\{x_1, x_2, x_3, x_4\}^{\uparrow} = \{y_3, y_4\}$
- $\{y_3, y_4\}^{\downarrow} = \{x_1, x_2, x_3, x_4\}$

2.2 Reticulado de conceitos

Para organizar hierarquicamente os conceitos formais, é necessário estabelecer a relação $subconceito - superconceito$ matematicamente, assim definida (WILLE, 2005):

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow (\Leftrightarrow B_1 \supseteq B_2)$$

Assim, a relação subconceito-superconceito pode ser de "inclusão da extensão" ($A_1 \subseteq A_2$) ou "inclusão das intensões, mas em ordem inversa" ($B_1 \supseteq B_2$). Ainda utilizando-se do exemplo da tabela 2.1, a Figura 2.1 ilustra a relação de ordem para os conceitos, no qual o conjunto de todos os conceitos formais de um contexto (G, M, I) junto com a relação de ordem formam um reticulado completo, chamado de reticulado de conceitos de (G, M, I), sendo denotado por $\mathcal{B}(G, M, I)$. Isso significa que para todo conjunto de conceitos existe um único maior subconceito (o ínfimo) e um único menor superconceito (o supremo).

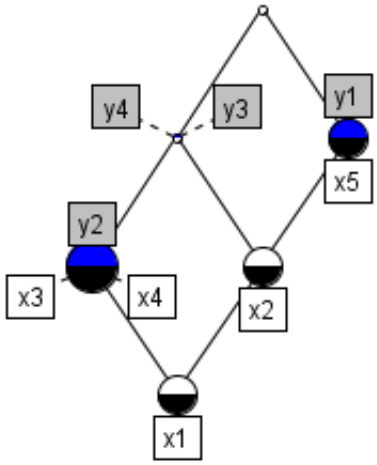

Figura 2.1 – Reticulado de conceitos

Cada conceito é representado por um nó do reticulado de conceitos. Os nós com a cor azul infere a existência de atributos ligados ao conceito. Nós coloridos em preto infere a existência de objetos ligados ao conceito e nós coloridos de azul e preto infere a existência de atributos e objetos ligados ao conceito.

2.3 Agrupamento Fuzzy

Agrupamento de dados é uma das formas de aprendizado estudada para se encontrar a similaridade entre os dados e assim permitir que esses dados pertençam a algum grupo que compartilham de características similares. O agrupamento de dados clássico, é um agrupamento rígido e o agrupamento fuzzy é mais flexível, mais parecido com o pensamento humano e com os problemas do mundo real. Porém, antes de falar sobre o agrupamento *fuzzy* e os algoritmos utilizados nesse trabalho de pesquisa, é necessário relembrar um pouco sobre conjuntos *fuzzy*, apresentada na subseção 2.3.1.

2.3.1 Conjuntos *Fuzzy*

O conjunto *fuzzy* é uma extensão dos conjuntos clássicos (*crisp*). Para qualquer conjunto tradicional, pode-se definir uma função denominada de função característica, que define quais elementos pertencem ou não a um determinado conjunto. Dado um elemento x de um conjunto universo X, a função característica γ_A pode ser representada por:

$$\gamma_A(x) = \begin{cases} 1 & \text{para } x \in A \\ 0 & \text{para } x \notin A \end{cases}$$

Um conjunto *fuzzy* é definido em um universo do discurso (conjunto base) e é caracterizado pela sua função característica. Como são conjuntos *fuzzy*, a função característica é denominada função de pertinência que tem como objetivo associar elementos de um conjunto base X a números reais do intervalo [0,1], que expressam a extensão com que o elemento se enquadra na categoria representada pelo conjunto *fuzzy*. Dado um conjunto X de elementos, uma das notações mais comuns de função de pertinência é:

$$\mu_A : X \longrightarrow [0, 1]$$

A teoria dos conjuntos *fuzzy* está fundamentada no conceito de pertinência, no qual, primeiramente qualquer função

$$\mu_A : x \longrightarrow [0, 1]$$

descreve uma função de pertinência associada a um conjunto *fuzzy* A . As funções de pertinência mais utilizadas na literatura possuem o formato triangular, trapezoidal e gaussiano. Essas funções, bem como seus respectivos gráficos são apresentados abaixo.

- Função Triangular: definida pelos parâmetros a, m e b, no qual $a \leq m \leq b$.

$$\mu_A(x) = \begin{cases} 0 & \text{se } x \leq a \\ \frac{x-a}{m-a} & \text{se } x \in (a, m) \\ 1 & se\ x = m \\ \frac{b-x}{b-m} & \text{se } x \in (m, b) \\ 0 & se\ x \geq b \end{cases}$$

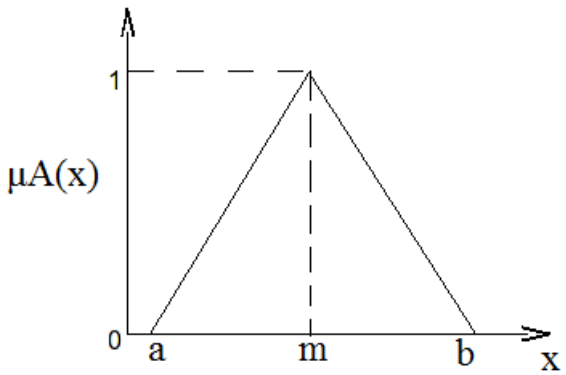

Figura 2.2 – Função de pertinência Triangular

- Função Trapezoidal: definida pelos parâmetros a, m, n e b, no qual $a \leq m < n \leq b$.

$$\mu_A(x) = \begin{cases} 0 & \text{se } x \leq a \\ \frac{x-a}{m-a} & \text{se } x \in (a, m) \\ 1 & se\ x \in (m, n) \\ \frac{b-x}{b-n} & \text{se } x \in [n, b] \\ 0 & se\ x \geq b \end{cases}$$

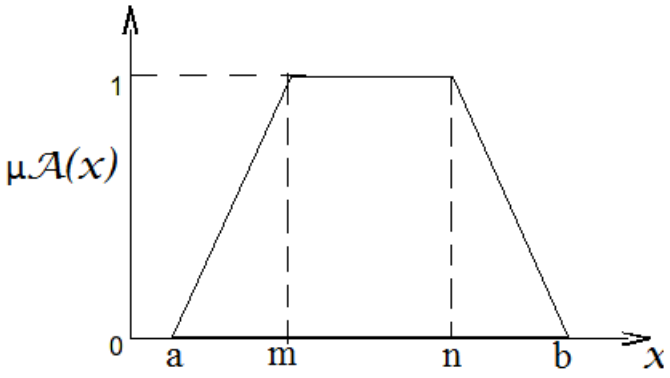

Figura 2.3 – Função de pertinência Trapezoidal

- Função Gaussiana: definida pelos parâmetros m e k, sendo $k > 0$.

$$\mu_A(x) = e^{-k(x-m)^2}$$

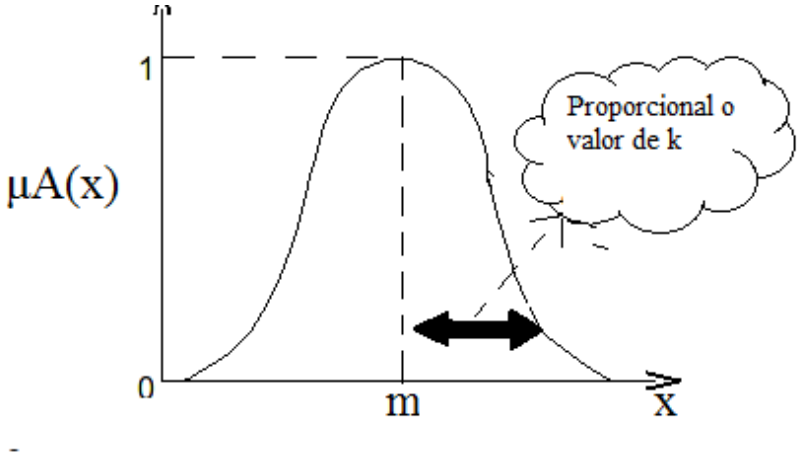

Figura 2.4 – Função de pertinência Gaussiana

O valor obtido por $\mu_A(x)$ representa o grau de pertinência do elemento x no conjunto *fuzzy* A. Quanto mais próximo de 1 for o valor de $\mu_A(x)$, então maior será o grau de pertinência do elemento x no conjunto A. Alternativamente, a função de pertinência

pode ser denotada pela letra que dá nome ao conjunto *fuzzy*, ou seja:

$$A : x \longrightarrow [0, 1]$$

sendo que $A(x)$ representa o grau de pertinência de x em A. Para exemplificar o conceito de conjunto *fuzzy* (KLIR; YUAN, 1995), 1995], três conjuntos representam os conceitos de jovem, adulto e idoso. A representação desses conjuntos é definida pela função de pertinência, sendo esta, trapezoidal μ_1 (jovem), μ_2(adulto) e μ_3(idoso), apresentada na figura 2.5. Essas funções são definidas no intervalo [0, 80], cuja representação gráfica é ilustrada na 2.5. A figura 2.6 apresenta as funções de pertinência para cada conjunto.

