Distributing deductive knowledge bases by clustering

MIRCEA PREDA
University of Craiova
Department of Computer Science
Al. I. Cuza 13, 200585 Craiova
ROMANIA
mpreda@acm.org

ANA-MARIA MIREA
University of Craiova
Department of Computer Science
Al. I. Cuza 13, 200585 Craiova
ROMANIA
mireeaq@yahoo.com

Abstract: Distributing knowledge bases is an important problem from at least two reasons: knowledge bases become larger over time and also it is common that a knowledge base to be built by several different cooperating agents. The paper proposes a new method for finding a distribution of a knowledge base that stresses on the locality principle. By locality, it is understand that elements usually involved together in reasoning should be placed if possible into the same node. The distribution is obtained by grouping the similar elements from the knowledge base by a hierarchical clustering algorithm.

Key–Words: Distributed databases, deductive databases, clustering methods, hierarchical systems, semantic networks

1 Introduction

Distribution is a common feature for the nowadays knowledge bases. Not only that the knowledge bases grow increasingly complex and become difficult to manage by the current hardware but also cooperation and merging between the several knowledge bases from different organizations is usual. When distribution is required, the main problem that is raised is how to divide the available pieces of information on several computing nodes. The purpose of the article is twofold: first metrics to measure the quality of a distribution are considered. The main quality criterion is locality: the pieces of knowledge needed by an inference process should be usually located together. After that, algorithms for automated distributing a knowledge base on a set of computing nodes are designed and evaluated. The knowledge structure and the employed inference methods are paramount for a successful distribution. Two scenarios are considered starting with minimal assumption over the knowledge base and ending with inheritance knowledge bases where the structure and the inference mechanisms are fully known.

Numerous research papers address the topic of the distributed knowledge representation, but few discuss the problem considering the semantics of the database, some major ones being enlisted below. [7] mathematically describes the knowledge base distribution problem in the semantic network context. In [2], an architecture is presented that allows sharing and integrating several knowledge bases. Meta knowledge is used to coordinate the local systems in a object oriented framework. [1] uses a comprehensive cost model of the data transfer that occurs when a query is processed to develop an iterative distribution algorithm that provides near optimal solutions. In [4] and [5], distributed knowledge bases are considered in a mobile environment with focus on data replication and conflict resolution.

The rest of the article continues as follows. In the next section, a distributed knowledge base scenario with minimal assumptions is considered. The scenario proposes metrics to evaluate the quality of a distribution based on locality principle and a clustering algorithm to automatically build near optimal distributions. Section 4 extends the study to knowledge bases that uses inheritance as main inference method. Quality distribution metrics are redefined taking account the inheritance graph. It is proved that the clustering algorithm introduced in previous section performs well in the new conditions. The article concludes with a summary of the newly introduced proposals and with directions for further work.

2 Distributing generic knowledge bases

Let \( K \) be a knowledge base composed from a set of objects. For the next two sections, all considered knowledge bases will be sets of objects. Let \( \mathcal{A} \) be the set of questions that can be submitted to the knowledge base \( K \) and \( A \in \mathcal{A} \) an interrogation.
The optimal distribution degree:

\[ GRAD^* = \sum_{A} \alpha(A) \]  

The optimal distribution degree is particularly attained by the trivial split \( I = \{k\} \). This split have a single subset \( K \). But to put all objects in only one node is not a practical decision for all situations.

3 Finding an optimal split for a knowledge Base

We suppose that the number of sets in partition \( m \) is given. Our purpose is to find a split with the smallest distribution degree.

We try to find \( I^* \in Part(K), |I^*| = m \) so that:

\[ GRAD(I^*) = \min\{GRAD(I) | I \in Part(K), |I| = m\} \]  

If we try to use the brute force, the approach that generates all splits has exponential complexity. Let us try instead to use a hierarchical clustering algorithm to find an approximation of the optimal split.

