1 Background

Formal Concept Analysis

Wille's knowledge-landscape metaphor[9] is used to describe the process through which information is explored using formal concept analysis[7, 6]. "Knowledge" is derived from the analysis of the structure of empirical data concerning the real-world. The task of knowledge exploration is seen as a dynamic one in which the computer is used as a medium through which aspects of the knowledge can be investigated. Central to the idea of formal concept analysis is the understanding that a fundamental unit of thought is a *concept*. The concept is constituted by its *intension* and its *extension*.

This understanding has been formalized by a *(formal)* context, $\mathbb{K}$ defined by a triple (G, M, I) where G and M are sets and I is a binary relation between G and M (i.e., $I \subseteq G \times M$. $(g, m) \in I$ is read as *"the object g has the attribute m"*. The *(formal)* concepts of $\mathbb{K}$ are the pairs (A, B) with $A \subseteq G$ and $B \subseteq M$ such that (A, B) is maximal with respect to the property $A \times B \subseteq I^1$. The set A is called the extent and B is called the intent of the concept (A, B). The set $\underline{\mathcal{B}}(\mathbb{K})$ of all concepts of a context $\mathcal{K}$ with order $(A_1, B_1) \leq (A_2, B_2) :\Leftrightarrow A_1 \subseteq A_2$ is always a complete lattice, and is called the *concept lattice* of the context $\mathbb{K}$.

In many applications, training data relates individual objects to attributes that take on several values. Such data is formalized as a many valued context (G, M, W, I)[8]. G is a set of objects, M is a set of attributes, W is a set of attribute values, and $I \subseteq G \times M \times W$ is a relation such that $(g_1, m_1, w_1) \in I$ and $(g_2, m_2, w_2) \in I$ implies that $w_1 = w_2$. Attributes m_i are understood as partial functions from G into W. A *(formal)* context in which there are no attribute values is often called a single-valued context. Multi-valued contexts can be mapped into formal contexts using *conceptual scales*. A conceptual scale for a set $Y \subseteq M$ is a single-valued context $(G_{\mathbb{S}}, M_{\mathbb{S}}, I_{\mathbb{S}})$ with $\times_{m \in Y} m(G) \subset G_{\mathbb{S}}$. The derived relation $J_{\mathbb{S}} \subseteq G \times M_{\mathbb{S}}$ is defined by $(g, n) \in H_{\mathbb{S}} :\Leftrightarrow ((w_m)_m \in Y, n) \in I_{\mathbb{S}}$ with $(g, m, w_m) \in I$ for all $m \in Y$.

The process of creating a conceptual scale is performed by using knowledge from the domain from which the data is taken. Often conceptual scales are created by hand, and their concept lattice, since they are represented by formal contexts, laid out by hand. Conceptual scales may also be used to impose an external ordering on the attributes, both for a multi-valued context, and for a single valued context. In our work, conceptual scales are created by the manipulation of a view of a medical taxonomy of terms. Manipulation of the view enables both the definition and the layout of the conceptual scale.

Hierarchies of Medical Terms

In order to infer meaning and draw interpretation using our knowledge representation, a language of the relevant symbols being study is needed. Our application, drawn from the medical domain, uses some of the common symbols used in medicine and contained in electronic dictionaries. These are briefly described in this section.

[1] Maximal is taken to mean $A = \{a \in G \mid \forall m \in B \; aIm\}$ and $B = \{b \in M \mid \forall g \in A \; gIb\}$.

The Unified Medical Language System, UMLS, is the result of an ongoing research and development effort by the National Library of Medicine. The central purpose of the UMLS is to facilitate the retrieval and integration of information from multiple machine-readable biomedical information sources. There are four sections to UMLS; the meta-thesaurus, the semantic network, a specialist lexicon, and an information sources map. The meta-thesaurus is the central vocabulary component of the UMLS. In 1996, the meta-thesaurus contained 252,892 distinct concepts and 542,723 different concept names from 30 vocabularies. The work described here is largely concerned with detecting the existence of medical concepts, in text documents, from the meta-thesaurus. Each vocabulary in the meta-thesaurus has a subsumption ordering defined on its concepts. A directed graph constructed from the covering relations[2] corresponding to the subsumption relations of the various vocabularies contains cycles. As a result we selected a single vocabulary from the meta-thesaurus.

Medical Subject Headings (MeSH) is a medical thesaurus contained in the UMLS meta-thesaurus. MeSH is commonly used for indexing and describing medical concepts. The Meta-thesaurus is organized by concept or meaning. The purpose is to link together concepts from its constituent vocabularies. Each concept in MeSH is given a unique concept identifier, CUI. Associated with each CUI is a set of alternative descriptions of that concept. These alternative descriptions are called terms. The set of terms associated with a concept in the meta-thesaurus encompass lexical variants.

