

# USO DA ANÁLISE FORMAL DE CONCEITOS PARA O CÁLCULO DE SIMILARIDADES COMO SUPORTE PARA UM SISTEMA ESPECIALISTA DE DIAGNÓSTICOS VIA WEB

VINÍCIUS M. **FERNANDES**<sup>1</sup>; KLEBER X. S. **SOUZA**<sup>2</sup>; SILVIA M. F. S. **MASSRUHÁ**<sup>3</sup>;  
LEONARDO M. **CARREIRO**<sup>4</sup>

Nº0903005

**Resumo:** O raciocínio para a descoberta de um diagnóstico tanto para um sistema automatizado quanto para um especialista consiste em achar a correspondência de um conjunto de sintomas numa matriz (esparsa) contendo um conjunto completo das doenças e os sintomas. Sempre que um sintoma se manifesta, o especialista ativa nessa matriz as possíveis doenças que podem manifestar este sintoma. O raciocínio prossegue até o ponto que um grupo muito pequeno de doenças (idealmente, apenas uma) seja conhecido como o causador daqueles sintomas. O processo de raciocínio que parte da causa até as consequências tem recebido atenção crescente desde a proposição da Teoria das Coberturas Parcimoniosas por Peng e Regia, no início da década de 90, em oposição aos sistemas especialistas comumente encontrados que partiam das consequências em direção à causa. Como geralmente o número de sintomas é grande, é necessário agrupar as doenças similares em conjuntos de forma que os sintomas mais freqüentes são perguntados em primeiro lugar. Dessa maneira, o raciocínio favorece a redução do universo de possíveis doenças, excluindo aquelas que não manifestam o sintoma. Este trabalho avalia a similaridade entre as doenças considerando um conjunto comum de doenças e distintos sintomas e propõe um método para estruturar a o espaço de pesquisa.

**Abstract:** Diagnostic reasoning either automated or conducted by diagnosticians relies on having a set of symptoms and trying to match them against a (sparse) matrix containing the complete set of diseases and their corresponding symptoms. Whenever a symptom manifests itself, the diagnostician activates in this matrix several possible diseases which could manifest that symptom. The reasoning proceeds up to the point in which a very small subset of diseases (ideally one) is known to cause those symptoms. The reasoning process that goes from the causes to the consequences has received increasing attention since the proposition of Parsimonious Covering Theory by

1. BOLSISTA CNPq: Graduação em Engenharia de Computação, IC/UNICAMP, Embrapa Informática Agropecuária, Campinas-SP, ✉ [vinicius@cnptia.embrapa.br](mailto:vinicius@cnptia.embrapa.br)

2. ORIENTADOR: Doutor em Telemática, Embrapa Informática Agropecuária, Campinas-SP, ✉ [kleber@cnptia.embrapa.br](mailto:kleber@cnptia.embrapa.br)

3. ORIENTADORA: Doutora em Computação Aplicada, Embrapa Informática Agropecuária, Campinas-SP, ✉ [silvia@cnptia.embrapa.br](mailto:silvia@cnptia.embrapa.br)

4. COLABORADOR: Graduação em Engenharia de Computação, IC/UNICAMP, Embrapa Informática Agropecuária, Campinas-SP, ✉ [lmachado@cnptia.embrapa.br](mailto:lmachado@cnptia.embrapa.br)

Peng and Reggia in early 90's, in opposition to the one commonly found in expert systems that were based from the consequences to the causes. As the number of symptoms is usually large it is necessary to group similar diseases together in such a way that the most frequent symptoms are asked first. In this way, the reasoning further reduces the space of possible diseases by excluding those that do not manifest that symptom. This paper evaluates similarities among diseases considering the set of common and distinct symptoms and proposes a method for structuring the space of search.

## **Introdução**

O raciocínio para a descoberta de um diagnóstico tanto para um especialista em diagnósticos ou um para um sistema automatizado é um complexo sistema cognitivo que consiste basicamente em achar a correspondência de um conjunto de sintomas com suas possíveis doenças. Outra complicação é a fato de que alguns sintomas ocorrem em determinados prazos e intensidades (Massruhá et al. 2004). Sistemas automatizados de diagnósticos, ou sistemas especialistas, exigem a construção de uma base de conhecimento contendo o conjunto de sintomas e doenças mais completo possível, que é utilizado em conjunto com um motor de inferência.