3.1 Hierarchical clustering

Many unsupervised learning methods use a measure of similarity between different patterns to obtain the grouping in clusters. When the features of the patterns are numeric, the distance measure can be ordinary Euclidian distance between two points in an \( n \)-dimensional space. This is an iterative method based on distance. Suppose we have a set, II of unlabeled training patterns. A simple hierarchical clustering algorithm (see [3]) functions as follows. First it is computed the Euclidian distance between all pairs of patterns in II. Suppose the smallest distance is between patterns \( X_i \) and \( X_j \), these patterns are collected into a cluster \( C \) that replaces them. The process continues iteratively until the clusters tree is obtained.

In order to use a clustering algorithm we need a distance between the objects from knowledge base. This distance will be defined regarding the interrogations from \( A \).

Figure 1: Clustering concepts to obtain a split of a knowledge base. The concepts are considered similar if they are used to answer to almost same questions.

Let us consider \( A' \) a set of training interrogations. For each \( A \in A' \), \( K(A) \) is known.
\textbf{Definition 6} The relationship (similarity) degree between two concepts Rel\((k_1, k_2)\) represents the number of questions that needs for answering either both concepts or none of them. \(\text{Rel} : K \times K \rightarrow \mathbb{N}\)

\[
\text{Rel}(k_1, k_2) = |\{A| k_1 \in K(A), k_2 \in K(A)\}| + |\{A| k_1 \notin K(A), k_2 \notin K(A)\}|
\] (7)

\textbf{Definition 7} The difference between two concepts \(\text{Diff}(k_1, k_2)\) represents the number of questions that need one concept, but do not need the other. \(\text{Diff} : K \times K \rightarrow \mathbb{N}\)

\[
\text{Diff}(k_1, k_2) = |\text{SDiff}(k_1, k_2)| \quad \text{where}
\]

\[
\text{SDiff}(k_1, k_2) = \{A| k_1 \in K(A), k_2 \notin K(A)\} \cup \{A| k_1 \notin K(A), k_2 \in K(A)\}
\] (8)

Another definition for \(\text{SDiff}(k_1, k_2)\) is

\[
\text{SDiff}(k_1, k_2) = \{A| \{k_1, k_2\} \notin K(A) \text{ and } \{k_1, k_2\} \cap K(A) \neq \emptyset\}
\]

The above functions can be converted to the \([0,1]\) range.

\[
\overline{\text{Rel}} : K \times K \rightarrow [0,1], \overline{\text{Diff}} : K \times K \rightarrow [0,1]
\]

\[
\overline{\text{Rel}}(k_1, k_2) = \frac{\text{Rel}(k_1, k_2)}{|A'|},
\]

\[
\overline{\text{Diff}}(k_1, k_2) = \frac{\text{Diff}(k_1, k_2)}{|A'|}.
\]

\textbf{The Properties of} \(\overline{\text{Rel}} \text{ and } \overline{\text{Diff}}\)

- \(\overline{\text{Rel}}(k, k) = 1\), a concept is the most alike himself.
- \(\overline{\text{Rel}}(k_1, k_2) = \overline{\text{Rel}}(k_2, k_1)\), \(\overline{\text{Diff}}(k_1, k_2) = \overline{\text{Diff}}(k_2, k_1)\), symmetry.
- \(\overline{\text{Diff}}(k, k) = 0\), there is no difference between a concept and himself.

\textbf{Proposition 8} \(\overline{\text{Diff}}\) is a distance over the concepts from \(K\).

