Using formal concept analysis to calculate similarities among corn diseases

K.X.S. de Souza, S.M.F.S. Massruhá and V.M. Fernandes
Embrapa Information Technology, Caixa Postal 6041, Campinas, SP, Brazil;
kleber@cnptia.embrapa.br; silvia@cnptia.embrapa.br; vinicius@cnptia.embrapa.br

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) are 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 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.

Keywords: knowledge representation, ontology, lattice

Introduction

Diagnostic reasoning either conducted by diagnosticians or by automated systems is a complex cognitive process. Basically, it involves the matching of a set of symptoms with their possible diseases (or malfunctions in case of machines). Another complication is the fact that some symptoms occur at certain time frames and intensities (Massruhá *et al.,* 2004). Automated diagnostic systems, or expert systems, require the construction of a knowledge base containing the most possible complete set symptoms and diseases, which is used together with an inference engine.

Expert systems, like Mycin (Buchanan and Shortliffe, 1984) for example, usually perform inference from the consequences to the causes. Others, like the one proposed by Massruhá (2003) and Massruhá *et al.* (2003) is based on a theoretical model that formalises abductive reasoning in diagnosis, the Parsimonious Covering Theory (Peng and Reggia, 1990). Abductive reasoning performs from the causes to the consequences. To explain the difference, in deductive inference, given the fact A and the rule A→ B, then B is true. In abductive reasoning, given the fact B and the rule A→ B, then A is plausible, because A may be one of the possible causes of B. In diagnosis, A may be one of the diseases causing the symptom B, provided that symptom B is present.

As the number diseases of causing a certain symptom may be large, Parsimonious Covering Theory organises the set of hypothesis such that they be large enough to explain the totality of the symptoms (coverage) and yet small enough to minimise the complexity of the explanation (parsimony).

This paper proposes the use of Formal Concept Analysis (FCA), a data analysis technique based on Lattice Theory and Propositional Calculus (Wille, 1982), as a supporting tool to help the identification of hidden relations among data. FCA is especially suitable for exploration of symbolic knowledge (concepts) contained in a formal context, such as a corpus, a database, or an ontology. For the application of FCA, the mathematical relation <diseases, symptoms> expressed in the above matrix is mapped to FCA's <objects, attributes>. As a result, the ordering algorithm produces a mathematical structure called concept lattice, which shows on the top the most common symptoms and in the bottom the least frequent ones. Diseases are attached to the point (node) that encompasses, in the lattice hierarchy, all respective symptoms. Using the lattice, a similarity measure evaluates how close diseases are by counting the number of structuring elements they have in common.

The application of FCA to this problem gave interesting results, as for instance, sets of symptoms that never occur in isolation, which indicate that perhaps the system should ask for only one of them since the other is implied. Another result was that diseases with greater similarity value coincided with those grouped together by a human expert (phyto-pathologist), an indication of the quality of the similarity measure.

The paper is organised as follows. Sections 2 and 3 explains in general lines how Parsimonious Covering Theory and Formal Concept Analysis. Each of these formalisms are grounded in solid mathematical basis and would require much more space for a complete explanation. Please refer to Reggia and Peng (1986) and Wille (1982) for further details. Then, Section 4 presents the application of the similarity measure developed in Souza and Davis (2004). Finally, the results obtained by the application of FCA to the problem presented in Reggia and Peng (1986) are discussed Section 5.

Parsimonious covering theory

The formal method of diagnostic reasoning in the Parsimonious Covering Theory represents knowledge via a network of causal associations (Reggia and Peng, 1986). In this theory, an explanation E^+ for S^+ is a set of diseases such that:

1. S^+ is a subset of *Symptoms*(E^+);
2. E^+ is parsimonious with respect to the Irredundancy criterion in which no proper subset of E^+ covers S^+.

These concepts can be better explained with a working example. This example was adapted from (Reggia and Peng, 1986) and will be used during the whole paper.