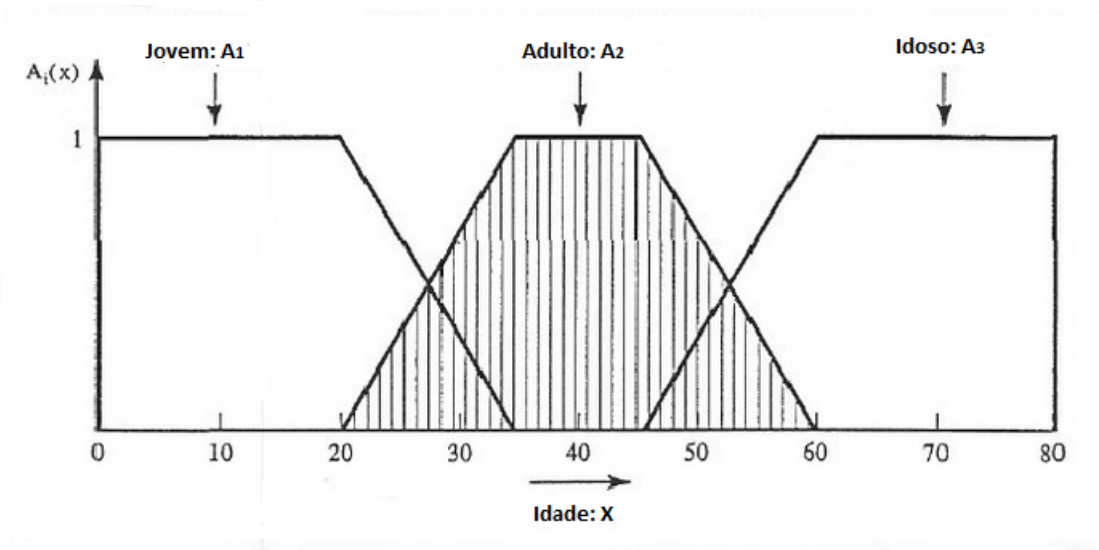

Figura 2.5 – Representação gráfica de função de Pertinência trapezoidal dos conceitos jovem, adulto e idoso (KLIR; YUAN, 1995)

$$A1(x)\begin{cases} 1 & Se\ x \leq 20 \\ \dfrac{35-x}{15} & Se\ 20 < x < 35 \\ 0 & Se\ x \geq 35 \end{cases}$$

$$A2(x)\begin{cases} 0 & Se\ x \leq 20\ ou\ x \geq 60 \\ \dfrac{x-20}{15} & se\ 20 < x < 35 \\ \dfrac{60-x}{15} & Se\ 45 < x < 60 \\ 1 & Se\ 35 \leq x \leq 45 \end{cases}$$

$$A3(x)\begin{cases} 0 & Se\ x \leq 45 \\ \dfrac{x-45}{15} & Se\ 15 < x < 60 \\ 1 & Se\ x \geq 60 \end{cases}$$

Figura 2.6 – Função de Pertinência trapezoidal dos conceitos jovem, adulto e idoso (KLIR; YUAN, 1995)

2.3.2 Algoritmos de Agrupamento Fuzzy

Agrupamento de dados é uma das questões fundamentais para o reconhecimento de padrões (KLIR; YUAN, 1996), sendo este responsável por encontrar padrões no dados e tornando possível o agrupamento desses dados em pseudo-partições ou partições (grupos) de acordo com as características encontradas. Então divide-se os objetos de dados em subconjuntos sem sobreposição, tal que cada objeto pertença exatamente a um subconjunto. O processo de agrupamento visa agrupar os dados de forma com que a similaridade entre os dados de um grupo é maximizada enquanto a similaridade seja minimizada entre dados de grupos diferentes.

Nos agrupamentos clássicos, os objetos pertencem a um grupo de forma disjunta, ou seja, pertence a apenas a um grupo. Alguns dados podem ter características similares com mais de um grupo, sendo então necessário introduzir a teoria de conjuntos *fuzzy* para tratar fatores de incerteza e imprecisão.

Tomando como exemplo uma base de dados $X = \{x_1, x_2, ..., x_j\}$ é representada por vetores n -dimensionais de atributos $X_j = \{x_{j1}, x_{j2}, ..., x_{jn}\}^T \in \mathbb{R}^n$. A base de dados X é denominada matriz de dados.

Tabela 2.2 – Exemplo de base de dados

	Atributo1	Atributo2	Atributo3
Objeto1	1	7	1
Objeto2	5	3	2
Objeto3	2	0	1

$$X = \{x_1, x_2, x_3\}\, x_i \in \mathbb{R}^3 \forall_i$$

- $x_1 = [1, 7, 1]^T$
- $x_2 = [5, 3, 2]^T$
- $x_3 = [2, 0, 1]^T$

No agrupamento *fuzzy*, as partições *crisp* dão lugar a partições *fuzzy*. Uma partição *fuzzy*, também conhecida por matriz de pertinência *fuzzy* $U = [u_{ij}]$ de tamanho kxN, onde k é o número de grupos e N é o número de objetos de X. Os valores de u_{ij} são correspondentes aos graus de pertinência entre o j-ésimo objeto e o i-ésimo grupo, seguindo as seguintes restrições

$$0 < \sum_{j=1}^{N} u_{ij} < N, i = 1, ..., k \tag{2.1}$$

Para as partições *fuzzy* probabilísticas, a soma dos graus de pertinência de cada grupo para com o respectivo objeto devem ser iguais a 1, como é apresentado na equação (2.2)

$$\sum_{i=1}^{k} u_{ij} = 1, j = 1, ..., N \tag{2.2}$$

$$0 < \sum_{j=1}^{N} u_{ij} < N, i = 1, ..., k \tag{2.3}$$

A tabela 2.3 apresenta um exemplo de partição *fuzzy* probabilística, no qual $k = 2$ e $N = 3$ e a soma das pertinências é 1.

Tabela 2.3 – Exemplo de Partição *fuzzy*

	objeto1	objeto2	objeto3
C_1	0.6	1	0.1
C_2	0.4	0	0.9
total	1	1	1

Então $U = \{C_1, C_2\}$ é uma partição 2-*fuzzy* de X. Bezdek e Pal (1992) exemplificou o processo de agrupamento *fuzzy* como apresenta a figura 2.7.

Para agrupar um conjunto de objetos, três grupos foram criados: MAÇÃS (M),PERAS (P) e LARANJAS (L). O processo proposto no exemplo é colocar sobre cada objeto um rótulo que indica a qual grupo ele pertence. Objetos marcados com P significa que pertence ao grupo das Peras,e assim com os outros grupos. O objeto oval O3 é um LIMÃO, representando uma anomalia nos dados, pois a rotulação definida neste exemplo não existe grupo LIMÃO.

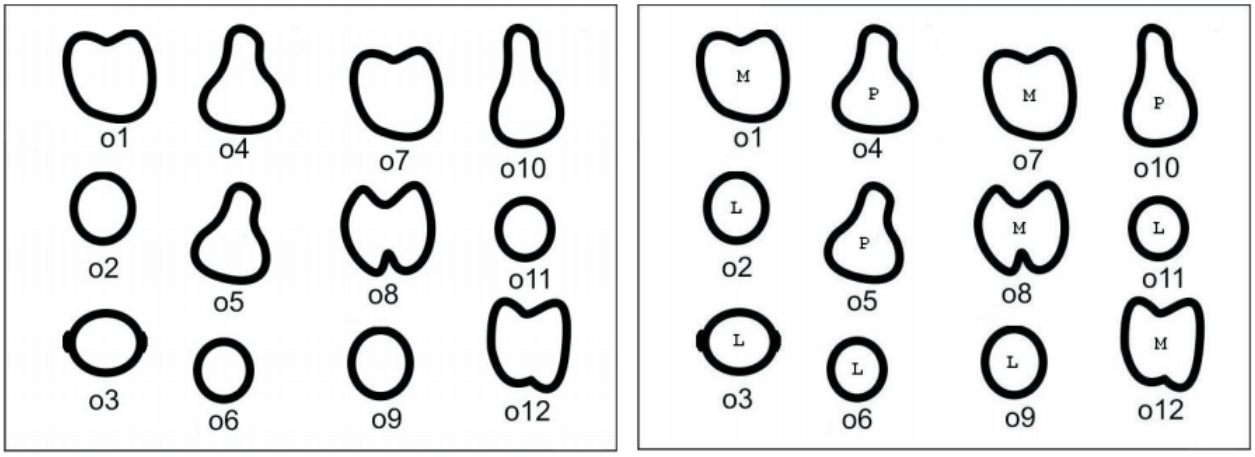

Figura 2.7 – Objetos físicos:(a) Objetos antes do agrupamento.(b) Objetos após o agrupamento. Adaptado de (BEZDEK; PAL, 1992)

As partições *crisp* e *fuzzy* são apresentadas nas tabelas 2.4 e 2.5 respectivamente. O resultado do agrupamento apresentado na figura 2.7 mostra que o objeto O3 foi rotulado como LARANJA (L), assim como no agrupamento *crisp*, onde o objeto pertence somente a esse grupo. No entanto, na partição *fuzzy* é possível observar que o objeto O3 também possui características com outros grupos (P e M) com graus de pertinência menores que o grupo L. O agrupamento *fuzzy* mostra ser mais flexível que o agrupamento *crisp*.

Tabela 2.4 – Partição *crisp* resultante do agrupamento

	O1	**O2**	**O3**	**O4**	**O5**	**O6**	**O7**	**O8**	**O9**	**O10**	**O11**	**O12**
P	0	0	**0**	1	1	0	0	0	0	1	0	0
L	0	1	**1**	0	0	1	0	0	1	0	1	0
M	1	0	**0**	0	0	0	1	1	0	0	0	1
total	1	1	**1**	1	1	1	1	1	1	1	1	1

Tabela 2.5 – Partição *fuzzy* resultante do agrupamento

	O1	**O2**	**O3**	**O4**	**O5**	**O6**	**O7**	**O8**	**O9**	**O10**	**O11**	**O12**
P	0.05	0	**0.06**	0.93	0.92	0	0	0	0.09	0.75	0.10	0.1
L	0	0.97	**0.85**	0	0	0.99	0.21	0.19	0.82	0.13	0.80	0.25
M	0.95	0.03	**0.09**	0.07	0.08	0.01	0.79	0.81	0.09	0.12	0.10	0.65
total	1	1	**1**	1	1	1	1	1	1	1	1	1

Entre os algoritmos existentes para a realização de agrupamento *fuzzy*, encontram-se o *Fuzzy c-Means* (FCM), originalmente proposto por Bezdek (1981) e o *Possibilistic Fuzzy C-Means* , desenvolvido por Pal et al. (2005). Ambos os algoritmos foram utilizados neste trabalho.