\textbf{Proof} a) \(\overline{\text{Diff}}(k, k) = 0\), \(\overline{\text{Diff}}(k_1, k_2) \geq 0\).

b) The triangle rule

\[
\overline{\text{Diff}}(k_1, k_2) \leq \overline{\text{Diff}}(k_1, k_3) + \overline{\text{Diff}}(k_3, k_2).
\]

Using the definition of \(\overline{\text{Diff}}\),

\[
\overline{\text{Diff}}(k_1, k_2) = \frac{\overline{\text{Diff}}(k_1, k_3)}{|A'|} + \frac{\overline{\text{Diff}}(k_3, k_2)}{|A'|}.
\]

We will prove that \(\text{SDiff}(k_1, k_2) \subseteq \text{SDiff}(k_1, k_3) \cup \text{SDiff}(k_3, k_2)\). Let \(A \in \text{SDiff}(k_1, k_2)\). There are the following possible cases:

1. \(k_1 \in K(A), k_2 \notin K(A), k_3 \in K(A)\), then \(A \in \text{SDiff}(k_3, k_2)\).
2. \(k_1 \in K(A), k_2 \notin K(A), k_3 \notin K(A)\), then \(A \in \text{SDiff}(k_1, k_3)\).
3. \(k_1 \notin K(A), k_2 \in K(A), k_3 \in K(A)\), then \(A \in \text{SDiff}(k_1, k_3)\).
4. \(k_1 \notin K(A), k_2 \in K(A), k_3 \notin K(A)\), then \(A \in \text{SDiff}(k_3, k_2)\).

Consequently,

\[
|\text{SDiff}(k_1, k_2)| \leq |\text{SDiff}(k_1, k_3)| + |\text{SDiff}(k_3, k_2)|.
\]

The last result proves the triangle inequality.

Based on the above result, \(\overline{\text{Diff}}\) can be used as a distance for a concept clustering algorithm.

Another possibility to consider is to attach to every object from the knowledge base an array of real numbers that will describe the object. Each real number from array will correspond to a characteristic of the object. In our settings, the requests from the training set will be considered to be characteristics and an object will receive the value 1 for a characteristic (request) if it is used to answer at the request and 0 otherwise.

Let us consider the function \(f : K \rightarrow \mathbb{R}^s\) where \(s\) is the cardinal of the requests training set and

\[
f(o) = (y_1, y_2, ..., y_s).
\]

If \(A_i\) is the \(i\)th member of the requests training set, then

\[
y_i = \begin{cases} 1 & \text{if } o \in K(A_i) \\ 0 & \text{if } o \notin K(A_i) \end{cases}
\]

\textbf{Proposition 9} The function \(d' : K \times K \rightarrow \mathbb{R}_+\)

\[
d'(o_1, o_2) = \sqrt{\sum_{i=1}^{s} (y_{i_1} - y_{i_2})^2}
\]

where \(f(o_1) = (y_{i_1}^{o_1}, y_{i_2}^{o_2}, ..., y_{i_s}^{o_1})\) and \(f(o_2) = (y_{i_1}^{o_2}, y_{i_2}^{o_2}, ..., y_{i_s}^{o_2})\) is a distance.
Proof The properties of distances can be easily verified by $d'$.

Example 10 Let $K = \{o_1, o_2, ..., o_{17}\}$ a knowledge base and $A' = \{A_1, A_2, ..., A_7\}$ a training set of interrogations where

\[
\begin{align*}
K(A_1) &= \{o_5, o_4, o_2, o_1\} \\
K(A_2) &= \{o_6, o_4, o_3, o_1\} \\
K(A_3) &= \{o_9, o_8, o_3, o_1\} \\
K(A_4) &= \{o_9, o_8, o_7\} \\
K(A_5) &= \{o_{11}, o_{10}, o_7\} \\
K(A_6) &= \{o_{17}, o_{14}, o_{13}, o_{12}\} \\
K(A_7) &= \{o_{16}, o_{15}, o_{12}\}
\end{align*}
\]