Figure 1 shows a set of diseases d_1, ... d_9, and a set of symptoms, $s_1,...,s_6$. In Parsimonious Covering Theory, given the set of symptoms $\{s_1, s_4, s_5\}$, there are 12 possible combinations of diseases which could manifest them:

$\{(d_1,d_7), (d_1,d_8), (d_1,d_9), (d_2,d_7), (d_2,d_8), (d_2,d_9), (d_3,d_8), (d_4,d_8), (d_3,d_5,d_7), (d_3,d_5,d_9), (d_4,d_5,d_7), (d_4,d_5,d_9)\}$

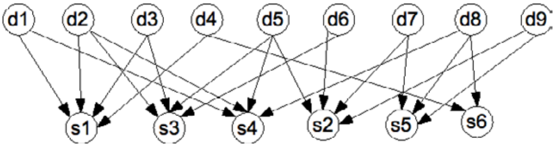

Figure 1. Graph relating diseases and symptoms.

In this case, one has to further investigate in order to come to a conclusion, and that is carried out by an inference engine. However, if the symptom were only s_6, there would be only two possible hypothesis, d_4 and d_8, and the inference process would stop.

Two factors greatly influence the precision of an inference system: the size of the knowledge base and how well the system matches the set of given symptoms with the knowledge stored. Unfortunately, considering the whole set of possible combinations of diseases and symptoms at every step of diagnostic process also implies the increase of computational complexity of the underlying algorithms. The combinatorial explosion would eventually require exponential execution time, what would not be acceptable from the user point of view.

Reggia and Peng (1986) pointed out that two problems have to be addressed during problem-solving to obtain a diagnosis, namely how questions should be generated to obtain additional information, and when problem-solving should terminate. There are many extensions of the theory to address these problems. Conditional probability and an entropy minimising metric have served as a basis for many of them. For further details, please refer to (Reggia and Peng, 1986; Peng and Reggia, 1990; Massruhá *et al.,* 2004).

Applying formal concept analysis

Formal Concept Analysis (FCA) is a technique based on lattice theory and propositional calculus, producing what is called a Concept Lattice. FCA has been applied in many domains, such as structuring of information systems, knowledge discovery in databases, political science and psychology. Due to space limitations, this paper cannot provide details of the application of FCA, because it involves a great amount of mathematics. Please see (Souza and Davis, 2004) for further information. However, a brief explanation of the process will be provided in the sequel.

FCA involves the analysis of a set of attributes S={s_i}, corresponding to the symptoms of the example adapted from (Reggia and Peng, 1986), a set of objects D={ d_i } (diseases) which contain these attributes (manifest these symptoms), and a binary relation *R* between *D* and *S*. The first step in the application of FCA is to create a formal context, displayed in Table 1. Taking Figure 1 as a representation of the relation between D and S, this context is created in such a way that whenever there is an arrow from a disease to a symptom the intersection between the two is marked with an 'x'.

Table 1. Objects and attributes represented in the lattice of Figure 2.

Objects	Attributes					
	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	

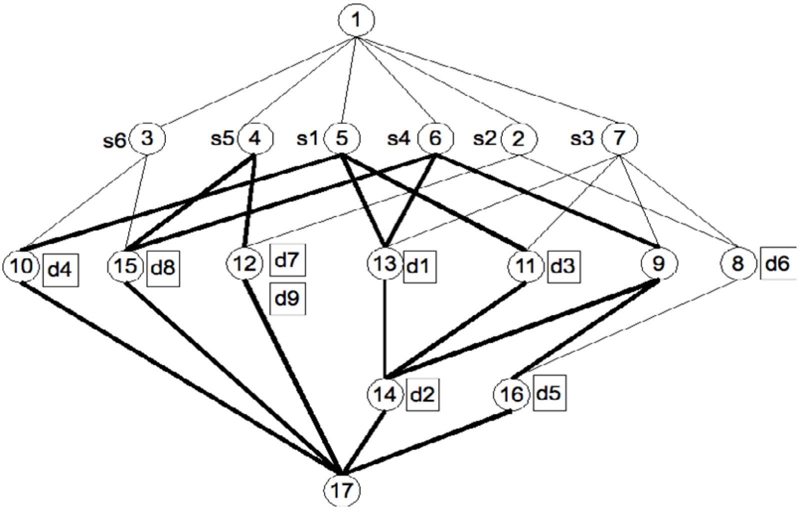

Figure 2. Hasse diagram corresponding to the lattice obtained from Table 2.