2.3.3 *Fuzzy C-Means* (FCM)

O algoritmo *Fuzzy C-Means* é uma extensão do Algoritmo K-Means (MACQUEEN et al., 1967) e tem como objetivo encontrar grupos *fuzzy* para um conjunto de dados minimizando uma função relativa às distâncias entre os dados e os centros dos grupos. Esses mesmos dados pertencem a um determinado grupo com algum grau de pertinência $u_{ij} \in [0, 1]$. A função objetivo do algoritmo J_{FCM} é responsável por encontrar a dissimilaridade intra-grupo. Quanto maior o valor de J_{FCM} maior será o grau de pertinência de um objeto para com o grupo mais próximo e graus de pertinência menores estão relacionados a grupos mais distantes.

$$J_{FCM} = \sum_{j=1}^{N} \sum_{i=1}^{k} (u_{ij})^m D_{ij} \tag{2.4}$$

O cálculo dos valores dos centros dos grupos se dá pela média ponderada entre X_j e as pertinências de C_i. Dada uma partição *fuzzy* $U = [u_{ij}]_{NXk}$, os protótipos (centros dos grupos) do i-ésimo grupo é calculado da seguinte forma:

$$v_i = \frac{\sum_{j=1}^{N} (u_{ij})^m X_j}{\sum_{j=1}^{N} (u_{ij})^m} \tag{2.5}$$

Existem diferentes medidas de distância que podem ser utilizadas no processo de agrupamento de dados, dependendo da base de dados que se deseja separar em grupos. Uma das mais conhecidas medidas de similaridade é a distância Euclidiana (também utilizado desse projeto) apresentada em (2.6)

$$D_{ij} = \sqrt{\sum_{j=1}^{N} \sum_{i=1}^{k} (X_j - v_i)^2} \tag{2.6}$$

Durante o processo do FCM de minimização da equação 2.4, os valores da matriz U é calculado e atualizado U^{T+1}, onde $1 \leq i \leq k, 1 \leq j \leq N$

$$u_{ij} = \left(\sum_{c=1}^{k} \left(\frac{D_{ij}}{D_{cj}} \right)^{\frac{1}{(m-1)}} \right)^{-1} \tag{2.7}$$

De uma forma geral, o processo de agrupamento do FCM é representado de maneira simples no algoritmo 14. Todos os parâmetros utilizados antes da linha de repetição são fornecidos pelo usuário e mantidos sem nenhuma modificação durante todo o processo de

agrupamento.

Algoritmo 1: FUZZY C-MEANS (FCM)

Entrada: m, X, ε

Saída: Partição *fuzzy* U, Matriz de Centróides V

início
- Determine o valor do parâmetro de fuzzificação m
- Determine a quantidade de partições *fuzzy* k
- Determine um valor pequeno e positivo para o erro máximo ε
- Inicialize a matriz de protótipos V aleatoriamente
- Inicialize o contador de iterações t como $t = 0$
- Inicializar a Matriz de pertinência U
- **repita**
  - $t++$
  - Atualiza a Matriz de centróides V usando 2.5
  - Atualizar a partição U usando 2.7
- **até** $\left\|U^{(t+1)} - U^{(t)}\right\| < \varepsilon$
- ;

fim

2.3.4 **Possibilistic Fuzzy C-Means** (PFCM)

O algoritmo PFCM é uma extensão do algoritmo *Possibilistic C-Means* (PCM) e pode ser considerado um melhoramento do algoritmo *Fuzzy Possibilistic C-Means*. Para o cálculo da função objetivo, utiliza-se a matriz de partição probabilística U e uma matriz de partição possibilística P. O Cálculo da distância D_{ij} pode ser qualquer medida de similaridade como apresentado em (2.6) e (**??**).

$$J = \sum_{i=1}^{k}\sum_{j=1}^{N}((\mu_{ij})^m + (p_i)^\gamma)D_{ij} \tag{2.8}$$

O algoritmo PFCM buscar relaxar a restrição 2.9 imposta no FPCM para valores de tipicidade.

$$\sum_{j=1}^{N} p_{ij} = 1 \tag{2.9}$$

No PFCM a nova função de minimização dada pela equação 2.10.

$$J = \sum_{i=1}^{k}\sum_{j=1}^{N}(a.(u_{ij})^m + b.(p_{ij})^\gamma)D_{ij} + \sum_{i=1}^{k}\eta_i\sum_{j=1}^{N}(1 - p_{ij})^\gamma \tag{2.10}$$

Os valores de fuzzificação são representados por $m > 0$ e $\gamma > 0$. $U = [u_{ij}]$ e $P = [p_{ij}]$ são referentes as partições probabilísticas (pertinências) e possibilísticas(tipicidades) respectivamente. As constantes $a > 0$ e $b > 0$ representam a importância que se quer dar na função de minimização. η_i é uma constante possibilística definida pelo usuário, determinando a importância do segundo termo da equação (2.10), por exemplo, se η_i for muito alto, valores de u_{ij}, serão tão maiores o quanto possíveis. O cálculo de η_i é o mesmo para o algoritmo PCM em (2.11), onde $1 \leq i \leq k$ e os valores de U são calculados em (2.12) assim como no algoritmo FCM, no qual $1 \leq i \leq k$ e $1 \leq j \leq N$.

$$\eta_i = \frac{\sum_{j=1}^{N}(u_{ij})^m D_{ij}}{\sum_{j=1}^{N}(u_{ij})^m} \tag{2.11}$$

$$u_{ij} = \left(\sum_{c=1}^{k} \left(\frac{D_{ij}}{D_{cj}} \right)^{\frac{1}{m-1}} \right)^{-1} \tag{2.12}$$

Os valores dos centros de grupos são calculados através da equação (2.13) e a atualização da matriz de tipicidade P que ocorre durante o processo de minimização da função objetivo é calculada em (2.14).

$$v_i = \frac{\sum_{j=1}^{N}(a.(u_{ij})^m + b.(p_{ij})^\gamma)x_j}{\sum_{j=1}^{N}(a.(u_{ij})^m + b.(p_{ij})^\gamma)} \tag{2.13}$$

$$p_{ij} = \left(1 + \left(\frac{b}{\eta_i} D_{ij} \right)^{\frac{1}{(\gamma-1)}} \right)^{-1} \tag{2.14}$$

De uma forma geral, o processo de agrupamento do FCM é representado de maneira simples no algoritmo 11. Todos os parâmetros utilizados antes da linha de repetição são fornecidos pelo usuário e mantidos sem nenhuma modificação durante todo o processo de

agrupamento.

Algoritmo 2: Possibilistic Fuzzy C-Means (PFCM)

Entrada: Todos os parâmetros do algoritmo FCM, γ, η_i, partição possibilística $P = [p_{ij}]$, importâncias a e $b > 0$

Saída: Partição U, partição P, centróides V

```
início
    t = 0
    repita
        Calcule os protótipos dos grupos (2.13)
        Calcule as distâncias (2.6)
        Atualize a partição possibilística P
        (2.14)
        Atualize a partição probabilística U (2.7)
        t++
    até ||U^(t+1) − U^(t)|| < ε;
fim
```

3

Redução de Reticulados de Conceitos por Agrupamento de Dados

Neste capítulo serão apresentadas três técnicas de redução de reticulado de conceitos encontradas na literatura, bem como a explanação de cada uma delas.

3.1 Técnicas de redução de reticulados de conceitos

O FCA pode desenvolver uma estrutura potencialmente complexa mesmo a partir de uma base de dados consideravelmente pequena. Dessa forma, fez-se necessário desenvolver uma maneira de reduzir esse problema. Dias e Vieira (2015) classificou as técnicas de redução de reticulado de conceitos em 3 categorias: remoção de informação redundante, simplificação e seleção.

3.1.1 Remoção de informação redundante

Existem aplicações onde deve-se manter o número de conceitos gerados pela Análise de Conceitos formais e consequentemente mantém o reticulado de conceitos com o alto nível de complexidade. Por exemplo, em um contexto formal têm-se objetos que possuem os mesmos atributos que outro objeto são redundantes e mesmo assim eles podem ser usados para a contagem do número de objetos da extensão de um conceito. Caso esse objetos e atributos redundantes sejam removidos, a contagem da extensão será modificada, prejudicando aplicações que precisam desse número. Vale ressaltar que para utilizar a técnica de remoção de informação redundante é necessário ter conhecimento prévio do domínio no qual se pretende trabalhar.

Aplicações que não precisam manter o número de objetos e atributos, pode-se então aplicar alguma técnica de redução. Uma das formas de redução é eliminar a informação redundante. Dado um objeto $g \in G$, um atributo $m \in M$ ou a incidência

$$i \in I$$

, é considerado informação redundante de conceitos quando removendo um desses dados o resultado final é um reticulado isomórfico ao original. Isso significa que se diminuir o número de objetos ou atributos, é possível manter a estrutura do reticulado original.

Uma das formas realizar a redução de redundância, e talvez a maneira mais simples é fazer a substituição de um conjunto de objetos que possuem os mesmos atributos por um objeto mais representativo. Também é possível substituir um conjunto de atributos que aparecem exatamente com os mesmos objetos por um atributo mais significativo. Dado um contexto formal 3.1 como exemplo:

Tabela 3.1 – Exemplo de Contexto Formal

	a	**b**	**c**	**d**	**e**	**f**
1	1	1	0	0	0	0
2	1	0	0	0	0	0
3	1	1	1	1	0	1
4	0	1	0	1	1	1
5	1	0	0	0	0	0

Os objetos 2 e 5 são iguais e os atributos d e f são iguais. Os objetos e atributos iguais são unidos e no reticulado de conceitos aparecem como um só objeto e um só atributo.

Outra forma de eliminar informação redundante é eliminar um atributo que pode ser descrito por um conjunto de atributos. Supondo que $m \in M$, $X \subseteq M$ e $m \notin X$, se $m' = X'$, o atributo c pode ser descrito pelos atributos a e d. Consequentemente, caso o atributo c for eliminado, o reticulado resultante do contexto formal no qual a coluna referente a c é eliminada é isomorfo ao contexto original.

A junção de objetos ou de atributos é baseada na estrutura do reticulado de conceitos e a partir disto, outras técnicas de redução foram propostas.

Em Liu e Mi (2008) é possível determinar famílias de atributos comuns aos objetos. Realizando a combinação de alguns elementos dessas famílias é capaz de produzir um reticulado de conceitos reduzido. Os resultados obtido em (LIU; MI, 2008) foram iguais a (ZHANG; WEI; QI, 2005), porém diferentemente de Liu e Mi (2008), Zhang, Wei e Qi (2005) utilizou uma matriz de discriminação para diminuir o custo computacional da redução do reticulado de conceitos.

Pei, Li e Mi (2011) propôs uma técnica de redução similar visando manter uma estrutura isomorfa ao original em uma extensão da ACF. Os autores utilizam a técnica de redução em contextos formais de decisão *fuzzy*.