In these settings, the function $f : K \rightarrow \mathbb{R}^7$ that maps objects into arrays of numbers has the values enlisted in the below table:

\[
\begin{align*}
f(o_1) &= (1, 1, 1, 0, 0, 0, 0) \\
f(o_2) &= (1, 0, 0, 0, 0, 0, 0) \\
f(o_3) &= (0, 1, 1, 0, 0, 0, 0) \\
f(o_4) &= (1, 1, 0, 0, 0, 0, 0) \\
f(o_5) &= (1, 0, 0, 0, 0, 0, 0) \\
f(o_6) &= (0, 1, 0, 0, 0, 0, 0) \\
f(o_7) &= (0, 0, 1, 1, 0, 0, 0) \\
f(o_8) &= (0, 0, 1, 1, 0, 0, 0) \\
f(o_9) &= (0, 0, 1, 1, 0, 0, 0) \\
f(o_{10}) &= (0, 0, 0, 0, 1, 0, 0) \\
f(o_{11}) &= (0, 0, 0, 0, 1, 0, 0) \\
f(o_{12}) &= (0, 0, 0, 0, 1, 1, 0) \\
f(o_{13}) &= (0, 0, 0, 0, 1, 0, 0) \\
f(o_{14}) &= (0, 0, 0, 0, 1, 0, 0) \\
f(o_{15}) &= (0, 0, 0, 0, 0, 0, 1) \\
f(o_{16}) &= (0, 0, 0, 0, 0, 0, 1) \\
f(o_{17}) &= (0, 0, 0, 0, 0, 0, 1)
\end{align*}
\]

By applying a hierarchical clustering algorithm, the cluster organization from figure 2 is obtained.

Let us consider the split $I_1 = \{a, b\}$ where the sets $a$ and $b$ are defined in the figure 2. Then $\text{GRAD}(I_1, A_1) = \ldots = \text{GRAD}(I_1, A_7) = 1$. In accordance with the definition 2, $I_1$ is optimum for interrogations regarding the training interrogation set $A'$. Let us analyze now $I_2 = \{c, d, b\}$. Then $\text{GRAD}(I_2, A_1) = \text{GRAD}(I_2, A_2) = \text{GRAD}(I_2, A_4) = \ldots = \text{GRAD}(I_2, A_7) = 1$ and $\text{GRAD}(I_2, A_3) = 2$. If the training interrogations have equal importance coefficients ($\frac{1}{7}$), then $\text{GRAD}(I_2) = \frac{5}{7}$ and $\text{GRAD}(I_1) = \frac{7}{7} = 1$.

Figure 2: The splits resulted by applying a hierarchical clustering algorithm to the knowledge base $K$ and training interrogation set $A$.

4 Distributing knowledge bases that use inheritance as primary reasoning method

Let us consider $K$ a knowledge base, $K = (\text{Obj}, \rho)$ where $\text{Obj}$ is the set of objects from knowledge base and $\rho \subseteq \text{Obj} \times \text{Obj}$ represents the inheritance relation. $(x, y) \in \rho$ indicates that $x$ is a parent of $y$.

To distribute $K$ on $m$ computational nodes, we should find a partition $I = \{I_1, ..., I_m\}$ such that

$$\bigcup_{i=1}^{m} I_i = K, I_i \cap I_j = \emptyset.$$ 

Before to devise an algorithm that can automatically distribute the knowledge base $K$ on $m$ nodes, we should establish what means an acceptable distribution. The main criterion will be the locality of the reasoning process, if an object $x$ is placed in a set $I_i$ from the partition, then it is recommendable that the ancestors of $x$ to be also placed in $I_i$. Let us denote by $\text{MissP}(x)$ the number of $x$'s ancestors that are missing from $I_i$. In these conditions, the lack of quality of a partition $I$ can be indicated by the value $\sum_{x \in \text{Obj}} \text{MissP}(x)$. The degree of cohesion for the partition elements $I_i$ can be also significant. To formalize the notion of cohesion, let us define, for a partition element $I_i$, two values:

- $\text{lack}\_\text{cohs}(I_i)$ represents the number of pair of objects $x, y \in I_i$ such that $x$ and $y$ have no common parent.
\[ \text{cohs}(I_i) \text{ represents the number of pair of objects } \]
\[ x, y \in I_i \text{ that have a common parent.} \]

Obvious, between these two values there is the following relation: \( \text{cohs}(I_i) = |\text{Obj}|^2 - \text{lack_cohs}(I_i) \).