FCA analyzes which subsets of the set of objects have the same attributes and, conversely, which subset of attributes is shared by the same objects. For example, the subset $\{s_2,s_3\}$ is present both in d_5 and in d_6, and s_4 exists only in d_5. So, $\{s_2,s_3\}$ is related to $\{d_5,d_6\}$ and $\{s_2,s_3,s_4\}$ is related to $\{d_5\}$. In Formal Concept Analysis, the abstraction of concepts present in human thoughts, in which concepts are classes of things having certain attributes, is structured in a lattice, the concept lattice. In this lattice, if a concept A is above a concept B, and the two are linked, concept A is more general than B and, as being such, it carries part of attributes of B. As a consequence, one can say that whenever B happens, A is also happening, which suggests a logical entailment. In the lattice, one can not only see a hierarchy of concepts, but also the whole set of binary relations present among concepts. That makes the visual analysis of data superior to the one can be obtained by looking at a hierarchy of classes. In Figure 2, every node in the graph is a concept.

In the concept lattice of Figure 2, the circles 1,...,17 are nodes, the strings besides the circles are attributes (symptoms), and the rectangles represent objects. In this lattice, s_1, s_2 and s_3 are attributes of object d_2, because they are positioned in nodes from the node labeled *14*, at which d_2 is positioned, up to the root node. Object d_5 has also attributes s_2 *and* s_3, but not s_1. In this way, one can say that d_2 shares two attributes with d_5 and that it has one attribute that d_5 does not have. It is using this information that the similarity evaluation is performed, as show in next section.

The similarity measure

If two objects are positioned in the same node (concept), they have the same attributes and are, therefore, instances of the same class of objects that have that set of attributes. The number of attributes in common can then be weighted against the number of attributes that are present only in one of the objects to measure the similarity between two objects. However, most times attributes may appear in pair or triplets with the consequence of being positioned in the same node in the lattice. That means that from the structural point of view the attributes are not adding relevant information to differentiate objects.

In order to circumvent this problem, a structural similarity measure was proposed in (Souza and Davis, 2004). It considers some special elements of the lattice called meet-irreducible elements. These elements can be identified easily in the lattice as those nodes having only one edge linking them to the upper layers of the lattice. In the lattice of Figure 2, nodes 2 through 7 are all meet-irreducible elements. All the six symptoms are positioned in the top of the lattice in this case. The similarity measure considers the number of meet-irreducible elements in common and the number of such elements that each node (n_i) has separately, as follows:

$$S(n_i, n_j) = Struct(n_i \cap n_j) / (Struct(n_i \cap n_j) + 0.5\ Struct(n_i - n_j) + 0.5\ Struct(n_j - n_i)) \quad (1)$$

In Equation 1, Struct($n_i \cap n_j$) represents the number of structural elements shared by n_i and n_j, and Struct(n_i-n_j) represents the number of structural elements in n_i but not in n_j. For example, the similarity between nodes 14 and 16 is calculated as:
Struct($n_{14} \cap n_{16}$) = 2, which correspond to nodes 6 and 7;
Struct(n_{14}-n_{16}) = 1, which corresponds to node 5;
Struct(n_{16}-n_{14}) = 1, which corresponds to node 2.
Then, S($n_{14} \cap n_{16}$) = 2/(2 + 0.5 + 0.5) = 0.67.

Table 2 shows the results of the calculation of the similarity measure for each pair of nodes of the lattice of Figure 2.

Table 2. Similarities among nodes in the lattice of Figure 2.

Node	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	0.00	0.67	0.00	0.67	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

