CHiC:
A Fast Concept Hierarchy Constructor for Discrete or Mixed Mode Databases

Ola Ågren
Department of Computing Science
Umeå University, Umeå, Sweden

E-mail: Ola.Agren@cs.umu.se

Abstract

In this paper we propose an algorithm that automatically creates concept hierarchies or lattices for discrete databases and datasets. The reason for doing this is to accommodate later data mining operations on the same sets of data without having an expert create these hierarchies by hand.

Each step of the algorithm will be examined; We will show inputs and output for each step using a small example. The theoretical upper bound of the complexity for each part of the algorithm will be presented, as well as real time measurements for a number of databases. We will finally present a time model of the algorithm in terms of a number of attributes of the databases.

1. Introduction

Data mining in large sets of data is not a trivial occupation. Choosing one type of algorithm over another for a certain problem can mean the difference between getting good (for some definition of good) or no, or even inconclusive, results. Almost all of the methods used today in data mining require either structured data (so that clustering can easily be performed) or a concept hierarchy that envisions the inner structure of the data [7, 5, 17, 6].

Concept hierarchies are descriptors over data sets, in such a way that all records in the corresponding data sets can be described by the hierarchies. Concept hierarchies are typically used in retrieval systems [13, 12, 3] and for data mining [6]. Each facet or aspect of the records has a corresponding hierarchy, often in the form of a tree. An example of such a hierarchy that describes a C source code could be Imperative $\rightarrow$ C $\rightarrow$ ANSI for the language facet and Ordered $\rightarrow$ Tree $\rightarrow$ Binary for the data type facet. Constructing a concept hierarchy manually is very time consuming and relies on a thorough understanding of the actual data at hand. Construction of concept hierarchies can be automated (as in this work). However, there are still some doubts about the soundness of automatically generated facets/dimensions, especially since no exact rules for what is a “good” concept hierarchy exist today.

The remainder of the paper is organized as follows; The second section is the background to the work. The third section gives the definitions necessary for understanding the later parts of this paper. The fourth section is the algorithm, while the fifth contains an example of running the algorithm. The sixth section contains an analysis of the algorithm and the seventh shows relationships with previous work. The last two sections contains the discussion and the experiences that we have had with CHiC, respectively.

2. Background

A few years ago we were faced with a large data set consisting of two parts; One relational and one consisting of a set of keywords. Feeding only the relational part of the data set into data mining systems revealed no new information about it. We realized that the only way of getting more information from the data set was to use the given set of keywords as means of additional clustering. We were unable to find any already published algorithm for finding structures in discrete data, so we had to create one ourselves. The result of our work was a concept hierarchy constructor, called CHiC.

CHiC will automatically find concept hierarchies (or lattices, as the case might be) in sets of discrete data. Our definition of discrete data is either keyword based or otherwise enumerated data, e.g. the sample data in Figure 1. Our algorithm knows nothing about differences between numbers, whether integer or floating point, so it cannot be used to cluster numeric or continuous data.

CHiC has been used by major divisions of at least two multinational companies to generate concept hierarchies from their mixed mode databases.
3. Definitions

Given \( k_i \in \text{keywords} \) (the keywords in the system) and \( \sigma_k \in \mathcal{P}\text{records} \) (the set of records in the database that contain keyword \( k_i \)), we can define the concepts used in this work.

A **bucket** is a placeholder for a set of keywords.

Two or more keywords that always appear together are **keyword equivalent**, i.e.

\[
k_i = k_j \iff \sigma_{k_i} = \sigma_{k_j}, \quad \text{where } k_i \neq k_j.
\]

This implies that they may be **folded** into one keyword, i.e. the name of the latter is added as a synonym of the first and all references to the latter are removed or ignored. Such removed keywords are said to belong to \( k_i \)'s **family**, i.e.

\[
k_i = k_j \implies k_i, k_j \in \text{\textit{fam}}_i, \text{\textit{fam}}_j \in \mathcal{P}\text{keywords}.
\]

Formally, **subsumption** is defined as an implicit subset/superset relationship between the interpretations of the two concepts \cite{2}. This means that if a keyword \( k_i \) never appears in a record without \( k_j \) appearing, but not vice versa, then \( k_i \) is **subsumed** by \( k_j \).

\[
k_i < k_j \iff \sigma_{k_i} \subseteq \sigma_{k_j}
\]

A keyword \( k_i \) that is not subsumed by another keyword is said to be a **vector**. Vectors are used as the bases of the concept hierarchies.

4. The Algorithm

The algorithm (as outlined in Figure 2) given here is straightforward, although some of the implementation details can be anything but trivial to implement. The only requirements are that the keywords must have a consistent enumeration order (as given in equation 4) and that the database has already been created. The different parts of the algorithm are otherwise independent of different implementations of sets, etc.

\[
\forall x, y \in \text{keywords} : x \neq y \implies \text{ord}_x \neq \text{ord}_y
\]

(4)

Further information about each part of the algorithm are given in Section 4.7 (correctness of the algorithms) and Section 6 (cost estimations based on code complexity).

4.1. Read Index Data

All keywords should be given an enumeration order and the inverted file (see \cite{9}) corresponding to each keyword is read from disk. The following must hold after the step in the algorithm has been executed:

\[
\forall x \in \text{keywords} : \sigma_x = \text{inverted DB file}_x
\]

(5)

\[
\forall x \in \text{keywords} : |\sigma_x| = y \implies x \in \text{bucket}_y
\]

(6)

**Algorithm 1 (ReadInvertedFiles)** This step of the algorithm reads the index/inverted data files from the disc, or must generate them (at some run time cost) if they do not exist. In this algorithm description (and in the cost estimation in section 6.1) we expect them to be on disc before execution.

**proc** ReadInvertedFiles() **equiv**

Empty all buckets;

**foreach** \( x \in \text{keywords} \) **do**

\[
\sigma_x \leftarrow \text{inverted DB File}_x;
\]

\[y \leftarrow |\sigma_x|;
\]

\[
\text{bucket}_y \leftarrow \text{bucket}_y \cup \{x\};
\]

done
4.2. Find and Fold Keyword Families

The second step of the algorithm finds keywords that are keyword equivalent, adds the latter to the formers family and removes the latter (according to the keyword enumeration order) from further computations.

The following must always hold true for the algorithm used:

$$\forall x, y \in \text{keywords} \bullet ord_x < ord_y \land \sigma_x = \sigma_y \rightarrow x =_k y \quad (7)$$

where $y$ should be excluded from further computations.

**Algorithm 2 (FindAndFoldFamilies)** This step of the algorithm compares each keyword in a bucket with all other keywords with a later enumeration order in the same bucket, removing the latter if the two are keyword equivalent.

**proc** FindAndFoldFamilies() ⊑

```java
foreach bucket ∈ {buckets} do
    while ∃ x, y ∈ bucket • ord_x < ord_y ∧ σ_x = σ_y do
        fam_x ← fam_x ∪ fam_y;
        bucket ← bucket \ fam_y;
        keywords ← keywords \ fam_y;
    done
done
```

The reason for performing family folding before finding subsets is to prune the solution space as quickly as possible, since all keywords in a family will subsume/be subsumed by exactly the same keywords.

4.3. Find All Subsets and Subsumptions

This step of the algorithm finds where one keyword is subsumed by another, while ignoring folded keywords altogether. The effect of this step can be summarized as:

$$\forall x, y \in \text{keywords} \bullet x \neq y \land \sigma_x \subset \sigma_y \rightarrow x <_k y \quad (8)$$

**Algorithm 3 (FindSubsets)** This step of the algorithm will find all subsumptions in the database. CHIC works in decreasing bucket size order to further optimize the algorithm.

**proc** FindSubsets() ⊑

```java
foreach x ∈ keywords do
    foreach y ∈ keywords • |σ_x| < |σ_y| do
        if σ_x ⊂ σ_y then x <_k y;
    done
done
```

The reason for performing family folding before finding subsets is to prune the solution space as quickly as possible, since all keywords in a family will subsume/be subsumed by exactly the same keywords.
4.4. Find All Vector Nodes

The goal of the next step is to find all vectors (keywords that are not subsumed by any other keyword) in the given dataset, i.e.:

$$\text{vectors} = \{x \in \text{keywords} \mid \exists y \in \text{keywords} \cdot x \prec_k y\} \quad (9)$$

Algorithm 4 (FindVectorNodes) This step of the algorithm will add all found vectors in an array, sorted in decreasing cardinality order. The order is important for the heuristics of the next step.

```proc
FindVectorNodes() \equiv
vectors \leftarrow \text{empty array};
foreach x \in \text{keywords} do
    if \exists y \cdot y \in \text{keywords} \land x \prec_k y
        then insert k in vectors;
done
```

4.5. Group Vectors into Different Dimensions/Facets

Partition the set of vectors into as few partitions (vectors) as possible, while maintaining:

$$\forall x, y \in \text{vectors}_i, x \neq y \land \sigma_i \cap \sigma_y = \emptyset \quad (10)$$

Algorithm 5 (ColourVectorNodes) The algorithm is a greedy implementation based on heuristics that maintains equation 10. CHiC contains two different versions of keyword selection: The first version chooses the keyword with the highest cardinality that fits and the second selects the keyword that has the highest number of conflicts with other keywords first. The latter version is much more costly in terms of CPU time as can be seen in section 6.1.

```proc
ColourVectorNodes() \equiv
dim \leftarrow 0;
while vectors \neq \text{empty array} do
    dim \leftarrow dim + 1;
    while \exists x \in \text{vectors} \cdot \sigma_i \cap \bigcup_{y \in \text{vectors}_{\text{dim}}} \sigma_y = \emptyset do
        Remove k from vectors;
        Add k to vectors_{\text{dim}};
    done
done
maxDimension \leftarrow dim;
```

There exists yet another option to CHiC that changes the behavior of this step slightly. It is possible to ask the program to fill the dimensions as much as possible by reusing vectors from one dimension in later dimensions if there is no conflict with vectors already added in that dimension.

The easiest way to do this is to mark vectors with no conflict with any remaining vector in vectors to be used in current dimension and all that follows it. It is also very easy to go back and check already used vectors for conflicts when setting up latter dimensions.

The reason for using a heuristics based approach rather than trying all possible combinations is the extreme processing cost that it would incur. Finding the optimal colouring for even a small data base of, e.g., 36 vectors and 6 dimensions would yield 6^{35} combinations. Given a computer system that would test 50,000 combinations per second that would take in the order of 3.44 \times 10^{22} seconds, or roughly 10^{15} years. Our system yields an answer in a fraction of a second for the same data base, but we can not conclusively say that it is an optimal solution (neither in the number of dimensions, nor in the spanning of each dimension).

4.6. Fill Out the Concept Hierarchies/Lattices According to the Subsumptions Found in Step 4.3

This step generates a number of graphs $G_i = (N_i, E_i)$ where the nodes and the edges are given by equations 11 and 12.

$$N_i = \{x \in \text{keywords} \mid x \in \text{vectors}_i, \forall \exists y \in \text{vectors}_i \cdot x \prec_k y\} \quad (11)$$

$$E_i = \{(y,x) \mid x,y \in \text{keywords} \land x \prec_k y \land \exists z \cdot (x \prec_k z \land z \prec_k y)\} \quad (12)$$

Algorithm 6 (GenerateDimensions, Hierarchy Version) This algorithm generates concept hierarchies for each dimension (see Figure 4 for an example), i.e. the desired output for later data mining on the data.

```proc
GenerateDimensions() \equiv
for i \leftarrow 1 \text{ to } \text{maxDimension} do
    N_i \leftarrow \text{vectors}_i;
    foreach x \in \text{vectors}_i do
        N_i \leftarrow N_i \cup \{y \in \text{keywords} \mid y \prec_k x\};
    done
    E_i \leftarrow \{(y,x) \mid \forall \exists y \in N_i \cdot x \prec_k y\};
    Perform topological sort on $E_i$;
    done
```

Algorithm 7 (GenerateDimensions, Lattice Version) This step of the algorithm generates concept lattices for each dimension. This means that there might be more than one path between a vector and a given node further down in the graph.

---

1 One vector locked in the first dimension and with many permutations of combinations occurring more than once but in different dimensions.
4.7. Correctness of the Algorithm

Equation 4 implies that each keyword should have an unique order number. This number is used when selecting which keyword to fold and which to keep in Algorithm 2 (see Section 4.2), and it is also used by CHiC to prune the solution space in Algorithm 3 (see Section 4.3). Our solution to the ordering number is to give each keyword a number from 1024 and up in the order that they were seen when generating the database, which implies that a keyword cannot be subsumed by a keyword with a higher number unless they appeared for the first time in the same record.

Equation 5 is trivial. The implication of this equation is that the values have to be available at later stages, but how these are made available is not important for the algorithm.

Equation 6 shows the bucket sorting stage of the algorithm. The algorithm will work even if the sorting stage is removed, but at a much higher computational cost (see [1]).

The slightly rewritten form of Algorithm 1 seen below shows that it does indeed fulfill Equations 5 and 6.

```plaintext
proc GenerateDimensions() ≡
  for i ← 1 to maxDimension do
    Ni ← vectors;
    foreach x ∈ vectors do
      Ni ← Ni ∪ {y ∈ keywords | y <k x};
    done
  foreach x ∈ keywords do
    foreach y ∈ keywords • y <k x do
      foreach z ∈ keywords • z <k y do
        Remove z <k x
      done
    done
  for i ← 1 to maxDimension do
    Ei ← {(y,x) | ∀x, y ∈ Ni • x <k y};
  done
```

5. Example of Execution

If we take the dataset in Figure 1 as input to our algorithm we will get the following results:

1. Reading inverted files yields (this is subset of the results):
   \[ σ_{make} = \{01\} \]
   \[ σ_{text} = \{01, 02, 05, 07 - 13, 15\} \]
   \[ σ_{ELF} = \{03, 14, 16\} \]
   \[ σ_{executable} = \{03\} \]

   The buckets are also filled in with the following contents:

<table>
<thead>
<tr>
<th>bucket #</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>make, commands, Bourne, shell, script, executable, dynamically, linked, directory, International, language, auxiliary, file, ASCII, PostScript, conforming, at, level, 2.0, \TeX{}, transcript, 2e, English, lex, description</td>
</tr>
<tr>
<td>2</td>
<td>\LaTeX{}, document, C, program, relocatable</td>
</tr>
<tr>
<td>3</td>
<td>ELF, 32-bit, MSB, SPARC, version, 1, not, stripped</td>
</tr>
<tr>
<td>11</td>
<td>text</td>
</tr>
</tbody>
</table>
2. The following non-trivial keyword families are found in the data:

<table>
<thead>
<tr>
<th>family</th>
<th>folded keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>make</td>
<td>commands</td>
</tr>
<tr>
<td>Bourne</td>
<td>shell, script</td>
</tr>
<tr>
<td>ELF</td>
<td>32-bit, MSB, SPARC, version, 1, not, stripped</td>
</tr>
<tr>
<td>executable</td>
<td>dynamically, linked</td>
</tr>
<tr>
<td>International</td>
<td>language</td>
</tr>
<tr>
<td>auxiliary</td>
<td>file</td>
</tr>
<tr>
<td>PostScript</td>
<td>conforming, at, level, 2.0</td>
</tr>
<tr>
<td>\TeX{}</td>
<td>transcript</td>
</tr>
<tr>
<td>C</td>
<td>program</td>
</tr>
<tr>
<td>lex</td>
<td>description</td>
</tr>
</tbody>
</table>

3. Subsumptions found are: \(2e <_k \text{document}\), \(2e <_k \text{text}\), \(2e <_k \LaTeX{}, \text{ASCII} <_k \text{text}\), \(\text{Bourne} <_k \text{text}\), \(C <_k \text{text}\), \(\text{English} <_k \text{text}\), \(\text{International} <_k \text{text}\), \(\text{PostScript} <_k \text{document}\), \(\text{PostScript} <_k \text{text}\), \(\text{auxiliary} <_k \LaTeX{}, \text{document} <_k \text{text}\), \(\text{executable} <_k \text{ELF}\), \(\text{lex} <_k \text{text}\), \(\text{make} <_k \text{text}\), \(\text{relocatable} <_k \text{ELF}\), and, finally, \(\LaTeX{} <_k \text{text}\).

The results of steps 2 and 3 can be seen in Figure 3.

4. The vector families in the sample are \text{text}, \text{ELF}, \LaTeX{}, and \text{directory} with 11, 3, 2 and 1 instances respectively.

5. The first dimension consists of the vectors \text{text}, \text{ELF} and \text{directory}, while the second dimension has only one vector, \LaTeX{}.