In Australia, MediMedia publishes the MIMS range of drug and medical references. The main use of MIMS is by medical practitioners in Australia who consult it for drug and prescription information. MIMS has been released in electronic form. Drugs in the MIMS database are described by product names, form names, and generic names. The form names describe the form in which the drug is produced, for example tablets. The drugs described by product names in MIMS are placed in a two level hierarchy. There is a link between product of a particular form and collections of generic names. Via this link the hierarchy may be extended to the generic names.

2 Automated Concept Matching

In an initial phase concepts are detected in the document via full text matching. That is, if a string is found in a document that exactly matches, without regard to punctuation, capitalization, or white space, a term description in MeSH or a drug description in MIMS, that concept is marked as being in that document. This matching was decided upon as an initial step because more liberal schemes yielded much lower precision, at the gain of a modest increase in recall. The recall for this initial matching is quite high, due in most part to the variety of lexical variants that exist within MeSH coupled with the fact that many concepts consist of a single word.

3 Structure and Content Mark-up

The structure of the documents was marked up using SGML. A utility for hand mark-up of the text was created and is shown in Figure 1. The interface firstly presents the user with the currently automatically recognized terms, and requests that the user either reject incorrectly identified concepts, or identify concepts that it has missed.

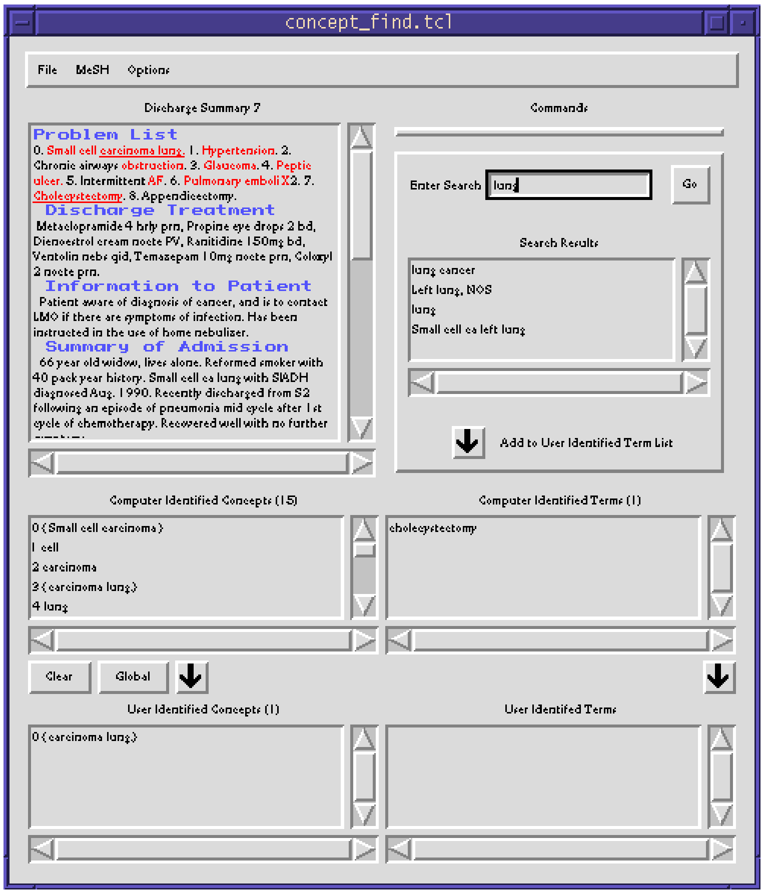

Figure 1: TFD: The Mark-up utility to help the clinician perform medical concept identification

When a new discharge summary is shown to the user, concept extraction is performed with the fresh information supplied by the user. The user is able to search the MeSH and MIMS database using either keywords or regular expressions. It was discovered that a user who is familiar with MeSH and MIMS, and a medical practitioner can perform almost all searches using keyword searching using a thesaurus of words for query expansion. The thesaurus is tailored and maintained by the user.

Two interfaces are provided for navigating in the hierarchical ordering of the terms in MeSH. Firstly, medical concepts revealed either by the automated recognition of concepts in MeSH, or by the users searching, may be added to a Folding Hasse Diagra by selecting them. The Folding Hasse diagram is maintained by the system, and is discussed in detail in the next section. Secondly when a MeSH term is selected, it is shown in a dialogue box together with its most immediate specializations and generalizations.