Como o número de doenças que causam certo sintoma pode ser grande, a Teoria das Coberturas Parcimoniosas organiza o conjunto de hipóteses de tal modo que seja abrangente o suficiente para explicar a totalidade dos sintomas (cobertura) e ainda pequeno o suficiente para minimizar a complexidade da explicação (parcimônia).

Este trabalho propõe a estruturação do espaço de pesquisa em um sistema de diagnóstico utilizando a Análise Formal de Conceitos (FCA), uma técnica de análise dos dados baseada na Teoria dos Reticulados e no Cálculo Proposicional (Wille, 1982), como uma ferramenta de apoio na identificação das relações ocultas entre os dados. A FCA é especialmente adequada para exploração de conhecimentos simbólicos (conceitos) contido em um contexto formal, como um banco de dados, ou uma ontologia.

Para a aplicação da FCA, a relação matemática <doenças, sintomas> expressa na matriz é mapeado para a FCA como <objetos, atributos>. Como resultado, o algoritmo de ordenação produz uma estrutura matemática chamada “Reticulado de Conceitos”,

que mostra na parte superior os sintomas mais comuns e na parte inferior os menos freqüentes. As doenças estão associadas ao ponto (nó), que engloba, na hierarquia do reticulado, todos os respectivos sintomas. Usando o reticulado, a similaridade medida avalia o quão perto estão as doenças através da contagem do número de elementos estruturantes que tem em comum.

A aplicação da FCA para este problema deu resultados interessantes, como por exemplo, conjuntos de sintomas que não ocorrem de forma isolada, o que indica que, talvez, o sistema possa perguntar apenas um deles uma vez que o outro está implícito. Outro resultado foi que doenças com maior valor de similaridade coincidiram com as agrupadas por um especialista humano (fitopatologista), uma indicação da qualidade da medida de similaridade.

## **Material e Métodos**

- Uso da ferramenta Galicia 3 para entrada da matriz de contexto contendo a relação entre as doenças e os sintomas, construção e visualização do reticulado de conceitos.
- Aplicação da Análise Formal de Conceitos.
- Programa, implementado em Java, que através do arquivo XML contendo o reticulado de conceitos: gera a similaridade numérica entre os conceitos, armazena a matriz de similaridades no formato CSV e sua visualização através de uma árvore hiperbólica no formato HTML
- Ferramenta Eclipse para desenvolvimento em Java.
- Base de conhecimento contendo 41 doenças do milho e 75 sintomas.

## **Resultados e Discussão**

Na Análise Formal de Conceitos, a abstração dos conceitos presentes no pensamento humano, no qual conceitos são objetos com determinadas classes de atributos, está estruturada em um reticulado, o Reticulado de Conceitos. Neste reticulado, se A é um conceito acima de um conceito B, e os dois estão ligados, o conceito A é um conceito mais geral do que B e, como tal, carrega parte dos atributos de B. Como consequência, é correto afirmar sempre que B acontece, A também está presente, sugerindo uma vinculação lógica. No reticulado, não se enxerga apenas uma

hierarquia de conceitos, mas também todo o conjunto binário de relações presentes entre os conceitos. Isso faz com que a análise visual dos dados possa ser obtida olhando para uma hierarquia de classes. Na Figura 1, cada nó no grafo é um conceito.

Tabela 1: Objetos e Atributos representados no diagrama da Figura 1.