From these considerations, the following measure of lack of quality can be devised for a partition:

\[ \text{MissQ} : \text{Part}(K) \to \mathbb{N} \]

\[ \text{MissQ}(I) = \sum_{x \in \text{Obj}} \text{MissP}(x) + \sum_{i=1}^{m} \text{lack_cohs}(I_i). \]

As in the previous section, we intend to use a clustering algorithm to automatically split the knowledge base. But, in order to do that, a measure of the distance between two objects should be introduced. Our aim is to put close objects into same set and distant objects into different sets from partition. The distance should be based on the inheritance relation because the inheritance usage was the only supposition made about the knowledge base.

Let \( \text{Initial}(K) \subseteq \text{Obj} \) be the set of objects from \( K \) that have no parents. We intend to use these objects as reference points to compare the elements of the knowledge base.

Let \( sp \in \text{Initial}(K) \) and \( o \in \text{Obj} \). The distance between \( sp \) and \( o \) is defined as

\[ d_{sp}(o) = \begin{cases} \text{the distance in the inheritance graph} & \text{between } o \text{ and } sp \\ \text{if } sp \text{ is an ancestor of } o, & +\infty \text{ otherwise.} \end{cases} \]

The distance between two objects \( o_1 \) and \( o_2 \) regarding to \( sp \) can be defined as

\[ d_{sp}(o_1, o_2) = |d_{sp}(o_1) - d_{sp}(o_2)| \]

Here, it is considered that \( |+\infty-+\infty|=0 \). With these developments, the following result can be proved:

**Proposition 11** The function \( d : \text{Obj} \times \text{Obj} \to \mathbb{R}_+ \)

\[ d(o_1, o_2) = \sum_{sp \in \text{Initial}(K)} d_{sp}(o_1, o_2) \]

is a distance.

**Proof** a) \( d(o, o) = 0, \forall o \in \text{Obj} \). \( d(o_1, o_2) \geq 0, \forall o_1, o_2 \in \text{Obj} \). These two affirmations are true from definition of the function \( d \).
b) Triangle inequality. \( \forall o_1, o_2, o_3 \in \text{Obj} \) then \( d(o_1, o_2) \leq d(o_1, o_3) + d(o_3, o_2) \).

Let us prove that for \( \forall sp \in \text{Initial}(K) \) then

\[ d_{sp}(o_1, o_2) \leq d_{sp}(o_1, o_3) + d_{sp}(o_3, o_2). \]

\[ |d_{sp}(o_1) - d_{sp}(o_2)| \leq |d_{sp}(o_1) - d_{sp}(o_3)| + |d_{sp}(o_3) - d_{sp}(o_2)|. \]

This is equivalent with:

\[ |d_{sp}(o_1) - d_{sp}(o_2)| \leq |d_{sp}(o_1) - d_{sp}(o_3)| + |d_{sp}(o_3) - d_{sp}(o_2)|. \]

Let us denote \( A = d_{sp}(o_1) - d_{sp}(o_3) \) and \( B = d_{sp}(o_3) - d_{sp}(o_2) \). Then, the last relation can be rewritten as \( |A + B| \leq |A| + |B| \), which is true.

Another approach to compare objects from knowledge base is to attach an array of real numbers to each object. In this case, the distance between two objects will be distance between the attached arrays of real numbers. Let us consider the function

\[ f : \text{Obj} \to \mathbb{R}^s, f(o) = (y_1, y_2, \ldots, y_s) \]

where \( s = |\text{Initial}(K)| \) and \( y_i = d_{sp_i}(o) \) where \( sp_i \) is the \( i \)th member of \( \text{Initial}(K) \).

In these conditions, the distance between two objects is the euclidian distance between their attached arrays.