Figure 2 is an example of the immediate specialization and generalizations of the term “carcinoma”. Selection of a child or parent will cause the window to be updated with the information of that child or parent.

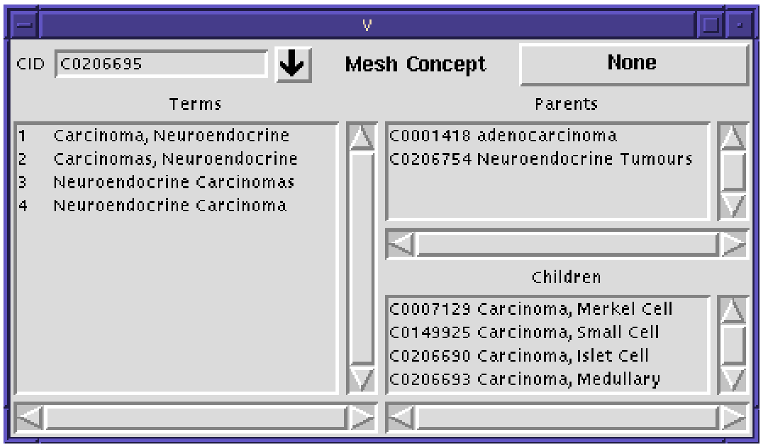

Figure 2: TFD: Displaying the Children and Parents of a MeSH Concept

4 Constructing Conceptual Scales

Once the medical concepts have been extracted from the document to a desired level, the knowledge contained in the documents can be explored using formal concept analysis. Knowledge exploration proceeds in formal concept analysis by the user specifying which medical concepts are relevant to their current query. For instance, the user may be interested in the relationship between *Asthma* and the prescription of a variety of drugs such as *Ventolin*.

A Hasse diagram is a graphical representation of a partially ordered set. The set of MeSH and MIMS concepts are organized in a partially ordered set. The Hasse diagram represents the covering relation, sometimes called a parent-child relation, as lines, and if one element is covered by another element, then the covered element is below the covering element in the diagram and there is a line between them. As a result of this, if one element is less than another in the partial order, it will appear below the element that it is less than, and there will be a path from the lesser element to the greater element.

The Hasse diagram of MeSH would contain over 100,000 elements, and be impossible to comprehend. In addition most of the elements would not be relevant to the user query. For example, a user interested in the relationship between *Asthma* and *Ventolin* may have no interest in whether or not the patients suffer from *Glaucoma*. Yet a relationship between these two terms would be inferred for display in the Hasse Diagram.

Another difficulty with the Hasse Diagram of the whole structure is that it is very large and presents computational difficulties. Instead, we present the user with a Folding Hasse Diagram. This is a diagram of the hierarchical relationship between just a set of concepts selected by the user. Since the process of creating a conceptual scale is a dynamic one, the user must be able to manipulate Hasse diagrams to reflect a changing representation of their information need.

We preserve the ordering relation of the elements, and generate a new covering relation induced by the smaller set of objects, and the same ordering relation. In order to perform this calculation efficiently a code is assigned to each element. This coding system uses sparse terms[3, 4] and described fully elsewhere[1].

Figure 3 is an example screen shot of the Folding Hasse Diagram. Each circle represents a concept from MeSH. Labels displaying the name of the concept are attached to concepts. A number of operations are presented in order to facilitate the manipulation of the diagram.

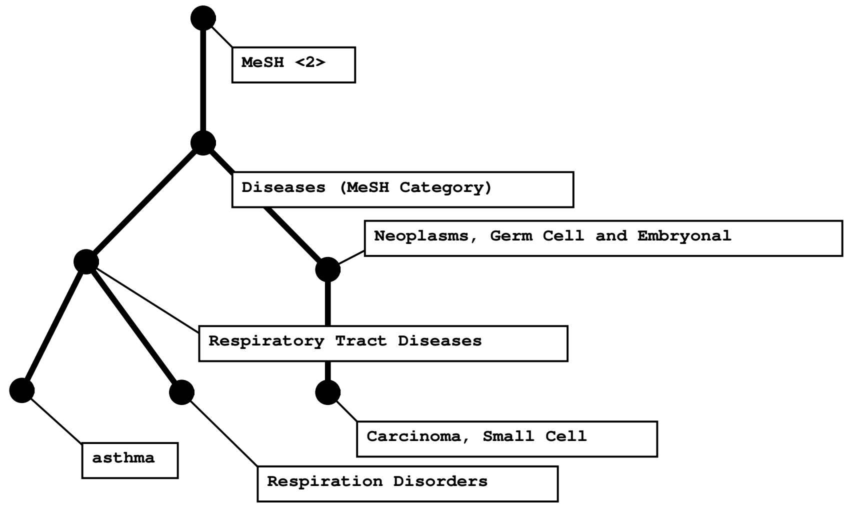

Figure 3: Folding Poset Viewer with MeSH Concepts

- **Insert Element** An element may be inserted into the diagram. This concept is selected by the user after performing a text query on the MeSH database. The element is inserted into the Hasse Diagram using topological search;
- **Insert Child Element** An element that is the child of an already displayed element may be displayed in the diagram;
- **Insert Parent Element** An element that is the parent of an already displayed element may be displayed in the diagram;
- **Remove Element** An element may be removed from the diagram. Any children that are not supported by other parents, are also removed in a recursive manner.

A join corresponds to the point at which paths from the two elements cross in the MeSH hierarchy. These joins can be calculated from the codes of the two elements. Every element pair has its join automatically added to the diagram — we call describe the diagram so drawn as *join-complete.*

The folding Hasse diagram uses some simple heuristics for layout. Firstly, it attempts to arrange elements on a grid which achieves parallel lines. Secondly it attempts to locate children underneath their parents. Thirdly it attempts to create symmetries, so that the same number of lines departing to the left, depart to the right of an element, and at the same angles. The user is able to move elements around to improve the layout. If the program decides to move an element in response to the insertion of a new element then this is performed by an animation to inform the user that the point is moving and allow them to preserve their mental map of the graph.

5 Building and Interpreting the Concept Lattice

Once the user has created a number of queries in the form of conceptual scales, these scales may be represented as contexts. From these contexts concept lattices are produced, and the diagrams of the conceptual scales augmented with new elements. These new elements are concepts in the concept lattice that correspond to the conjunction of a number of medical concepts in the scale produced by the user.

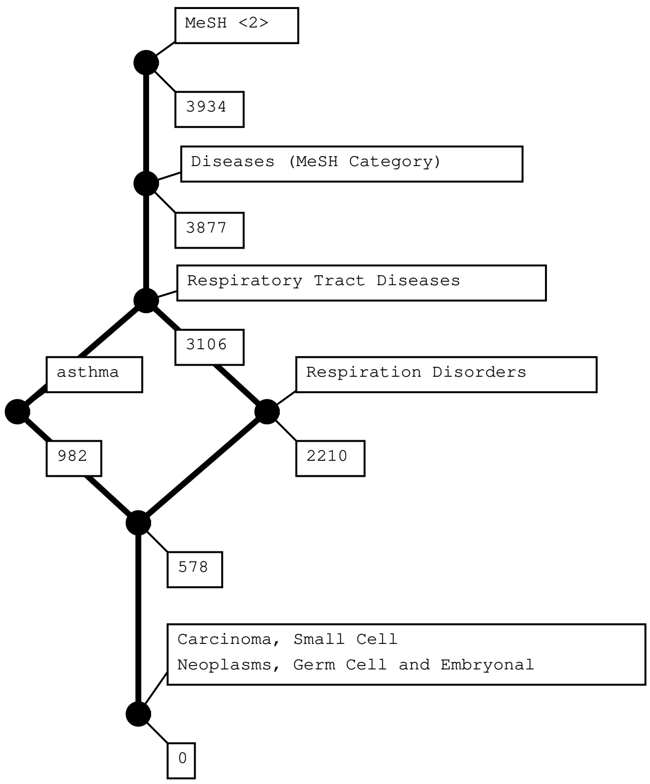

Figure 4: A concept lattice corresponding to the conceptual scale in figure 3

The creation of the lattices for each conceptual scale is performed using Ganter's Algorithm[5]. Figure 4 shows an example of a lattice that results from the conceptual scale in figure 3. This scale has elements labeled

"MeSH", "Disease", "Glaucoma", "Asthma" and "Carcinoma". These elements correspond to medical concepts in MeSH, and become the attributes in the concept lattice shown in figure 4.

The elements shown in figure 4 as circles are concepts in the concept lattice. These concepts have both an intent, and an extent. The intent is the set of attributes which are shared by all objects in the extent of the concept. The extent of the concept is the set of all objects that possess all of the attributes in the intent. In this case, the attributes are medical concepts from MeSH, and the objects are patients described by discharge documents. The ordering of concepts in the lattice is by attributes. The attributes in the intent of a concept, shown by a circle, may be obtained by collecting all the attributes assigned to ancestors of that concept in the diagram. So the concept labeled "Asthma" has "Asthma", "Respiratory Tract diseases", "Disease", and "MeSH" in its intent. The concepts that have generated new circles in the lattice correspond to the conjunction of attributes defined in the conceptual scale. These are point are were concepts "meet".

Two scales may be combined in a nested line diagram. When this occurs, the concept lattice of one scale is drawn inside the blown up circles of the concept in the concept lattice of the second scale. This indicates that a lattice product operation has been performed, and the attributes of the concept of the inner scale, can be determined by trace upwards to the top in both the inner scale, and upper scale.

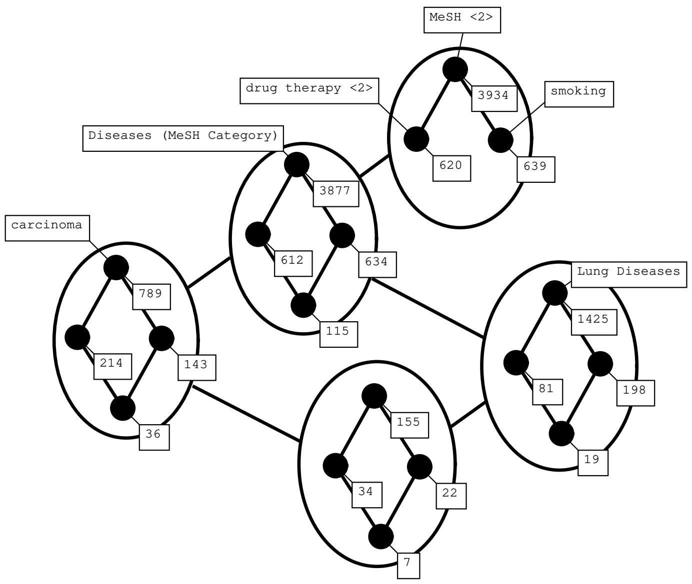

Figure 5: A nested line diagram exploring the relationship between Carcinoma, Chemotherapy and Smoking, numbers attached to vertices indicate the number of documents that cluster to that vertex

Figure 5 is an example of a nested line diagram. The outer scale has three attributes "Disease", "Carcinoma" and "Lung Diseases" while the inner scale has two attributes "Smoking" and "drug therapy". The intents of the concepts, represented by filled black circles, can be determined by following lines back upwards in the diagram.

Assigned to each small circle in the diagram is a number in a box. This number refers to the size of the extent of that concept. These numbers are the number of patients that correspond to that concept. It is possible to read implications and partial implications from the concept lattice. If one attribute label appears below another in the diagram, and there is a path between the concepts to which the attribute labels are attached, then we know that if an object has the lower attribute, it will also have the higher attribute.

Partial implications can be read by considering extents. For example consider the concept with extent "Carcinoma". This is the top concept in the far left hand large circle in Figure 5. The size of the extent of this concept is 789, i.e. 789 patients have "Carcinoma". Now consider the concept for "Carcinoma" and "Smoking". This is the concept below (and left) "Carcinoma" in Figure 5 and has 143 patients in its extent. By considering the size of the extents of these two concepts we can determine that of 789 patients with "Carcinoma", 143 were "Smoking".

6 Conclusion

This paper has presented a description of a technique and software tools used for the analysis of information as extracted from medical texts describing patients. Direct manipulations of a medical hierarchy allow a domain expert to construct scales to investigate this data. The method demonstrates how a knowledge representation language, and its diagrammatic rendering, can be used for document retrieval filtering.

References

[1] R. Cole and P. Eklund. Scalability in formal concept analysis. *Computational Intelligence*, 15(1), 1999.

[2] B. A. Davey and H. A. Priestly. *Introduction to Lattices and Order.* Cambridge Mathematical Textbooks. Press Syndicate of the University of Cambridge, 1990.

[3] Andrew Fall. Sparse term encodings for dynamic taxonomies. *Conceptual Structures: Knowledge Represenation as Interlingua, Proceedings of the 4th International Conference on Conceptual Structures*, pages 277–292, August 1996.

[4] Andrew Fall. *Reasoning with Taxonomies.* PhD thesis, Department of Computer Science, Simon Frasor University, Canada, 1997.

[5] B Ganter. Two basic algorithms in concept analysis. Technical report, Hochschule Darmstadt, 1984.

[6] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations.* Springer, 1998.

[7] R. Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. *Ordered Sets*, pages 445–470, 1982.

[8] R. Wille. Concept lattices and conceptual knowledge systems. *Computers and Mathematical Applications*, 23(6-9):493–515, 1992.

[9] R. Wille. Landscapes of knowledge: A pragmatic paradigm for knowledge processing. In G. Mineau and A. Fall, editors, *Proceedings of the second international symposium on Knowledge Retrieval, Use and Storage for Efficiency*, pages 2–13, 1997.