3.1.2 Simplificação

A técnica de simplificação baseia-se em abstrair diferenças não-essenciais seguindo algum critério, entre conceitos, objetos ou atributos. Nessa técnica de redução, ter conhecimento prévio ou não do domínio é opcional, dependendo do algoritmo que se pretende utilizar.

Um dos métodos de simplificação utilizado é o agrupamento. O agrupamento pode ser realizado no conjunto de atributos, objetos ou conceitos formais. Uma das formas de agrupamento é utilizar a técnica de *Singular Valeu Decomposition* (SVD) buscando reduzir a dimensionalidade do Contexto Formal. A técnica SVD é uma de muitas técnicas de Decomposição de Matriz da álgebra linear para transformar uma matriz de alta dimensionalidade para uma de baixa dimensionalidade. A SVD consiste em construir classes de objetos equivalentes. Aplica-se a relação de equivalência nas matrizes reduzidas geradas pela SVD. Um objeto será equivalente a outra se e somente se a distância entre eles for maior que um limiar. Se os objetos forem equivalentes então ocorre a junção dos mesmos.

Cheung e Vogel (2005) realiza o agrupamento no conjunto de objetos utilizando SVD para construir classes de objetos equivalentes. Nesse caso, cada documento é tratado com um objeto e serem equivalentes se e somente se o cosseno dos ângulos entre eles for superior a um limiar preestabelecido.

A tabela 3.2 mostra um contexto formal antes da aplicação da técnica SVD para a redução do reticulado de conceitos e a tabela 3.3 mostra o resultado do agrupamento dos objetos equivalentes do contexto formal original.

Tabela 3.2 – Contexto Formal da matriz Termo-Documento antes da aplicação de SVD extraído de (CHEUNG; VOGEL, 2005)

Termo/Doc	D1	D2	D3	D4	D5	D6	D7
Baby(y,ies,y's)		x			x		x
Child(ren's)		x		x			
Guide						x	x
Health							
Home		x		x			
Infant	x						
Proofing					x	x	
Safety				x			
Toddler	x						

Tabela 3.3 – Contexto Formal da matriz Termo-Documento depois da aplicação de SVD extraído de (CHEUNG; VOGEL, 2005)

Termo/Doc	(D1,D2,D3,D4)'	D5	D6	D7
Baby(y,ies,y's)	x	x		x
Child(ren's)	x			
Guide			x	x
Health	x			
Home	x			
Infant	x			
Proofing		x	x	
Safety	x			
Toddler	x			

Um dos problemas em se utilizar métodos algébricos para a redução nos contextos formais é o alto custo computacional (KUMAR; SRINIVAS, 2010). Uma possível solução é utilizar outros tipos de algoritmos de agrupamento. Kumar e Srinivas (2010), utilizou o agrupamento *fuzzy* de dados. O algoritmo escolhido para os experimentos foi o*Fuzzy C-Means*. Com o agrupamento *fuzzy*, um objeto pertence a mais de um grupo com algum grau de pertinência. O autor defende o uso da teoria de conjuntos *fuzzy* para deixar o agrupamento mais flexível e próximo do pensamento humano, além de obter resultados melhores se comparado a algoritmos de agrupamento clássicos. Diferente de (CHEUNG; VOGEL, 2005), as dimensões do contexto formal não mudam, porém os relacionamentos entre os objetos e atributos podem mudar.

A tabela 3.4 apresentam os resultados obtidos pelo autor utilizando esse tipo de agrupamento, comparando a técnica proposta por Cheung e Vogel (2005). Os resultados dos experimentos mostraram que utilizando o algoritmo *Fuzzy C-Means* o número de conceitos e arestas gerados foram menores que os números obtidos com o SVD.

Tabela 3.4 – Comparação dos resultados das técnicas de agrupamento *Fuzzy C-Means* e SVD (KUMAR; SRINIVAS, 2010)

	Conceitos	**Arestas**	**Altura**
Contexto Original	55	115	11
FCM k=9	32	56	10
FCM k=5	29	32	7
SVD k=9	54	112	10
SVD k=5	24	41	7

Diferente de (CHEUNG; VOGEL, 2005) e (KUMAR; SRINIVAS, 2010) que não precisam ter conhecimento prévio do domínio de problema, Dias e Vieira (2013) incorpora conhecimento adicional ao contexto formal. Nesse caso, substitui-se grupos de objetos similares por objetos que os representem, com base em uma avaliação qualitativa de seus atributos. Cada atributo do contexto formal é associado a um peso (variando entre 0 e

1), onde esse peso pode ser representado como a relevância do atributo (peso igual a 0 significa nenhuma relevância e peso igual a 1 significa relevância máxima). A similaridade entre os objetos é a soma ponderada dos pesos dos atributos que ambos compactuam (possuem ou não possuem).

3.1.3 Seleção

Em reticulados de conceitos com alto nível de complexidade podem existir conceitos considerados irrelevantes, dependendo da aplicação. A relevância pode estar relacionada a cardinalidade da intenção ou extensão do conceito, por exemplo.

Por definição, a técnica de seleção seleciona a partir de um contexto formal ou reticulado de conceitos, um subconjunto de atributos, objetos ou conceitos que satisfazem um conjunto de restrições. Nesse caso, é importante e necessário ter um conhecimento prévio do conjunto de atributos e conjunto de objetos, pois auxilia no processo de redução. Por exemplo, se o conjunto de atributos for bem conhecido, é possível atribuir pesos aos atributos que possuem maior relevância.

Belohlavek e Macko (2011) atribui um peso a cada atributo para expressar sua relevância e, em seguida, seleciona conceitos formais considerados relevantes. A importância de uma conceito é medido pela soma dos pesos de seus atributos (referente a intenção do conceito) dividida pela cardinalidade de sua intenção como Zhang et al. (2012) fez em seu trabalho. Tendo o contexto formal 3.1 como exemplo e supondo que os atributos agora possuem pesos $a = 1$, $b = 1$, $c = 0$, $d = 0.5$, $e = 0$ e $f = 0.5$. Utilizando um gerador mínimo (BELOHLAVEK; MACKO, 2011) e um limiar de 0.75, os conceitos formais relevantes foram:

- $C_1 = (\{1, 2, 3, 5\}\,\{a\})$
- $C_2 = (\{1, 3, 4\}\,\{b\})$
- $C_3 = (\{1, 3\}\,\{a, b\})$
- $C_4 = (\{3\}\,\{a, b, c, d, f\})$

Um problema observado ao se utilizar a técnica de seleção é a necessidade de calcular os conceitos formais antes de verificar se sua relevância está acima do limite estabelecido. Em alguns casos, onde o número de objetos ou de atributos são altos resultando em um contexto formal grande é inexecutável a geração de todos os conceitos formais.

Outro problema também identificado é o alto custo computacional quando se utiliza geradores e geradores mínimos, pois é necessário identificar entre todos os conceitos os quais atingiram o limiar estabelecido pelo usuário.

3.1.4 Análise das técnicas de redução

Em linhas gerais, as técnicas de eliminação de informação redundante produzem reticulados isomorfos aos originas, além de ser necessário ter conhecimento prévio do domínio do problema. A exigência do isomorfismo limita o nível de redução, impossibilitando o uso dessa técnica em muitas aplicações.

As técnicas de simplificação produzem simplificações do reticulado original, objetivando preservar apenas aspectos considerados relevantes, deixando opcional ter conhecimento prévio do domínio. Particularmente, técnicas de simplificação diminuem o espaço de conceitos e consequentemente possuem maior número de aplicações em casos onde é necessário reduções extremas. Outro ponto a ser observado é que como as técnicas são aplicadas no contexto formal, elas possuem uma complexidade inferior as outras técnicas classificadas nesse capítulo, tornando mais viável a manipulação do conhecimento representável por contextos grandes e que levam a reticulados complexos.

Por fim, as técnicas de seleção, selecionam atributos, objetos ou conceitos seguindo algum critério. A desvantagem dessas técnicas é quem podem ocorrer perdas de conceitos importantes. Outra desvantagem é o alto custo computacional, pois todo o espaço de conceitos é explorado. Além disso, para se utilizar essas técnicas é a necessidade de se ter o conhecimento prévio do domínio a ser trabalhado.

4

Redução de Reticulado de Conceitos usando Agrupamento Fuzzy

Este capítulo apresenta uma técnica de redução baseada em algoritmos de agrupamento fuzzy capaz de gerar um reticulado de conceitos simplificado.

4.1 Proposta

O número de conceitos gerados utilizando o ACF pode ser elevado até mesmo para uma base de dados pequena. Uma das formas de redução do reticulado de conceitos é aplicar alguma técnica de redução que foi apresentada no Capítulo anterior.

As técnicas de simplificação se mostram mais flexíveis se comparadas com a demais técnicas, pois além de não ser necessário ter o conhecimento prévio do domínio do problema, não possui limite de redução e não explora todo o espaço de conceitos. Outro ponto a ser destacado são as diferentes formas e algoritmos de agrupamento disponíveis na literatura. Por isso, no trabalho de pesquisa desenvolvido e descrito nesse documento optou-se utilizar uma das técnicas de simplificação, no caso algoritmos de agrupamento.

Assim como Kumar e Srinivas (2010), o presente projeto apresenta uma proposta de redução do contexto formal utilizando agrupamento. Diferentemente do trabalho de Kumar e Srinivas (2010), essa pesquisa utilizou dois algoritmos de agrupamento fuzzy em diferentes bases de dados, com o número de objetos e atributos altos, enquanto que os autores realizaram experimentos em apenas uma base de dados com número de objetos e atributos menores. Outra diferença a ser destacada é a utilização de diferentes números de grupos para realizar o agrupamento e a utilização de métricas de validação dos resultados, uma vez que em (KUMAR; SRINIVAS, 2010) foram utilizados apenas dois valores de grupos sem nenhuma forma de validar os resultados obtidos no agrupamento.

O trabalho aqui proposto divide-se em cinco etapas, onde cada etapa é dependente da anterior e foram denominadas da seguinte maneira:

1. Transformação do conjunto de dados em um Contexto Formal
2. Agrupamento do contexto formal
   a) Algoritmos de agrupamento FCM e PFCM
   b) Validação do agrupamento
3. Cálculo de Decomposição de matriz utilizando matriz de centróides
4. Transformação para valores binários
5. Aplicação da ACF no contexto formal reduzido
   a) Geração do reticulado de conceitos reduzido

Na Figura 4.1, cada retângulo de cor preta é uma etapa do projeto. Os retângulos pontilhados de cor azul é uma sub-etapa que ocorre dentro do retângulo preto. As setas ilustram a dependência entre as etapas e as setas pontilhadas ilustram a dependência entre as sub-etapas. As etapas e sub-etapas serão apresentadas nas próximas sessões.

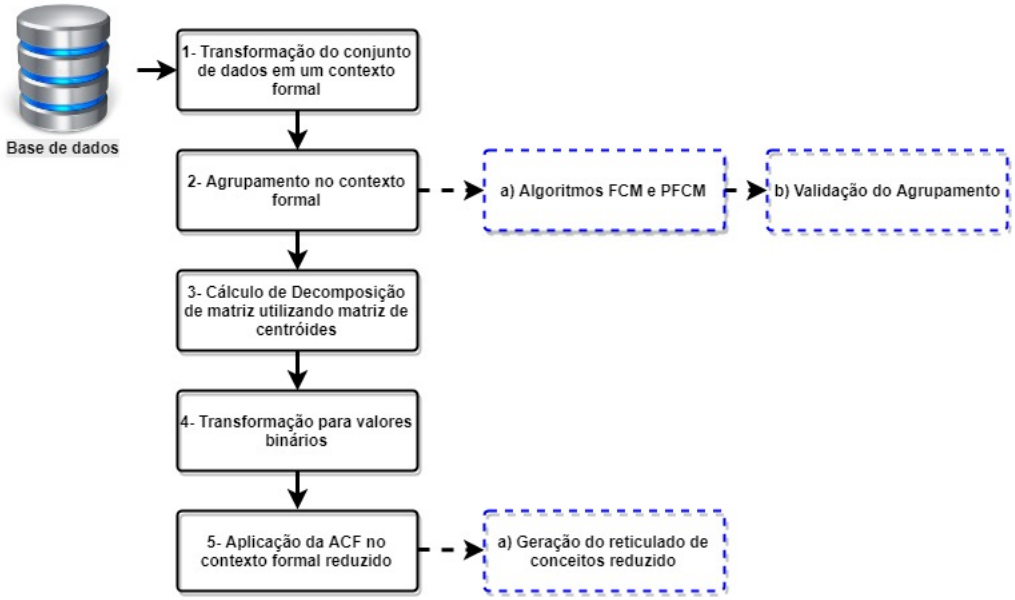

Figura 4.1 – Ilustração das etapas do projeto proposto

4.1.1 Transformação do conjunto de dados em um Contexto Formal

No primeiro passo do desenvolvimento transforma-se a base de dados de entrada para a ACF para o formato de contexto formal. Os dados devem estar organizados em forma de matriz, onde cada linha do contexto representa um objeto e cada coluna do contexto representa um atributo. Para exemplificar o processo de transformação da base de dados em um contexto formal, a Figura 4.2 mostra o trecho de uma base de dados original referente a documentos e termos.

```
@ATTRIBUTE fuzzy_sets   NUMERIC
@ATTRIBUTE computer NUMERIC
@ATTRIBUTE knowledge_data   NUMERIC
@ATTRIBUTE university   NUMERIC
@ATTRIBUTE triple   NUMERIC
@ATTRIBUTE triple_gmi   NUMERIC
@ATTRIBUTE degree   NUMERIC
@ATTRIBUTE membership   NUMERIC
@ATTRIBUTE references   NUMERIC
@ATTRIBUTE formal_context   NUMERIC
@ATTRIBUTE staab    NUMERIC
@ATTRIBUTE set_attributes   NUMERIC
@ATTRIBUTE number   NUMERIC
@ATTRIBUTE systems  NUMERIC
@ATTRIBUTE gmi  NUMERIC
@ATTRIBUTE terms    NUMERIC
@ATTRIBUTE workbench    NUMERIC
@ATTRIBUTE set  NUMERIC
@ATTRIBUTE low  NUMERIC
@ATTRIBUTE real NUMERIC
@ATTRIBUTE document NUMERIC
@ATTRIBUTE international    NUMERIC
@ATTRIBUTE intensional  NUMERIC
@DATA
0.0,0.0,0.0,0.03,0.0,0.0,0.12,0.0,0.0,0.0,0.0,0.0,0.01,0.07,0.0,0.1,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.69,0.15,0.0,0.01,0.0,0.0,0.0,0.0,0.0,0.09,0.0,0.12
0.0,0.02,0.5,0.0,0.0,0.0,0.06,0.02,0.0,0.0,0.09,0.0,0.06,0.06,0.0,0.18,0.0,0.04,0.0,0.0,0.0,0.0,0.0,0.56,0.0,0.01,0.0,0.0,0.0,0.0,0.08,0.04,0.45
0.0,0.26,0.0,0.0,0.0,0.0,0.08,0.01,0.04,0.0,0.05,0.0,0.1,0.14,0.0,0.29,0.38,0.07,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.05,0.0,0.0,0.24,0.0,0.02,0.06,0.0
0.0,0.06,0.17,0.16,0.0,0.0,0.41,0.02,0.0,0.0,0.0,0.0,0.0,0.03,0.04,0.0,0.2,0.0,0.04,0.0,0.0,0.0,0.0,0.0,0.0,0.81,0.0,0.0,0.0,0.0,0.0,0.0,0.12,0.27,0.0,0
0.0,0.15,0.0,0.08,0.0,0.0,0.0,0.12,0.0,0.0,0.05,0.0,0.0,0.21,0.0,0.21,0.0,0.02,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.01,0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.0,
0.0,0.0,0.0,0.0,0.0,0.0,0.01,0.01,0.04,0.0,0.0,0.15,0.04,0.11,0.0,0.32,0.0,0.06,0.06,0.04,0.0,0.0,0.0,0.0,0.05,0.0,0.0,0.07,0.0,0.0,0.0,0.0,0.12,0.0,0.2
```

Figura 4.2 – Exemplo de base de dados de Documentos e Termos

Na tabela (4.1) é possível observar o mesmo trecho da Figura 4.2 transformado em um contexto formal. Cada documento é referente a um objeto e cada termo é referente a um atributo.

Tabela 4.1 – Trecho do contexto formal sobre documentos e termos

	fuzzy-Sets	computer	knowlegde_data	university	triple	triple_gmi	...	degrees
doc1	0.0	0.0	0.0	0.03	0.0	0.0	...	0.0
doc2	0.0	0.02	0.5	0.0	0.0	0.0	...	0.21
doc3	0.0	0.26	0.0	0.0	0.0	0.0	...	0.0
...	...	...	...	...	...	...	...	...
doc40	0.0	0.0	0.0	0.03	0.0		...	0.04

4.1.2 Agrupamento do contexto formal

O contexto formal é a entrada para os algoritmos de agrupamento, então a aplicação do agrupamento nessa etapa é muito importante, pois é nessa etapa que se inicia o processo de redução do reticulado de conceitos. Os algoritmos de agrupamento escolhidos foram o agrupamento *Fuzzy C-Means* (FCM) e o *Possibilistic Fuzzy C-Means*

4.1.2.1 Algoritmos FCM e PFCM

Ambos os algoritmos têm como objetivo encontrar grupos em um conjunto de dados buscando agrupar dados similares. A principal diferença entre esses algoritmos se encontra na flexibilidade entre as pertinências de um objeto para com os grupos.A soma das pertinências de um objeto para com os grupos não necessitam somar 1 no PFCM, enquanto que no FCM, a soma das pertinências obrigatoriamente devem ser iguais a 1, fazendo com o que o algoritmo seja mais sensível a ruídos. Em contraponto, o algoritmo

FCM é mais simples que o PFCM, pois os cálculos de atualização da matriz de pertinência e de centróides são menos complexos, além de não ser tão sensível aos parâmetros inicias. O principal desafio nas tarefas de agrupamento é a análise e definição do número ideal de grupos. Esse problema pode ser abordado usando métricas de avaliação de agrupamento.

4.1.2.2 Validação do agrupamento

Para a avaliação dos agrupamentos *fuzzy*, foram usadas as medidas de Silhueta *Fuzzy* (CAMPELLO; HRUSCHKA, 1975), *Partition Coefficient* (BEZDEK, 1974), *Partition Entropy* (BEZDEK, 1975), *Xie-Beni* (XIE; BENI, 1991). As métricas de avaliação são utilizadas para determinar o número de grupos e a partição mais adequada dentre as encontradas para o conjunto de dados desejado. O índice de *Partition Entropy* (PE) possui valores no intervalo $[0, \log_a k]$, no qual é uma função do tipo minimizadora, ou seja, o valor encontrado de PE tende para 0 quando se tem grupos bem definidos para o conjunto de dados X. Caso os valores encontrados forem próximos a 1, indica-se a ausência de estruturas de grupos no conjunto de dados ou a incapacidade do algoritmo de obtê-las. O PE é calculado da seguinte forma:

$$PE(U) = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{k}\mu_{ij}\log_a \mu_{ij}, 1 < a < \infty \tag{4.1}$$

Onde k indica o número de grupos, n é o número de elementos da base X e μ_{ji} são as pertinências da partição *fuzzy* U resultante do agrupamento *fuzzy*.