Objetos	Atributos					
	s1	s2	s3	s4	s5	s6
d1	x			x		
d2	x		x	x		
d3	x		x			
d4	x					x
d5		x	x	x		
d6		x	x			
d7		x			x	
d8				x	x	x
d9		x			x	

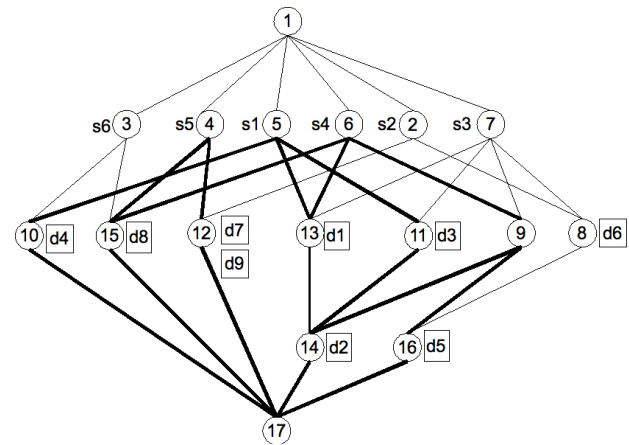


Figura 1: Diagrama correspondente ao reticulado obtido com a relação da tabela 1.

No reticulado de conceitos da Figura 1, os círculos de 1 a 17 são nós, as legendas ao lado dos círculos são atributos (sintomas), e os retângulos representam objetos. Neste reticulado, s1, s2 e s3 são atributos do objeto d2, porque eles estão posicionados em nós a partir do nó 14, no qual d2 que está posicionado, até o nó raiz. Objeto d5 também possui os atributos s2 e s3, mas não s1. Dessa maneira, pode-se dizer que d2 compartilha dois atributos com d5 e que possui um atributo que d5 não tem. É a partir dessas informações que a avaliação similaridade é realizada

Se dois objetos foram colocados no mesmo nó (conceito), eles possuem os mesmos atributos e são, portanto, instâncias da mesma classe de objetos que possuem aquele conjunto de atributos. O número de atributos em comum pode então ser ponderado em função do número de atributos que está presente em apenas um dos objetos para medir a similaridade entre dois objetos. Contudo, muitas vezes atributos podem aparecer em pares ou trios com a consequência de serem posicionados no mesmo nó do reticulado. Isso significa que do ponto de vista estrutural, os atributos não estão adicionando informações relevantes para diferenciar objetos.

A fim de contornar este problema, uma medida de similaridade estrutural foi proposta em (Souza e Davis, 2004). Essa medida considera alguns elementos especiais do reticulado, chamados *meet-irreducibles*. Estes elementos podem ser identificados facilmente no reticulado como os nós que possuem apenas uma aresta com as camadas superiores do reticulado. No reticulado da Figura 1, dos nós 2 ao no 7 reúne-se todos os elementos irredutíveis. Todos os seis sintomas são posicionados no topo do reticulado. A similaridade medida considera o número de elementos *meet-irreducible* em comum, bem como o número de tais elementos que cada nó ( $ni$ ), tem em separado, como seguinte:

$$S(ni, nj) = \text{Struct}(ni \cap nj) / (\text{Struct}(ni \cap nj) + 0.5 \text{Struct}(ni - nj) + 0.5 \text{Struct}(nj - ni))$$

Na equação 1,  $\text{Struct}(ni \cap nj)$  representa o número de elementos estruturantes compartilhados por  $ni$  and  $nj$ , e  $\text{Struct}(ni - nj)$  representa o número de elementos estruturantes em  $ni$  mas não em  $nj$ .