6. Only 2 edges are removed in the topological sorting of this small dataset, \((\text{text}, \text{PostScript})\) and \((\text{text}, 2e)\). The end result of running the algorithm is seen in Figure 4.
6. Algorithm Analysis

The algorithm given in Section 4 is a step by step description of what should be done in order to get a consistent concept hierarchy for later data mining. The steps are to be performed one after another and some of the steps can easily be broken down in parts to be executed in parallel (e.g. all dimensions can be handled simultaneously in steps 6/7).

Two things have to be considered though, and that is the maximal cost of running the algorithm and the actual running times on real data. The complexity of the algorithm is calculated in Section 6.1 and the running time of the prototype is given in Section 6.2.

6.1. Cost Estimation

Given \( m \) (number of records) and \( n \) (number of keywords), we can give the following upper bounds of the algorithmic complexities:

\[
\begin{align*}
\text{proc ReadInvertedFiles()} & \equiv O(mn) \\
\text{proc FindAndFoldFamilies()} & \equiv O(mn) \\
\text{proc FindSubsets()} & \equiv O(mn) \\
\text{proc FindVectorNodes()} & \equiv O(mn) \\
\text{proc ColourVectorNodes()} & \equiv O(mn) \text{ or } O(mn^{2.5}) \\
\text{proc GenerateDimensions()} & \equiv O(mn + n^2) \text{ or } O(mn^2) \\
\text{total} & \equiv O(mn^2) \text{ or } O(mn^{2.5}).
\end{align*}
\]

This includes the cost of set operations that we estimate to be linear to the number of possible elements in a set, e.g. \( O(m) \) or \( O(n) \).

6.2. Actual Execution Times

Some of the data bases used for testing the algorithm are summarized in Table 1. The first three are prototypical for the discrete databases that we have encountered elsewhere. The last two are more academic in their nature, since they have unusually high number of keywords and vectors for their size.

The execution times of the prototype as given in Table 2 are from a Sun Blade-1000 workstation with a 750 MHz UltraSPARC-III processor, 8 MB level 2 cache and 512 MB of memory. CHIC was compiled using gcc version 2.95.2.

Modeling the execution time in terms of the values given in Table 1 and the time values in Table 2 (together with multiple other data sets) yielded a simple model for each step. We have not modeled the extra time added in step 5a/5b if reusing of vectors in later dimensions is used (as per Section 4.5), but since it is in the order of 0.2% of the execution time (for all but the smallest of our data bases, where it increases the time 0.7%) it is almost negligible.

\[
\begin{align*}
t_1 & \approx 1.77 \times 10^{-8} mn + 3.19 \times 10^{-4} n \\
t_2 & \approx 1.31 \times 10^{-10} mn^2 \\
t_{3\&4} & \approx 2.30 \times 10^{-10} mn^2 \\
t_5 & \approx 6.40 \times 10^{-10} mn^2 \\
t_6 & \approx 1.36 \times 10^{-6} \ln(m)n^{2.333} \\
t_6 & \approx 1.77 \times 10^{-7} m'n \\
t_{\text{fill+6a}} & \approx 2.44 \times 10^{-11} m'n^2 + 3.51 \times 10^{-6} m'n' \\
t_7 & \approx 4.04 \times 10^{-8} m'n + 1.79 \times 10^{-7} n^2
\end{align*}
\]
The number of vectors and subsets can often be hard to guess before calculations have been done, but we have found that \( n! \approx n^2/m \) and \( n! \approx \sqrt{mn} \) are useful approximations. They will usually yield results within 25\% from the correct value for almost all data bases\(^2\).

7. Related Work

The work reported in this article is closely related to semantic and knowledge indexing. It is a widely spread research area that includes such diverse topics as finding keywords for hierarchical summarization [10], knowledge acquisition tools [4] and automatic indexing of system commands based on their manual pages on a UNIX system [15].

Conceptual clustering systems [16] are also quite closely related to this work, but from the viewpoint of the data mining system rather than at the preprocessing stage.

Automatic generation of concept hierarchies as defined in [14] is related but works on a probabilistic, rather than a discrete, view of the data. The rule used to find subsumptions in that work is that \( x \) would subsume \( y \) if \( P(x|y) \geq 0.8, P(y|x) < 1 \). This means that transitivity will not work for subsumptions; Given three keywords \( x, y \) and \( z \) such that \( P(x|y) \geq 0.8, P(y|x) < 1, P(y|z) \geq 0.8, P(z|y) < 1 \) and \( P(x|z) \not< 0.8, P(z|x) < 1 \) would yield a very problematic state since \( z \) is subsumed by \( y \) which in turn is subsumed by \( x \) but \( z \) is not subsumed by \( x \). Whether this is acceptable or not is up to the user of the system.

The concept hierarchies generated by our algorithm have strong resemblances to the feature-oriented classification trees used by Salton [13], Prieto-Díaz [12] and Börstler [3]. We believe that this is not a coincident, and that such structures evolve naturally when working with closely related pieces of data like software assets.

8. Discussion

One method of speeding up the average case while keeping the worst case the same has been used in a newer version of CHIC. The keywords are sorted (using bucket sort) in decreasing cardinality order in step 1. This speeds up the algorithm considerably for the average case, since keyword equality implies the same cardinality and subset a lower cardinality. This implies that the number of keywords to be checked in steps 2 & 3 was decreased significantly (with factors of approximately \( 1/\log(n) \) and at least \( 1/2 \), respectively).

The grouping of vectors into dimensions uses a heuristic algorithm rather than a best fit or even back track based graph colouring algorithm [11, 8]. We have experimented with multiple versions of backtracking algorithms, but were not able to find any that were both fast enough and gave sufficiently better results than our heuristic algorithm.

Some of the non-vector keywords are subsumed by more than one vector, thereby generating a lattice structure with rather complex properties. This is both a weakness and a strength: A weakness since it will generate the same set of vertices more than once and a strength since such keywords are excellent candidates for finding correlations and interesting data points in later Data Mining operations. It would be rather simple to update CHIC so that it excludes all keywords after they have been used for the first time in a dimension if that would fit a certain problem.

\( \begin{align*}
\text{country} & \quad \text{province_or_state} \\
\text{city} & \quad \text{month} \\
\text{street} & \quad \text{day} \\
\text{year} & \quad \text{quarter} \\
\text{week} & \quad \text{month}
\end{align*} \)

\[ \text{Figure 5. Hierarchical (a) and lattice (b) structures of attributes.} \]

Previous version of CHIC generated a conceptual hierarchy for each dimension even though the data indicated that a lattice would be more fitting (see Figure 5 for examples of both). The reason for this was the topological sorting done in the last step of GenerateDimensions (Algorithm 6). Removing the correct vertices from the set to generate a lattice was fortunately not that hard (see Algorithm 7); Subsumption is a transitive function, e.g. if keyword \( a \) subsumes both \( b \) and \( c \) while \( b \) subsumes \( c \) then the vertex \((a,c)\) can safely be removed from the resulting set. This technique proved to be very useful (especially when the keywords are already in decreasing cardinality order after Algorithm 1 is done), but requires that the chosen Data Mining system can handle lattices rather than hierarchies.

The time required for generating the concept hierarchy for our biggest dataset seems rather high (just over 80 minutes of computation, see Section 6.2), but since this is done
only once for each dataset before doing data mining we believe it to be satisfactory anyway. One way of speeding this up is to create an extra dataset that contains no duplicate records and use this dataset when constructing the concept hierarchy. The time for concept hierarchy generation on our large dataset (with 38187 unique records) falls to around 11 minutes execution time. Data mining is often done on static databases, e.g. data warehouses, so the cost of concept hierarchy generation should generally be amortized over the number of times that the resulting data is later used. Our approach should probably not be used in a constantly evolving database with major upgrades going on simultaneously.

9. Experiences

All work so far on the prototype have been fruitful in the form of data relations both directly from concept hierarchy generation and also from later data mining. The concept hierarchies generated have generally been of good quality (i.e. logically connected keywords tend to be in the same dimension or even subsumed, etc.) with some minor glitches and we hope to see future use of our algorithm in the works of others.

The divisions of two multinational companies that have used CHIC have been very pleased with the results obtained from the program, and have found other, more novel, uses for it as well. One of the comments that we have received is that “The only other option [available to us] would have been to hire a very expensive expert in the field, and she would probably have come up with something remarkably similar to what we get from CHIC.”

References