O *Partition Coefficient* (PC) também identifica grupos bem definidos como no PE. Os valores de PC variam no intervalo $\left[\frac{1}{k}, 1\right]$, onde valores próximos de 1 indicam grupos bem definidos. Valores próximos de $\frac{1}{k}$, indica a ausência de grupos. A equação 4.2 mostra o cálculo de PC.

$$PC(U) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{k}\mu_{ij}^2 \tag{4.2}$$

O índice de *Xie-Beni* (XB) é usado para identificar a compactação e a separação dos grupos de modo a encontrar um número ótimo de grupos, definida como a razão (4.5)entre a compactação (4.3) e a separação (4.4) da partição *fuzzy*. Um valor pequeno para XB significa pouca compactação dos grupos.

$$Comp = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\mu_{ij}^m \left\|v_j - x_i\right\|^2 \tag{4.3}$$

$$Sep = min_{i\neq j}\left\|v_j - x_i\right\|^2 \tag{4.4}$$

$$XB = \frac{Comp}{Sep} \tag{4.5}$$

A Silhueta *Fuzzy* (SF) é fundamentada na similaridade entre os objetos de um grupo e na dissimilaridade destes objetos em relação ao grupo vizinho mais próximo, permitindo estimar o melhor número de k grupos. A SF é definida como:

$$SF = \frac{\sum_{i=1}^{N}(\mu_{pi} - \mu_{qi})^{\alpha} s_i}{\sum_{i=1}^{N}(\mu_{pi} - \mu_{qi})^{\alpha}} \tag{4.6}$$

α é um coeficiente de ponderação, no qual $\alpha > 0$ (valor padrão é 1). μ_{pi} e μ_{qi} é o primeiro e o segundo maior valor de pertinência da $i = ésima$ coluna da matriz de pertinência *fuzzy* respectivamente. s_i é a silhueta do objeto i definida como:

$$s_i = \frac{b_{pi} - a_{pi}}{max\left\{b_{pi}, a_{pi}\right\}} \tag{4.7}$$

$i = \{1, ..., n\}$ e $p = \{1, ..., k\}$. a_{pi} é a distância média do objeto i em relação a todos os objetos pertencentes ao mesmo grupo p e b_{pi} é a dissimilaridade do objeto i ao grupo vizinho mais próximo, ou seja, b_{pi} é o valor mínimo da distância média d_{qi} do objeto i a todos os objetos de um outro grupo q, onde $q \neq p$ e $q = \{1, ..., k\}$. Os Valores de SF encontram-se no intervalo $-1 \leq SF \geq 1$, onde valores de SF próximos de 1 significam o quão boa é a atribuição do objeto i ao grupo p.

4.1.3 Cálculo de Decomposição de matriz utilizando matriz de centróides

Para aplicar a ACF os dados devem estar organizados em um contexto formal, porém o resultado do agrupamento é uma matriz de pertinência e uma matriz de centróides. Então para se ter uma estrutura de composta por objetos (linhas), atributos(colunas) e os relacionamentos (objeto-atributo) utiliza-se a Decomposição de matriz (DM) (DHILLON; MODHA, 2001). O resultado final é uma aproximação de matriz e é calculado como mostra a equação 4.8.

$$D_k = C_k Z^* \tag{4.8}$$

Onde C_k é a matriz de centróides nxk (n é o número de objetos e k o número de grupos) escolhida seguindo uma das métricas de avaliação do agrupamento citado na etapa anterior, denominada matriz de conceitos $C_k = [v_1, v_2, ..., v_k]$. Z^* é calculado da seguinte forma:

$$Z^* = \left(C_k^t C_k\right)^{-1} C_k^t A \tag{4.9}$$

A é o contexto formal original (utilizado na etapa de agrupamento) e t é a matriz transposta de C_k.

4.1.4 Transformação para valores binários

Antes da aplicação da ACF, a matriz D_k precisa ser transformada em uma matriz binária. A transformação é realizada usando um valor de limiar, no qual valores abaixo do limiar recebem 0 e acima recebem o valor 1. Após transformar os valores dos relacionamento para binário, a matriz D_k volta a ser um contexto formal, porém diferente do contexto original, um contexto formal reduzido.

4.1.5 Aplicação da ACF no contexto formal reduzido

Após todas as etapas anteriores, essa última etapa é responsável pela aplicação a ACF no contexto formal reduzido.

4.1.5.1 Geração do reticulado de conceitos reduzido

Por fim, o resultado esperado é um reticulado menor, com número de arestas e conceitos menores se comparado ao original.

5

Experimentos e Análise dos Resultados

Esse capítulo tem como objetivo apresentar os resultados dos experimentos realizados com os algoritmos de agrupamento Fuzzy C-Means e Possibilistic Fuzzy C-Means utilizando métricas de avaliação. E apresentar de maneira quantitativa o reticulado de conceitos gerado após o agrupamento.

5.1 Experimentos

Optou-se pela linguagem de programação R [1] para a implementação dos algoritmos PFCM e o FCM, pois a linguagem oferece diferentes pacotes em problemas de aprendizado de máquina, como o pacote **e1071**[2] que possui a implementação completa do algoritmo FCM. Os índices de avaliação utilizados também se encontram disponíveis no pacote **fclust**[3] na linguagem R.

Para a geração do reticulado de conceitos foi utilizado a ferramenta **ConExp** (YEVTUSHENKO, 2000) no qual possui a funcionalidade básica necessária para o estudo e pesquisa da ACF.

Para validação da proposta, foram utilizados cinco bases de dados (Tabela 5.1) fornecidos por Rios (2013b) para avaliar os resultados obtidos pelos algoritmos de agrupamento utilizados. Todas as bases de dados passaram por um processo de extração de descritores de grupos, sendo que para esse trabalho, os descritores são os atributos e os documentos são os objetos. Os valores de relacionamento entre objeto e atributo são frequências numéricas.

[1] https://www.r-project.org/
[2] https://cran.r-project.org/web/packages/e1071/index.html
[3] https://cran.r-project.org/web/packages/fclust/index.html

Tabela 5.1 – Base de dados utilizadas nos experimentos

Bases	Objetos	Atributos
Iaarticles	40	60
Opinosis	51	60
Reuters	1052	440
Newyorktimes	18	100
Newsgroup	2000	80

A base de dados *Iaarticles* é composta por artigos do repositório *IEEE* coletada e rotulada manualmente Rios (2013a). A base de dados *Newyorktimes* foi coletada e rotulada manualmente, porém ela é composta por reportagens do site *New York Times* Rios (2013a).

A base de dados *Opinosis* é composta por comentários de clientes sobre as características de alguns produtos. Os comentários dos clientes foram obtidos nos sites: *Tripadvisor.com*, no qual consumidores comentaram sobre hotéis; *Amazon.com*, onde consumidores comentaram sobre aluguéis de carros e por fim, *Edmunds.com*, onde os consumidores falaram sobre produtos eletrônicos. A coleção *Opinosis* está disponível repositório da UCI [4].

A base de dados *Reuters* originalmente era uma coleção de documentos obtida durante o processo de desenvolvimento do sistema de categorização de documentos CONSTRUE (HAYES; WEINSTEIN, 1991) pela *Carnegie Group, Inc. and Reuters, Ltd.*

A base *Newsgroup* originalmente foi composta por coleções de documentos, possuindo 20000 documentos divididos em aproximadamente 20 categorias. A base de dados está disponível também do repositório UCI [5]e para os testes desse trabalho foi escolhido a categoria *science*.

Todas as bases de dados utilizadas foram transformadas para o formato de contexto formal. Para cada base de dados foram realizados 10 execuções dos algoritmos realizando a média aritmética para cada índice de avaliação do agrupamento. Foram escolhidos de forma empírica os valores de grupos $k = 5$ e $k = 9$ utilizados no trabalho de Kumar e Srinivas (2010) e adicionados o valores de $k = 3$, $k = 4$ e $k = 7$. O parâmetro de fuzzificação escolhido foi o mesmo para ambos os algoritmos de agrupamento, sendo $m = 1.2$ e para o algoritmo possibilístico o valor de γ também foi 1.2.

5.2 Avaliação dos agrupamentos

Realizou-se experimentos com os agrupamento *Fuzzy C-Means* e *Possibilistic Fuzzy C-Means*. As melhores matrizes de centróides escolhidas foram de acordo com o índice de

[4] http://archive.ics.uci.edu/ml/index.php
[5] https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

Silhueta *Fuzzy* (SF), onde valores de SF próximos de 1 significam que foram encontrados grupos bem separados (indica o quão distantes dois grupos estão) e compactos (mede a proximidade dos elementos de um grupo). Utilizou-se também os índices de *Xie Beni*, *Partition Entropy* e *Partition Coefficient* para avaliar o resultado dos agrupamentos. Os resultados dos índices de avaliação dos agrupamentos foram organizados por tabelas, na qual as linhas presentam um valor para o número de grupos e cada coluna representa cada índice de avaliação do agrupamento utilizado. Os resultados mostraram que em todas as bases utilizadas, os valores dos índices de avaliação ficaram distantes do ideal. Esse fato pode ser atribuído a distribuição dos objetos e atributos no espaço de exemplos, além da possibilidade da ocorrência de ruídos. Em todas as bases foram constatados sobreposição dos dados, dificultando assim a descoberta de grupos.

Observou-se na base de dados *Iaarticles* que o valor para o número de grupos considerado melhor entre todos utilizados no experimentos foi para $k = 3$ para todos os índices de avaliação com valores para $PE = 0.4804$, $PC = 0.73363$, $XB = 1.515$ e $SF = 0.30597$. Outro fato importante a se destacar é que aumentando o valor do número de grupos, os resultados dos índices foram piorando.Esse fato ocorre devido ao problema da sobreposição de dados e o algoritmo FCM é vulnerável a ruídos como mostra a tabela 5.2.

Tabela 5.2 – Média dos índices de validação de FCA para a base de dados Iaarticle

	PE	PC	XB	SF
K=3	0.4804	0.73363	1.515	0.30597
k=4	0.63082	0.62889	5.181	0.23667
k=5	0.5825	0.66599	4.475	0.21929
k=7	0.52269	0.7185	3.453	0.21239
k=9	0.51539	0.73244	2.992	0.19218

Assim como no FCM, no algoritmo PFCM o melhor resultado entre todos os valores de grupos usados foi para $k = 3$ com valores $PE = 1.089$, $PC = 0.33912$e $SF = 0.06756$. Porém os valores de XB, PE, e SF foram diferentes e discrepantes dos encontrados no FCM, significando que o algoritmo não conseguiu identificar grupos bem definidos como mostra a tabela 5.3.

Tabela 5.3 – Média dos índices de validação de PFCM para a base de dados Iaarticle

	PE	PC	XB	SF
K=3	1.089	0.33912	76.679	0.06756
k=4	1.376	0.25479	69.519	-0.00026
k=5	1.598	0.20421	65.951	-0.00529
k=7	1.933	0.1465	57.908	0.03669
k=9	2.185	0.11388	39.240	-1

Na base de dados *Opinosis*, obteve-se os melhores valores de PE, PC e XB para $k = 7$. No entanto, como foi mencionado nessa dissertação, para escolher a matriz de centróides foi utilizado o valor de SF, então analisando os valores de SF para todos os valores de k, a matriz escolhida foi para $k = 3$ como mostra a tabela 5.4.

Tabela 5.4 – Média dos índices de validação do FCM para a base de dados Opinosis

	PE	PC	XB	SF
K=3	0.2726	0.8252	2.068	0.4709
k=4	0.2966	0.82291	2.356	0.34581
k=5	0.26532	0.86441	1.314	0.32581
k=7	0.24909	0.8815	1.081	0.34683
k=9	0.25791	0.87555	1.286	0.33953

Com o algoritmo PFCM, os melhores valores dos índices foram para $k = 3$ exceto para o índice XB, onde o melhor valor, apesar de ser discrepante do ideal, foi para $k = 7$. Assim como na base de dados *Iaarticles*, observou-se a dificuldade em se identificar grupos bem separados como mostram os resultados na tabela 5.5.

Tabela 5.5 – Média dos índices de validação do PFCM para a base de dados Opinosis

	PE	**PC**	**XB**	**SF**
K=3	0.63621	1.528	175.839	0.4474
k=4	1.365	0.2604	134.822	0.16015
k=5	1.580	0.21145	114.095	0.14399
k=7	1.913	0.15253	76.651	0.23158
k=9	2.160	0.11976	85.456	-1

A Base de dados *Reuters* pode ser considerada uma base de dados grande com 1052 objetos e 440 atributos, no qual os dados estão muito próximos no espaço de distribuição dos dados. Como pode ser observado na tabela 5.6 os resultados dos índices mostram a impossibilidade de identificação de grupos.

Tabela 5.6 – Média dos índices de validação do FCM para a base de dados Reuters

	PE	**PC**	**XB**	**SF**
K=3	1.098	0.3333	73693069.263	-1
k=4	1.326	0.26816	587982330.879	-1
k=5	1.585	0.20585	6408947253.500	-1
k=7	1.856	0.1603	5237618379485	-1
k=9	2.125	0.12331	5,5627E+16	-1

Assim como no FCM, o algoritmo PFCM obteve dificuldades em identificar grupos separados como mostra a tabela 5.7. No entanto, encontrou-se grupos separados com baixo

índice de SF. Para a escolha da matriz de centróides foi escolhido para o número de grupos $k = 3$.

Tabela 5.7 – Média dos índices de validação do PFCM para a base de dados Reuters

	PE	**PC**	**XB**	**SF**
K=3	1.098	0.3334	62605.710	0.10168
k=4	1.386	0.25003	54933.699	0.06489
k=5	1.609	0.20003	36633.664	0.00145
k=7	1.945	0.1429	35357.259	-0.04253
k=9	2.197	0.1111	31556.284	-0.00692

A base de dados *Newyorktimes* é composta por 18 objetos e 100 atributos. Pode ser considerada uma base de dados pequena, e como pode ser observado na tabela 5.8 os melhores valores de PE e XB foram para os valores de $k = 9$; e para os índices PC e SF foram os valores de $k = 3$.

Tabela 5.8 – Média dos índices de validação do FCM para a base de dados Newyorktimes

	PE	PC	XB	SF
K=3	0.19432	0.90006	1.102	0.3274
k=4	0.27072	0.85084	1.613	0.25548
k=5	0.31363	0.82809	1.543	0.21463
k=7	0.19409	0.91159	0.75544	0.19891
k=9	0.16787	0.92937	0.60519	0.14202

Os resultados com o algoritmo PFCM para a base de dados *Newyorktimes* mostraram como pode ser observado na tabela 5.9 que os melhores resultados entre os valores de grupos foi para $k = 3$ para os índices PE, PC e SF.

Tabela 5.9 – Média dos índices de validação do PFCM para a base de dados Newyorktimes

	PE	PC	XB	SF
K=3	1.071	0.3513	22.162	0.07309
k=4	1.353	0.26673	13.726	-1
k=5	1.587	0.21016	15.367	-1
k=7	1.909	0.1537	12.998	-1
k=9	2.159	0.11993	11.791	-1

Por fim, os resultados tanto para o algoritmo FCM e PFCM mostraram mais uma vez a dificuldade de identificação de grupos compactos e bem separados para a base de dados *Newsgroups* como nas demais bases de dados utilizadas nos experimentos. Como pode ser verificado na tabela 5.10 que apesar dos valores serem inferiores, o melhor número de grupos entre os demais foi para $k = 3$.

Tabela 5.10 – Média dos índices de validação do FCM para a base de dados Newgroups

	PE	PC	XB	SF
K=3	0.82194	0.49566	49178.663	0.04749
k=4	1.149	0.36521	5344876.830	-0.12036
k=5	1.380	0.29857	499059603.430	0.00188
k=7	1.697	0.23439	3038855721.223	-1
k=9	1.858	0.20751	5,25546E+15	-1

No algoritmo PFCM, assim como no FCM para a base de dados *Newsgroups* os melhores valores para PE, PC e SF foram para $k = 3$ como mostra a tabela 5.11.

Tabela 5.11 – Média dos índices de validação do PFCM para a base de dados Newgroups

	PE	PC	XB	SF
K=3	1.098	0.3334	13493.681	0.05721
k=4	1.386	0.25008	10563.795	0.0027
k=5	1.609	0.20005	10451.622	0.00184
k=7	1.945	0.1429	7829.009	-0.0481
k=9	2.197	0.11113	7316.326	-0.03907

Como pode ser constatado a dificuldade de ambos os algoritmos de identificarem grupos compactos e separáveis em todas as bases de dados utilizados nesses experimentos. Os baixos valores dos índices de avaliação dos agrupamentos se deve a distribuição dos dados das bases de dados e não aos algoritmos empregados nesse trabalho de pesquisa. Como mencionado no inicio desse Capítulo, o índice de avaliação escolhido para determinar a matriz de centróides a ser utilizada no cálculo de Decomposição de matriz foi a Silhueta *Fuzzy* (SF) e em todas as bases de dados, segundo o índice de SF, o melhor número de grupos foi $k = 3$. O algoritmo FCM obteve os melhores valores de SF nas bases de dados *Iaarticles*, *Opinosis* e *Newyorktimes* enquanto o algoritmos PFCM obteve os melhores resultados de SF nas bases de dados *Reuters* e *Newsgroups*, onde o número de objetos passaram de 1000. Apesar da diferença dos resultados do índice SF entre as bases de dados, o algoritmo PFCM conseguiu se adaptar melhor quando se tem um número de objetos alto, isso se deve ao fato da flexibilidade das pertinências, porém é um algoritmo sensível aos parâmetros iniciais, resultando em resultados piores como foi no caso das bases de dados onde o FCM foi melhor.

5.3 Geração do reticulado de conceitos

Logo após a escolha das matrizes de centróides e cálculo de Decomposição de matriz, transformou-se os valores resultantes em valores binários utilizando um limiar. O valor do limiar foi definido utilizando a média aritmética entre os valores resultantes da

decomposição de matriz, pois a escolha de um limiar fixo para todas as bases de dados poderia causar a perda de muitos objetos para a geração do reticulado de conceitos.

A tabela 5.12 mostra o número de conceitos e arestas gerados antes da realização dos agrupamentos para cada base de dados. Como citado durante essa pesquisa, em alguns casos pode ocorrer a impossibilidade da geração do reticulado devido ao alto consumo de recursos computacionais e o mesmo foi constatado na base de dados *Reuters* e *Newsgroups*. Foi determinado então a escolha do número de objetos e atributos aceitáveis pela ferramenta utilizada para a aplicação da ACF e não para a realização do agrupamento (para os agrupamentos foram mantidos os números de objetos e atributos originais). Foram escolhidos os primeiros 452 objetos e 120 atributos para a base de dados *Reuters*; e os primeiros 400 objetos e 80 atributos para a base de dados *Newsgroups*.

Tabela 5.12 – Número de conceitos e arestas do reticulado de gerados antes do agrupamento *fuzzy*

Bases	Conceitos	Arestas
Iaarticles	23172	114168
Opinosis	268	676
Reuters	237160	1122515
Newyorktimes	160	436
Newsgroup	291872	1585123

As tabelas 5.13 e 5.14 referem-se ao número de conceitos e arestas do reticulado gerado após a aplicação do algoritmo FCM e PFCM respectivamente. Como pode ser observado na 5.13, não foi gerado o reticulado de conceitos, pois não foi identificado grupos separáveis e compactos na base de dados *Reuters* segundo a medida de Silhueta *Fuzzy* (SF).

Tabela 5.13 – Número de conceitos e arestas dos reticulados gerados após a aplicação do algoritmo FCM

Bases	Conceitos	Arestas
Iaarticles	1 247	3 974
Opinosis	140	287
Reuters	-	-
Newyorktimes	152	378
Newsgroup	36 436	163 116

Tabela 5.14 – Número de conceitos e arestas dos reticulados gerados após a aplicação do algoritmo PFCM

Bases	Conceitos	Arestas
Iaarticles	1 400	4 594
Opinosis	211	498
Reuters	13 228	52 251
Newyorktimes	99	214
Newsgroup	36 436	163 116

É possível observar a redução do número de conceitos e arestas após utilizando o agrupamento no contexto formal antes da geração do reticulado de conceitos analisando o número de arestas e conceitos.Observado o número de conceitos e arestas para a base de dados *Newsgroup* foram iguais em ambos os algoritmos de agrupamento, porém não significa que foi gerado um reticulado de conceitos idêntico. Durante o processo de geração dos reticulados, tentou-se gerar também graficamente o reticulado de conceitos, porém sem êxito. Apesar do número de conceitos e arestas estarem menores que o original, a ferramenta ConExp não conseguia mostrar de forma clara os conceitos e suas relações, e em alguns caso não conseguia gerar o reticulado graficamente. Uma análise visual e qualitativa poderia mostrar se ocorreu a diferença entre o relacionamento entre os conceitos ou se realmente foram gerados reticulados iguais (como no caso na base de dados *Newsgroup*), porém o escopo do trabalho limitou-se apenas a uma análise quantitativa dos resultados.

6

Conclusão

Esse capítulo apresenta as considerações finais do trabalho apresentado, bem como apontar os próximos trabalhos.

6.1 Considerações finais

A análise de conceitos formais (ACF) é um formalismo muito utilizado para a extração, representação e análise do conhecimentos utilizando uma estrutura de reticulado. Um problema importante identificado é a complexidade e a demanda de muitos recursos computacionais para a geração de um reticulado conceitual, no qual uma estrutura altamente complexa pode ser gerada utilizando até mesmo uma base de dados pequena. Alguns trabalhos na literatura tentam tratar o problema de completude do reticulado de conceitos de diferentes formas, onde notou-se uma técnica simples e eficiente denominada simplificação.

A pesquisa apresentada nessa dissertação foi baseada no trabalho de Kumar e Srinivas (2010), onde os autores utilizaram o agrupamento *fuzzy* para realizar a redução do reticulado de conceitos mostrando que a estrutura reduzida é homomórfica ao original. Porém, os trabalhos de Kumar e Srinivas (2010) e este projeto diferem-se ao se utilizar dois algoritmos de agrupamento *fuzzy*, variação no número de grupos e a validação dos agrupamentos utilizando medidas de avaliação.

Apresentou-se dois algoritmos de agrupamento *fuzzy*, o *Fuzzy C-Mean* (FCM) e o *PossibilisticFuzzy C-Means* (PFCM), onde o algoritmo FCM possui características probabilísticas e PFCM possui características probabilísticas e possibilísticas. Também apresentou-se algumas medidas de avaliação de agrupamento *fuzzy* disponíveis na literatura, como o índice de *Xie e Beni*, *Partition Entropy*, *Partition Coefficient* e a Silhueta *Fuzzy* utilizados para medir a separabilidade e compacidade dos grupos. As medidas de avaliação foram empregadas tendo como hipótese que o melhor número de grupos levaria ao melhor agrupamento.

Os experimentos mostraram a dificuldade dos algoritmos de identificarem grupos compactos e separados em bases de dados onde ocorrem sobreposição de dados e a má distribuição desses dados. Observou-se também que para essas bases de dados, o melhor número de grupos foi igual a 3 em todos os casos, considerando apenas o índice de Silhueta *Fuzzy*, exceto para a base de dados *Reuters*, no qual com o algoritmo FCM não foi encontrado nenhuma estrutura de grupo.

Os agrupamentos resultaram em reticulados reduzidos, levando em consideração apenas o número de conceitos e arestas do reticulado, porém observou-se a necessidade não somente de uma avaliação quantitativa, mas também uma avaliação qualitativa, pois sabe-se que qualquer tipo de simplificação causa perda de informação. Alguns trabalhos (incluindo o trabalho em que essa pesquisa se baseou) provaram que a estrutura do reticulado reduzido mantém-se homomórfica ao reticulado original, porém observou-se a necessidade de um especialista para analisar a qualidade do reticulado reduzido.

6.2 Trabalhos Futuros

Como trabalho futuro pretende-se realizar uma análise qualitativa dos reticulados reduzidos, pois como já mencionado, é de fundamental importância analisar a qualidade desses reticulados. Realizando a análise da qualidade dos reticulados de conceitos, pretende-se gerar automaticamente ontologias à partir desses reticulados. Outro ponto a ser destacado como trabalho futuro é a investigação do uso de paralelismo na condução dos experimentos, principalmente na geração do reticulado. Com um número de objetos e atributos elevado, como foi possível constatar, ocorreram dificuldades e até mesmo a impossibilidade de realizar experimentos e gerações de reticulados de forma gráfica em algumas bases de dados utilizadas nesse trabalho.

Por fim, pretende-se estender todo o processo descrito neste trabalho para *fuzzy*, utilizando a Análise de Conceitos Formais *Fuzzy* (ACFF) em outras bases de dados para posteriormente gerar automaticamente ontologias *fuzzy*, fazendo com que os domínios sejam tratados de formas mais reais e próximas do pensamento humano.

Referências

BAIN, M. Inductive construction of ontologies from formal concept analysis. In: SPRINGER. *Australasian Joint Conference on Artificial Intelligence*. [S.l.], 2003. p. 88–99.

BELOHLAVEK, R.; MACKO, J. Selecting important concepts using weights. In: _____. *Formal Concept Analysis: 9th International Conference*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 65–80.

BEZDEK, J. Mathematical models for systematics and taxonomy. *In Proceedings of 8th International Conference on Numerical Taxonomy*, p. 143–166, 1975.

BEZDEK, J. C. Cluster validity with fuzzy sets. *Journal of Cybernetics*, p. 58–72, 1974.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

BEZDEK, J. C.; PAL, S. K. *Fuzzy models for pattern recognition: methods that search for structures in data*. [S.l.]: IEEE press, 1992.

CAMPELLO, R.; HRUSCHKA, E. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, p. 2858 – 2875, 1975.

CHEUNG, S. K.; VOGEL, D. Complexity reduction in lattice-based information retrieval. *Information Retrieval*, v. 8, n. 2, p. 285–299, 2005.

CIMIANO, P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387306323.

DAVEY B. A.;PRIESTLEY, H. A. Introduction to lattices and order. *Cambridge University Press*, p. 310, 1990.

DHILLON, I. S.; MODHA, D. S. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 42, n. 1-2, p. 143–175, 2001.

DIAS, S. *Redução de Reticulados Conceituais*. Tese (Tese (doutorado)) — UFMG, Maio 2016.

DIAS, S. M.; VIEIRA, N. J. Applying the jbos reduction method for relevant knowledge extraction. *Expert Systems with Applications*, Elsevier, v. 40, n. 5, p. 1880–1887, 2013.

DIAS, S. M.; VIEIRA, N. J. Concept lattices reduction. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 42, n. 20, p. 7084–7097, 2015.

GANTER, B.; WILLE, R. *Formal Concept Analysis: Mathematical Foundations*. Berlin: Springer, 1999.

GODIN, R.; SAUNDERS, E.; GECSEI, J. Lattice model of browsable data spaces. *Information Sciences*, v. 40, n. 2, p. 89–116, 1986.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: _____. *Handbook on Ontologies*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 1–17.

HAAV, H.-M. A semi-automatic method to ontology design by using fca. In: CITESEER. *CLA*. [S.l.], 2004.

HAYES, P. J.; WEINSTEIN, S. P. Construe/tis: A system for content-based indexing of a database of news stories. In: *Proceedings of the The Second Conference on Innovative Applications of Artificial Intelligence*. [S.l.]: AAAI Press, 1991. (IAAI '90), p. 49–64.

KLIR, G.; YUAN, B. *Fuzzy sets and fuzzy logic: Theory and applications*. Upper Saddle River, New Jersey: Prentice Hall PTR, 1995. ISBN 0-13-101171-5.

KLIR, G. J.; YUAN, B. *Fuzzy sets, fuzzy logic, and fuzzy systems*. [S.l.]: World Scientific, 1996.

KUMAR, C. A.; SRINIVAS, S. Concept lattice reduction using fuzzy k-means clustering. *Expert systems with applications*, Elsevier, v. 37, n. 3, p. 2696–2704, 2010.

LIU, J.; MI, J.-S. A novel approach to attribute reduction in formal concept lattices. In: SPRINGER. *International Conference on Rough Sets and Knowledge Technology*. [S.l.], 2008. p. 426–433.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.: s.n.], 1967. v. 1, n. 14, p. 281–297.

MAIO, C. D. et al. Towards an automatic fuzzy ontology generation. In: IEEE. *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on*. [S.l.], 2009. p. 1044–1049.

MORAES, S. M. W. *Construção de Estruturas Ontológicas a partir de Textos: Um Estudo Baseado no Método Formal Concept Analysis e em Papéis Semânticos*. Tese (Doutorado) — Tese de Doutorado. PUC-RS, Porto Alegre, 2012.

OBITKO, M. et al. Ontology design with formal concept analysis. In: *CLA*. [S.l.: s.n.], 2004. v. 110.

PAL, N. R. et al. A possibilistic fuzzy c-means clustering algorithm. *IEEE transactions on fuzzy systems*, IEEE, v. 13, n. 4, p. 517–530, 2005.

PEI, D.; LI, M.-Z.; MI, J.-S. Attribute reduction in fuzzy decision formal contexts. In: IEEE. *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*. [S.l.], 2011. v. 1, p. 204–208.

PRISS, U. Formal concept analysis in information science. *Annual review of information science and technology*, Wiley Online Library, v. 40, n. 1, p. 521–543, 2006.

REZENDE, S. O. *Sistemas Inteligentes: fundamento e aplicações*. [S.l.]: Manole, 2003.

RIOS, T. *Organização flexível de documentos*. Tese (Tese (Doutorado em Ciências de Computação e Matemática Computacional)) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2013.

RIOS, T. N. *Organização flexível de documentos*. Tese (Tese de Doutorado) — Instituto de Ciências Matemáticas e de Computação (ICMC-USP), abril 2013.

TOUZI, A. G.; MASSOUD, H. B.; AYADI, A. Automatic ontology generation for data mining using fca and clustering. *arXiv preprint arXiv:1311.1764*, 2013.

WILLE., B. G. e R. Applied lattice theory: Formal concept analysis. In: . [S.l.: s.n.], 1997. p. 14.

WILLE, R. Formal concept analysis as mathematical theory of concepts and concept hierarchies. In: GANTER, B.; STUMME, G.; WILLE, R. (Ed.). *Formal Concept Analysis*. [S.l.]: Springer Berlin Heidelberg, 2005. v. 3626, p. 1–33.

WOLF, K. E. A first course in formal concept analysis - how to understand line diagrams. In: FAULBAUM, F. (Ed.). *Advances ha Statistical Software*. [S.l.]: Gustav Fischer Verlag, Stuttgart, 1991. p. 10.

XIE, X.; BENI, G. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 841– 847, 1991.

YEVTUSHENKO, S. A. System of data analysis concept explorer. *Proceedings of the 7th National Conference on Artificial Intelligence KII-2000*, p. 127–134, 2000.

ZHANG, S. et al. A completeness analysis of frequent weighted concept lattices and their algebraic properties. *Data Knowledge Engineering*, v. 81–82, p. 104–117, 2012.

ZHANG, W.; WEI, L.; QI, J. Attribute reduction theory and approach to concept lattice. *Science in china series F: Information sciences*, Springer, v. 48, n. 6, p. 713–726, 2005.