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,67	0,00	0,00	0,00	0,67	0,00	0,00	0,00	0,50	0,29
3	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,67	0,00	0,00	0,00	0,00	0,50	0,00	0,29
4	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,67	0,00	0,00	0,50	0,00	0,29
5	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,67	0,67	0,00	0,67	0,50	0,00	0,00	0,29
6	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,67	0,00	0,00	0,00	0,67	0,50	0,50	0,50	0,29
7	0,00	0,00	0,00	0,00	0,00	0,00	1,00	0,67	0,67	0,00	0,67	0,00	0,00	0,50	0,00	0,50	0,29
8	0,00	0,67	0,00	0,00	0,00	0,00	0,67	1,00	0,50	0,00	0,50	0,50	0,00	0,40	0,00	0,80	0,50
9	0,00	0,00	0,00	0,00	0,00	0,67	0,67	0,50	1,00	0,00	0,50	0,00	0,50	0,80	0,40	0,80	0,50
10	0,00	0,00	0,67	0,00	0,67	0,00	0,00	0,00	0,00	1,00	0,50	0,00	0,50	0,40	0,40	0,00	0,50
11	0,00	0,00	0,00	0,00	0,67	0,00	0,67	0,50	0,50	0,50	1,00	0,00	0,50	0,80	0,00	0,40	0,50
12	0,00	0,67	0,00	0,67	0,00	0,00	0,00	0,50	0,00	0,00	0,00	1,00	0,00	0,00	0,40	0,40	0,50
13	0,00	0,00	0,00	0,00	0,67	0,67	0,00	0,00	0,50	0,50	0,50	0,00	1,00	0,80	0,40	0,40	0,50
14	0,00	0,00	0,00	0,00	0,50	0,50	0,50	0,40	0,80	0,40	0,80	0,00	0,80	1,00	0,33	0,67	0,67
15	0,00	0,00	0,50	0,50	0,00	0,50	0,00	0,00	0,40	0,40	0,00	0,40	0,40	0,33	1,00	0,33	0,67
16	0,00	0,50	0,00	0,00	0,00	0,50	0,50	0,80	0,80	0,00	0,40	0,40	0,40	0,67	0,33	1,00	0,67
17	0,00	0,29	0,29	0,29	0,29	0,29	0,29	0,50	0,50	0,50	0,50	0,50	0,50	0,67	0,67	0,67	1,00

## Conclusão

Após testes realizados com doenças do milho que apresentaram resultados muito bons comparando a árvore de decisão desenhada por um fitopatologista e as similaridades calculadas usando essa medida de similaridade, foi elaborado um algoritmo que faz uso dessa medida para seleção da próxima pergunta.

O algoritmo desenvolvido resultou na implementação de um aplicativo em Java que percorre o reticulado, interpretado pelo programa através de um arquivo XML, considerando como ponto de partida para o início das perguntas um atributo

estruturante selecionado pelo usuário. Após a seleção desse sintoma, todos os nós conceitos que o possuem como pai são armazenados e ordenados de acordo com a similaridade e dá-se início as perguntas através dos sintomas (atributos) de um conceito (agrupamento de atributos) com maior similaridade em relação ao sintoma estruturante. Caso confirmados, um a um, os sintomas de um conceitos, armazena-se os nós que possuem aquele conceito como pai e continua-se a percorrer recursivamente o reticulado e caso um dos sintomas não seja confirmado, o próximo nó a ser visitado é o que apresenta segunda maior similaridade com o pai e assim por diante até se esgotarem todas as possibilidades.

## REFERÊNCIAS BIBLIOGRÁFICAS

- Ganter, B., Wille, R. (1999) Formal Concept Analysis: Mathematical Foundations. Springer, Berlin - Heidelberg - New York
- Massruha, S.M.F.S., Sandri, S.A., Wainer, J., 2003. Fuzzy Covering Theory: an alternative approach for diagnostic problem-solving. Proceedings of the Efitia 2003 Conference. Budapest, pp. 768-775.
- Massruha, S.M.F.S., 2003. Uma teoria de coberturas nebulosas para diagnostico, investigacao e tratamento. PhD Thesis, CAP/INPE. Sao Jose dos Campos, Brazil.
- Massruha, S., Sandri, S., Wainer, J., 2004. Ordering manifestations for investigation in incomplete diagnosis. Proceedings of the Information Processing and Management of
- Uncertainty in Knowledge-based Systems (IPMU 2004), Perugia, Italy.
- Peng, Y., Reggia, J.A., 1990. Abductive inference models for diagnostic problem-solving. Springer-Verlag.
- Souza, K.X.S., Davis, J. 2004. Aligning ontologies and evaluating concept similarities. In: On The Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, Lanarca, Cyprus. Proceedings. Number 3291 in Lecture Notes in Computer Science, Springer-Verlag Heidelberg, pp. 1012-1029.
- Wille, R. 1982. Restructuring lattice theory: An approach based on hierarchies of concepts. In: Rival, I. (Ed.). Ordered Sets. Volume 83 of NATO Advanced Study Institute Series C. Reidel, Dordrecht, pp. 445-470.