Analysis of the results

One important fact about Concept Lattices (proved in Theorem 1 in Ganter and Wille (1999)) is that the *infimum* (meet) and *supremum* (join) between every pair of objects are defined in terms of the usual set operators ($\cap$, $\cup$, $\subset$, $\supset$). Moreover, the *supremum* of two elements serves as a basis of comparison between them because it contains all the common attributes of these two elements. For example, the objects d_2 and d_5 have in common the attributes s_4 and s_3, because $join(n_{14}, n_{16}) = n_9$ and n_9 accumulates the attributes coming from n_6 and n_7, which are s_4 and s_3, respectively. Conversely, it is also possible to discover by inspection on the lattice which objects share a set of attributes by using the meet operation. For instance, the objects that have s_1 and s_3 as attributes are d_3 and d_2, because $meet(n_5,n_7) = n_{11}$ and the objects bellow n_{11} in lattice hierarchy are d_3 and d_2. Concept Lattices also have been used in data mining because its direct inspection allows for the calculation of support and confidence measures. For example, the support of attribute s_4 is 4/9, because from the total of 9 objects there are 4 objects bellow n_6 in the lattice. This results from the fact that the Concept Lattice considers all the possible combinations of the attributes (Powerset of the set of attributes) before calculating the existing infima and suprema. This information can be used to improve the selection of the next question by the inference engine.

The exploration of combinations of attributes and objects of Concept Lattices makes them especially useful together with Parsimonious Covering Theory. In accordance with Reggia and Peng (1986), the 12 combinations obtained by the application of Parsimonious Covering Theory in Section 2, when the symptoms $\{s_1, s_4, s_5\}$ were present, could be generated by the following generators:

$\{d_1, d_2\} \times \{d_7, d_8, d_9\}$
and
$\{d_8\} \times \{d_3, d_4\}$
and
$\{d_5\} \times \{d_3, d_4\} \times \{d_7, d_9\}$

These generators could also be obtained from the lattice of Figure 2. The edges marked with darker lines represent those activated by the same symptoms. One immediate information that can be seen in the lattice is that all the diseases, except d_6, show these symptoms, because they are positioned from the meet-irreducible elements 4, 5 and 6. These are precisely the nodes at which s_1, s_5 and s_4 are positioned, respectively.

Moreover, node 13 associates nodes 5 and 6 (meet operation on the lattice). From that, one can say that d_1 and d_2, which are positioned at or bellow 13, manifest symptoms s_1 and s_4. Since s_1 and s_4 have been considered, one has to look at the other paths in the lattice to determine what is missing to account for s_5 in order com complete the symptoms. The answer comes from nodes 12 and 15 and its diseases attached: d_7, d_8 and d_9. Now, if one makes the cross product between d_1, d_2 and d_7, d_8, d_9 the first generator above is obtained.

Table 2 give the similarities among nodes in the lattice of Figure 2. From this table, one can see that diseases d_2 and d_5, which are near from each other, are 0.67% similar, whereas d_5 and d_4 are 0.00% similar. In the latter case, this happens because the two diseases have no attributes in common. Tests carried out with corn diseases showed very good results when comparing the similarities calculated using this similarity measure and a decision tree designed by a phyto-patologist. Currently, the diagnostic system designed and built in Embrapa Agricultural Informatics, which uses an entropy measure to select the next question in the diagnosis process is evaluating the use of this similarity measure to improve its performance.

Conclusion

This paper proposed an evaluation of the use of similarities for structuring the space of search in a diagnosis system. The similarity measure is based on Formal Concept Analysis, a method grounded on Lattice Theory. This method considered the set of common and distinct symptoms and grouped similar diseases together in such a way that diseases that do not manifest symptoms are avoided whereas those sharing same symptoms are considered first. The exploration of combinations of attributes and objects over Concept Lattices makes them especially useful for joint use with Parsimonious Covering Theory, because the generators calculated in the latter could also be obtained from the Concept Lattice. Tests carried out with corn diseases showed very good results comparing the decision tree constructed by the expert, which is based on grouping of symptoms, with the similarity among nodes calculated from the Concept Lattice.

References

Buchanan. B.G., Shortliffe, E.H., 1984. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.

Ganter, B., Wille, R., 1999. Formal Concept Analysis: Mathematical Foundations. Springer, Berlin.

Massruhá, S.M.F.S., Sandri, S.A., Wainer, J., 2003. Fuzzy Covering Theory: an alternative approach for diagnostic problem-solving. Proceedings of the Efita 2003 Conference. Budapest, pp. 768-775.

Massruhá, S.M.F.S., 2003. Uma teoria de coberturas nebulosas para diagnóstico, investigação e tratamento. PhD Thesis, CAP/INPE. São José dos Campos, Brazil. (in Portuguese).

Massruhá, 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.