**Proposition 12** The function \( d' : \text{Obj} \times \text{Obj} \to \mathbb{R}_+ \)

\[ d'(o_1, o_2) = \sqrt{\sum_{i=1}^{s} (y_{si} - y_{oi})^2} \]

is a distance, where \( f(o_1) = (y_{1o_1}, y_{2o_1}, \ldots, y_{so_1}) \) and \( f(o_2) = (y_{1o_2}, y_{2o_2}, \ldots, y_{so_2}) \).

**Proof** The properties of the distances can be easily verified by \( d' \).

**Example 13** Let us consider the hierarchical knowledge base from figure 3. This knowledge base was used in [7] to illustrate the distribution of hierarchical knowledge. In these conditions, \( \text{Initial}(K) = \{o_1, o_7, o_{12}\} \).

![Image 3](https://example.com/image3.png)

**Figure 3:** The inheritance graph of the hierarchical knowledge base \( K \).
The function $f : \text{Obj} \rightarrow \mathbb{R}^3$ that attaches arrays of real numbers to the objects from $K$ is described by the following table:

<table>
<thead>
<tr>
<th>$o_1$</th>
<th>$o_7$</th>
<th>$o_{17}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(o_1)$</td>
<td>$0^+\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$f(o_2)$</td>
<td>$1^+\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$f(o_3)$</td>
<td>$1^+\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$f(o_4)$</td>
<td>$2^+\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$f(o_5)$</td>
<td>$3^+\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$f(o_6)$</td>
<td>$3^+\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$f(o_7)$</td>
<td>$+\infty$</td>
<td>$0^+\infty$</td>
</tr>
<tr>
<td>$f(o_8)$</td>
<td>$2$</td>
<td>$1^+\infty$</td>
</tr>
<tr>
<td>$f(o_9)$</td>
<td>$3$</td>
<td>$2^+\infty$</td>
</tr>
<tr>
<td>$f(o_{10})$</td>
<td>$+\infty$</td>
<td>$1^+\infty$</td>
</tr>
<tr>
<td>$f(o_{11})$</td>
<td>$+\infty$</td>
<td>$2^+\infty$</td>
</tr>
<tr>
<td>$f(o_{12})$</td>
<td>$+\infty$</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>$f(o_{13})$</td>
<td>$+\infty$</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>$f(o_{14})$</td>
<td>$+\infty$</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>$f(o_{15})$</td>
<td>$+\infty$</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>$f(o_{16})$</td>
<td>$+\infty$</td>
<td>$+\infty$</td>
</tr>
<tr>
<td>$f(o_{17})$</td>
<td>$+\infty$</td>
<td>$+\infty$</td>
</tr>
</tbody>
</table>

By applying a hierarchical clustering algorithm to the data from the above table, the splits from figure 4 are obtained. If the split $I_1 = \{a, b\}$ is considered, this split closely follows the inheritance graph of the knowledge base. The set $a$ represents the left tree from the inheritance graph and the set $b$ the right tree of the same graph. $MissQ(I_1) = 0 + lack\_cohs(a) + lack\_cohs(b) = 0 + 18 + 0 = 0$.

Figure 4: The splits resulted by applying a hierarchical clustering algorithm to the hierarchical knowledge base $K$ from figure 3.

5 Conclusion

The purpose of the paper was to study the possibilities of automatic distribution of knowledge on several computing nodes regarding the locality criterion of the resources used in the inferences. There are provided distribution methods both for the general case of knowledge bases without any other assumptions and for the special case of a knowledge base that uses inheritance as primary reasoning method. Both methods were exemplified using examples also employed by other distributed knowledge bases papers. Future developments include usage of the slots (attributes) from the inheritance knowledge base to develop new distribution criteria. We are also interested to verify the applicability of the proposed methods in the context of some practical applications from e-learning and Web recommendations domains. Details about the application possibilities on these topics can be found in [6].

References:


