Semantically-guided Evolutionary Knowledge Discovery from Texts

John A. Atkinson-Abutridy

Doctor of Philosophy

Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh

2003

Abstract

This thesis proposes a new approach for structured knowledge discovery from texts which considers both the mining process itself, the evaluation of this knowledge by the model, and the human assessment of the quality of the outcome.

This is achieved by integrating Natural-Language technology and Genetic Algorithms to produce explanatory novel hypotheses. Natural-Language techniques are specifically used to extract genre-based information from text documents. Additional semantic and rhetorical information for generating training data and for feeding a semi-structured Latent Semantic Analysis process is also captured.

The discovery process is modeled by a semantically-guided Genetic Algorithm which uses training data to guide the search and optimization process. A number of novel criteria to evaluate the quality of the new knowledge are proposed. Consequently, new genetic operations suitable for text mining are designed, and techniques for Evolutionary Multi-Objective Optimization are adapted for the model to trade off between different criteria in the hypotheses.

Domain experts were used in an experiment to assess the quality of the hypotheses produced by the model so as to establish their effectiveness in terms of novel and interesting knowledge. The assessment showed encouraging results for the discovered knowledge and for the correlation between the model and the human opinions.

Acknowledgements

I wish to express my sincere gratitude for my supervisors Professor Dr. Chris Mellish of the *Institute for Communicating and Collaborative Systems* and Dr. Stuart Aitken of the *Centre for Intelligent Systems and their Applications* for their insightful discussions to deepen and structure my understanding of this research theme.

I am indebted to the Ministerio de Planificación from the Chilean government (MIDEPLAN) for providing me with the scholarship to carry out this research.

I would also like to thank Dr. Alex Lascarides of the *Human Communication Research Centre* for her helpful comments and suggestions on my early ideas and the drafts of the thesis and Dr. Peter Wiemer-Hastings for his suggestions on LSA and for providing me with key data and tools to use on my research.

My special thanks are due to Dr. John Levine of the *Centre for Intelligent Systems and their Applications* and Dr. John Tait of the *School of Computing And Technology, University of Sunderland* for reading the draft of the thesis and making a number of suggestions.

At every step in preparing this thesis, my wife, Anita Ferreira-Cabrera, has always supported and encouraged me to tackle this challenge. To her: thank you for your patience, favors and all the other things that make it si worthwhile to know you.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*John A. Atkinson-Abutridy*)

List of Figures

List of Tables

Table of Contents

Chapter 1

Introduction

Like gold, information is both an object of desire and a medium of exchange. Also like gold, it is rarely found just lying about. It must be mined, and as it stands, a large portion of the world's electronic information exists as numerical data. Data Mining technology (Han and Kamber, 2001) can be used for the purpose of extracting "nuggets" from well-structured collections that exist in relational databases and datawarehouses (Abrams, 2002; Polanco and Francois, 1998). However, 80% of this portion exists as text and is rarely looked at: letters from customers, email correspondence, technical documentation, contracts, patents, etc.

An important problem is that information in this unstructured form is not readily accessible to be used by computers. This has been written for human readers and requires, when feasible, some natural language interpretation. Although full processing is still out of reach with current technology (Jurafsky and Martin, 2000; Manning and Schutze, 1999), there are tools using basic pattern recognition techniques and heuristics that are capable of extracting valuable information from free text based on the elements contained in it (e.g., keywords). This technology is usually referred to as **Text Mining**, and aims at discovering unseen and interesting patterns in textual databases (Feldman and Dagan, 1995; Hearst, 2000).

These discoveries are useless unless they contribute valuable knowledge for users who make strategic decisions (i.e., managers, scientists, businessmen). This leads then to a complicated activity referred to as **Knowledge Discovery from Texts** (KDT) which, like *Knowledge Discovery from Databases* (KDD), correspond to *"the non-*

trivial process of identifying valid, novel, useful and understandable patterns in data" (Fayyad et al., 1996).

With this in mind, it turns out that many applications have been misleadingly called TM or KDT. Traditionally, TM has been perceived as an Information Retrieval (IR) related application, a problem in Natural-Language Processing (NLP) or an Information Extraction task. Some researchers and industrialists in these areas have claimed to be doing text mining. By no means can it be said that these problems are easy ones, and they can even be part of a major KDT system. However, for each of them, it would be better being called by its own name. In the context of KDT, the knowledge extracted has to be grounded in the real world (i.e., users), and will modify the behavior of a human or mechanical agent. Unlike NLP, KDT does not aim at doing text understanding, but at discovering unsuspected relations holding in a body of texts. In addition, discovery in KDT means that induction is used, while in NLP research has not generally interested in inductive processing, except in applying machine learning techniques to NLP.

A general scheme to make this distinction clear is shown in figure 1.1 where the different levels of perceptions of the TM problem in terms of the information captured are highlighted. The first level (IR) can be seen as a first step in retrieving relevant, but known, information from text databases which helps users to find documents that satisfy their information needs. The second level consists of extracting meaningful information from the documents (i.e., via Information Extraction techniques) in order to find relevant facts (Gaizauskas and Wilks, 1997). Although more specific, this phase tends to extract relevant information within the text that cannot be identified by the first level but at the same time, is of a form already known by the human authors. However, since this is a time-consuming task in terms of analysing huge amounts of data by hand, automatic techniques like these are provided. A last level, knowledge discovery, aims at performing intelligent analysis of the extracted facts in order to discover unseen, non-trivial and interesting patterns in them (e.g., predictive rules). Note that the source documents can be used for each level in an independent way, which is the case for many current stand-alone applications that use their own corpus of text documents (i.e., the raw set of original document may directly feed the process of discovery with

no need of previous processing). From this perspective, it is worth wondering what may a typical outcome of a Text Mining activity look like, and which limitations (if any) could this have.

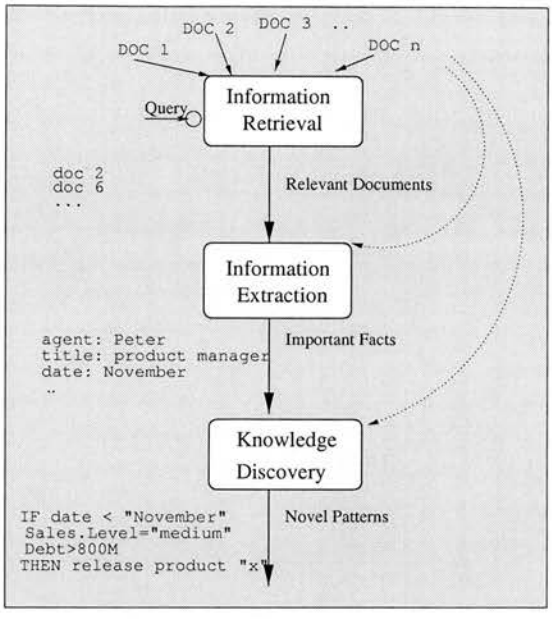

Figure 1.1: Typical Levels of Information Captured by different Applications

In order to address these issues, consider the following situation: Suppose that a knowledge worker (*Intelligence Analyst*) works for the company *"TopComputers"* (computer manufacturer and dealer) based in *Germany*. Part of his work is to make strategic decisions by investigating new trends in the market, watching the competitors, finding and exploring new possibilities for products, etc.

For this, he already knows that a valuable source of knowledge that should definitely be analysed is the company's huge repository of text documents containing summarized information regarding customer surveys, reports of the competitors, product descriptions, customers feedback, etc.

As the whole process of "mining" and evaluating the discovered knowledge is a very complex and extremely time-consuming task, he may be advised to obtain a text mining tool. Specifically, the kind of tool which performs cluster analysis seems to be appropriate as its results are easy to visualize. Besides, this kind of output is typical

for the state-of-the-art clustering-based TM systems.

Once his company's text documents have been cleaned up properly, the tool is applied to perform the analysis itself. What this does is basically to assume that the texts can be represented as a set of keywords. From these, the mining via statistical techniques is performed by creating and visualizing clusters of related concepts. Some of the tool's outcome produced from its text data can be seen in figure 1.2.

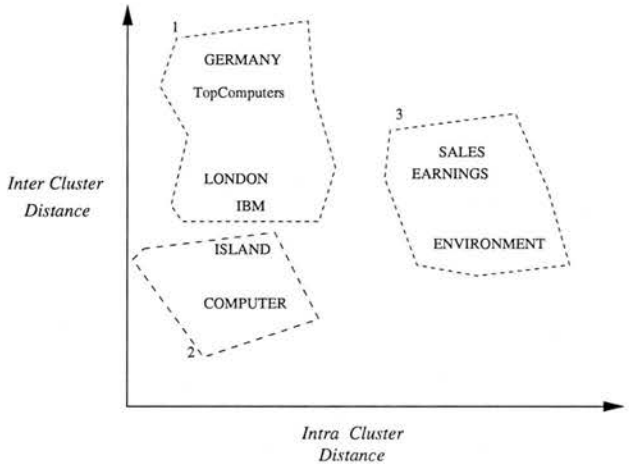

Figure 1.2: Cluster Analysis for "TopComputers"

Here, three relevant clusters of concepts have been produced in a two-dimensional map. The horizontal axis and the vertical axis represent the distance between concepts inside a cluster (internal associations) and the distance between clusters (external associations), respectively. Note that the map shows not only the clusters but also provides information on how close/distant the concepts are (the shorter the distance, the closer are the concepts). In addition, each concept in a cluster is linked to the set of documents where the concept appears.

From this conceptual map, the analyst wishes to mine for plausible and valuable knowledge in the form of, for example, unseen relations. Thus, he may want to investigate the association between two concepts, which from now on will be referred to as the **Target Concepts**. Some of the "discoveries" involving this kind of concepts are worth observing:

- There seems to be a close relation between *TopComputers* and *Germany*. The

analyst already knows this so it does not represent anything interesting to him. However, these concepts have been clustered in the same group as *London* and *IBM*. While IBM is also a computer manufacturer/dealer, it is intriguing that *TopComputers* has (apparently) nothing to do with IBM and *London* but they are in the same cluster. In fact, his company has no current business in *London*, so it would be potentially useful and interesting to find out the specific connections between *TopComputers* and *London* (and *IBM*).

- Some relations in the clusters might be obvious for the analyst, e.g., *Germany* and *TopComputers*, *Sales* and *Earnings*. However, other cases such as *Germany* and *London* which can be regarded as obvious for other people, might become interesting and even novel, in the context of his company (e.g.,, Might the company be missing out an important trading market in London?).

- There is a relatively close intra-cluster connection between *Island* and *Computer*. Of course, the analyst has no idea where this connection comes from and so he would like to have more information about it. As it is, the connection may hide a key association between those target concepts from different clusters: these concepts in cluster 2 are very related to target concepts *London* and *IBM* in cluster 1 (Could London-based IBM be expanding its business to some island whose potential the analyst was not aware of? Why is it that the company's competitor is expanding there?).

- There is an interesting connection which does not seem to be obvious: *Earnings* (and *Sales*) and *Environment*. Of course the analyst knows that the company's business has nothing to do with environmental activities, but the discovery suggests that there is something interesting with *Environment* that he does not know, and this might be somewhat affecting the company's earnings. Although this is not clear, this is definitely something the analyst needs to figure out (e.g., Could the manufacturing of the company's products have a negative effect on the environment which in return has an incidence on the sales?).

In this simple example, there are key issues that need to be addressed so as to turn them into strategic decisions. However, with only the information provided by the

cluster map, the analyst is unable to carry out a deeper analysis so to answer some of the questions above. Some reasons for this can be highlighted as follows:

- As the "knowledge" of the concepts is extracted in isolation, it is completely unclear what the nature of the relations/associations are. This suggests that providing keyword-based information is insufficient to understand and to make good use of the discovered relations.

- As no linguistic knowledge beyond keywords is considered, the mining process is usually rather limited to traditional statistical analysis of data. This will then fail to find information concerning the context of the knowledge or specific relationships contained in it.

- If the analyst really wishes to find out the complex relation(s) between target concepts, he/she will have to go through the hundreds/thousands of documents linked to each of these concepts. Next, he/she should read and analyse them in depth so as to come up with the hidden patterns. There is so much information and the task is so complex that he/she may not be able to do it at all.

- As the analyst has to "navigate" through the set of documents to investigate the hidden connections, the tool turns out to be a "information filtering tool" rather than a proper mining system. Indeed, the concepts in the map only represent a starting point to search for relevant information within the documents.

- The degree of interestingness, novelty and usefulness of the discovered relations is up to the user (i.e., analyst) who ultimately decides what is valuable and what is not. Indeed, before carried out a deep analysis, the example above shows that some relations can be regarded as interesting by the user (e.g., *London* and *Island*, etc). However, in large-scale text mining applications, the amount of discovered knowledge can be such that the user can not afford to carry out this assessment. In this case, the system should provide ways to evaluate the discovered knowledge so it is "filtered" before being delivered for its final assessment, in terms of criteria such as novelty, interestingness, usefulness, etc.

In order to deal with these issues, many current KDT approaches show a tendency to start using more structured or deeper representations than just keywords to perform further analysis so to discover informative and (hopefully) unseen patterns. Some of these approaches attempt to provide specific contexts for discovered patterns (e.g., "it is very likely that if X and Y occur then Z happens.."), whereas others use external resources (lexicons, ontologies, thesaurus) to discover relevant unseen semantic relationships which may "explain" the discovered knowledge, in restricted contexts, and with specific fixed semantic relationships in mind (Nahm and Mooney, 2000a; Hearst, 1998; Jacquemin, 2001; Polanco and Francois, 1998). Sophisticated systems also use these resources as a commonsense knowledge base which along with reasoning methods can effectively be applied to answering questions on general concepts.

Despite their advantages, most of the scope of application of these systems is still restricted as they fail to handle domain-specific terminology and/or to match patterns other than those pre-defined (Hearst, 1998; Harabagiu and Moldovan, 1999). Besides, in most cases, the usefulness, novelty and interestingness of the discovered knowledge have not been proved.

In this thesis, a model is proposed for knowledge discovery from texts (KDT) in which:

- evolutionary computation techniques of search and optimization are used to search for high-level and structured unseen knowledge in the form of explanatory hypotheses from previously extracted genre-based information from texts and from semantic and rhetorical training information.
- different strategies are proposed so as to automatically evaluate the multiple quality goals to be met by these hypotheses (interestingness, novelty, etc).
- and, a computer prototype is built to search for the best set of optimum hypotheses. Furthermore, experiments with human domain experts are carried out to assess the discovered hypotheses in terms of their quality.

Accordingly, the aim of this thesis is to produce a KDT model independent on domain resources which makes use of the underlying rhetorical information so to represent text documents for text mining purposes, satisfying the following aims:

1. To achieve a plausible search of novel/interesting knowledge based on an evolutionary knowledge discovery approach which makes effective use of the structure and genre of texts.

2. To establish evaluation strategies which allows for the measuring the effectiveness in terms of the quality of the outcome produced by the model and which can correlate with human judgments.

3. To produce novel knowledge which contributes additional information to help one better understand the nature of the discovered knowledge, compared to Bag-Of-Words (BOW) approaches.

4. To evaluate its adequacy in terms of the quality of the produced novel knowledge and its correlation with human judgments.

To this end, the following claims are investigated:

1. Additional linguistic information beyond keywords can be extracted from text documents to produce high-level representations and therefore the basis for effective discovered knowledge. This aims at not only helping the creation of novel and interesting knowledge but also at supporting the provision of explanations to help the user to better understand the relationship between target concepts when compared with keyword-based text mining systems.

2. Given the complexity in handling textual information across different domains and contexts, exploiting genre-based information (e.g., for scientific abstracts) can be useful to represent text documents. This information can capture semantic and rhetorical aspects of the underlying knowledge in a domain-independent fashion but at the same time, taking advantage of the genre.

3. Search techniques which look for the optimum set of hypotheses by exploring the complete search space are plausible. These strategies should be able to work in a resource-independent way so as to avoid any bias to some particular concept organization, and also in a way that takes into account the hypotheses' underlying linguistic knowledge.

4. Domain-independent strategies to evaluate the quality and plausibility of the discovered knowledge can be conceived. This evaluation can be used by the KDT model to guide search and optimization of possible explanatory hypotheses.

1.1 Organization of this Thesis

The thesis demonstrates that using additional genre-based linguistic information and semantically-guided search can be effective for knowledge discovery from texts.

- In Chapter 2, common approaches to keyword-based text mining and knowledge discovery from texts, along with the main techniques used are examined. This chapter also shows some representative tools that are used or that might be used by systems as part of structured discovery tasks.

- In Chapter 3, a new model for knowledge discovery from text that uses evolutionary techniques is developed. Some techniques/models in previous research are extended and new strategies for representation and evaluation are proposed in a way that the search for novel knowledge is strongly guided by semantic and pragmatic information previously obtained from the source text data.

- In Chapter 4, a prototype of the model is built and the produced outcome (hypotheses) is used in an experiment with domain experts. To confirm the effectiveness of the model, the experts assessed these hypotheses in terms of their quality from a KDD viewpoint. Further analyses are carried out to establish correlations between the experts and the model, and explanations are provided for some particular assessments.

- In Chapter 5, the thesis is concluded and further directions of research are suggested.

Chapter 2

Related Work

2.1 Data Mining

Data Mining is the process of discovering valuable information from large amounts of data stored in databases, datawarehouses, or other information repositories (Zhang, 2002; Han and Kamber, 2001).

Data Mining (DM) differs from traditional statistical analysis in that formal statistical inference is assumption-driven in the sense that a hypothesis is formed and validated against the data, whereas Data Mining is discovery-driven, that is, patterns and hypotheses are automatically extracted from data.

Popularly, DM has been treated as a synonym for *Knowledge Discovery in Databases* (KDD). However, a widely accepted definition for KDD (Fayyad et al., 1996) states it as *"the non-trivial process of identifying valid, novel, useful and understandable patterns in data"*, which points to KDD as a complicated process as part of which data mining only involves the task of discovering unseen patterns. Additional steps are then needed to establish whether these patterns are really novel, useful, etc. In this regard, four main steps can be identified in the full process of KDD (Han and Kamber, 2001):

- *Data Preprocessing:* involves the activities of Data Cleaning (noisy, erroneous, missing and/or irrelevant data are handled), Data Selection (data relevant to the analysis task are retrieved from the database), and Data transformation (data are transformed into forms appropriate for mining by performing summary or

aggregation operations).

- *Data Mining:* involves the essential process where intelligent methods (e.g., Machine Learning) are applied in order to extract data patterns. The user can significantly aid the data mining method by correctly performing the preceding step.

- *Pattern Evaluation:* identifies the truly interesting patterns representing knowledge based on some interestingness measures, tests the model for accuracy on an independent dataset, and assesses the sensibleness of the model.

- *Knowledge Presentation:* uses visualization and knowledge representation techniques to present the mined knowledge to the user.

The different DM techniques are usually associated with specific statistically-based methods (Han and Kamber, 2001) or Machine Learning (ML) strategies (Mitchell, 1997) depending on the nature, requirements and complexity of the problem. Popular tasks have included: finding *Association Rules*; where specific kinds of associations between terms are captured in terms of *Support* and *Confidence* (Han and Kamber, 2001; Reinartz, 1998), *Prediction*; where the prediction of missing or relevant data is carried out by using traditional methods (e.g., regression analysis, generalized linear model, correlation analysis), or ML techniques (e.g., Genetic Algorithms, Neural networks, decision trees, etc), *Term Clustering*, and *Classification*; where unknown data patterns are assigned to established classes based on a training model (e.g., class prediction) in which only the features are known, commonly performed via decision trees, k-nearest neighbour, and neural networks.

Although there is a significant number of very efficient ML algorithms (Mitchell, 1997) that have been proved to be valuable in a variety of real-world applications, these along with the other DM techniques have limitations. These assume that the data have been carefully collected into a single database with specific DM tasks in mind (Mitchell, 1999). In addition, despite the fact that this generation of DM techniques works well with numeric and symbolic features, there still lack effective and robust algorithms for learning from data that is represented by other media or a combination of these (e.g., images, texts, etc).

2.1.1 Evaluation in KDD

A critical aspect of KDD is that the discovered knowledge should be somehow interesting (*pattern evaluation*), where the term *"interestingness"* is argued related to the properties of surprisingness, usefulness and novelty of the new knowledge.

Although KDD is defined as the process of identifying valid, novel, useful and understandable patterns in data, most literature is just about validity, and very little is about novelty, utility and understandibility. An examination of many popular works on Data Mining by Pazzani (2000) shows that none of the approaches are devoted to making sure that the knowledge is really novel, useful, and understandable. While some KDD systems cover these topics, most contain unfounded assumptions about "comprehensibility" or "interestingness". Besides, no metrics used by the different techniques have proved to correlate with user judgements of what is interesting. In fact, as regards "comprehensibility", there has been no study that shows that people find smaller models more comprehensible or that the size of a model is the only factor that affects its comprehensibility (Pazzani et al., 1997). In this regard, (Gaines, 1996) points out "Psychological studies of the nature of comprehensibility of knowledge structure are necessary to give substance to the intuitions". It is claimed then (Pazzani, 2000) that the benefits of KDD can be realized by paying attention to the cognitive factors that make the resulting models coherent, credible, easy to use, and easy to communicate to others.

Hence, by taking the human cognitive process into account, we might be able to increase the usefulness of KDD systems as it is ultimately the people's perception of novelty, utility and understandibility which determine the acceptance of data mining.

Some existing approaches to finding subjectively interesting knowledge (i.e., in the form of prediction rules) ask the user to explicitly specify what types of rules are interesting and uninteresting. The system then generates or retrieves those matching rules. However, given the huge number of possible rules generated by the system, it is unclear which rules should be filtered so as to be assessed by the user, and what the system's "interestingness" means for this user.

For many other approaches, the lack of human evaluation has been dealt by assuming that the interestingness can be regarded as the degree of "unexpectedness" of the

discovered knowledge to the user. Measuring this kind of criterion usually involves a combination of objective measures (statistically-based metrics) and subjective measures (based on the user's beliefs or expectations for the data). Differences in these methods lie in what sort of "previous" knowledge the discovered patterns should be compared with so as to measure the surprisingness. Accordingly, representative approaches can be characterized by the previous knowledge that they use and the analysis performed to determine this degree of "surprisingness", as follows:

- *Templates:* in one template-based approach, the user specifies interesting and uninteresting rules using templates (Klementinen, 1994). A template describes a set of rules in terms of items occurring in the conditional and the consequent parts. The system then retrieves the matching rules from the set of discovered rules. However, the degree of surprisingness is implicit in the specification of the interesting rules. That is, the user is asked to state what interestingness would be for him/her.

- *Beliefs:* here the methods for discovering unexpected patterns consider a set of expectations or beliefs about the problem domain (Padmanabham and Tuzhilin, 1998). However, this kind of method is generally not as efficient or flexible as post-analysis methods (Liu et al., 2000) unless the user can specify his or her beliefs about the domain completely beforehand, which is very difficult, if not impossible. Typically, the user must interact with the system to provide a more complete set of expectations and find more interesting rules.

- *Information Gain:* surprisingness, for some people, is also seen as a relation to the amount of information conveyed by the discrete items (attributes) contained in a discovered rule. In this regard, (Freitas, 1998) proposes evaluating the interestingness of discovered pattern (e.g., rule) as a combination of its predicting accuracy.

  The predictive accuracy is measured for the attributes of a rule antecedent by using information-theory based metrics, specifically ones based on *Information Gain* (Cover and Thomas, 1991). This is calculated for each predicting attribute

of the rule. Attributes with high information gain are expected to be good predictors of class, when these attributes are considered individually. It is assumed that a user would tend to be more surprised if he/she saw a rule containing attributes with low information gain (Jaroszeqicz and Simovici, 2001). Therefore, rules whose antecedent contains attributes with low information gain are more surprising than rules with attributes of high information gain.

In order to measure the rule consequent's degree of interestingness, the idea is that the larger the relative frequency of the value being predicted by the consequent, the less interesting it is (Noda et al., 1999). Basically, this considers the proportion of items in a database that satisfy both the antecedent and consequent over the number of items that satisfy only the rule antecedent.

- *Databases:* a statistically-based approach is taken by (Radcliffe and Surry, 1994), in which the interestingness of a mined pattern (predicted rule) is computed in terms of its quality, accuracy (Liu et al., 2000) and coverage, and its degree of generality. Unlike (Noda et al., 1999), the measures are evaluated according to the items covered in a database. Specifically, accuracy is defined as the proportion of items in the database that satisfy both the antecedent (A) of the rule that also satisfy the consequent (C), that is $\frac{|A \cap C|}{A}$. Whereas coverage is seen as the proportion of items that satisfy the consequent of a rule that also satisfy the antecedent of this (i.e., rules that apply to a larger fraction of the database are useful), that is $\frac{|A \cap C|}{C}$. Generality is promoted with a function that monotonically increases with either or both $| A \cap C |$ and $| A' \cap C' |$, where A' is the set of items not in A.

  All these metrics are then combined in a nonlinear way as an overall criterion of goodness. A distinguishable feature in Radcliffe's approach is the assumption that interestingness of the predicted rules can be captured by statistical correlations with the databases. However, the outcome (predicted rules) has not proved to be effective or useful in terms of human judgements.

- *Actionability and Statistical Deviations:* Some efforts in specific domains have dealt with the problem of assessing interestingness in terms of interesting devia-

tions found in data and their actionability; when the user can act on it to her/his advantage (Mitra, 2002; Silbershatz and Tuzhilin, 1996). In this kind of system, a set of "key findings" (Piatesky-Shapiro and Matheus, 1994) is discovered whose interestingness (regarded as a degree of actionability) is then evaluated by a domain knowledge-based system (e.g., expert system). The measure is realized by using a promising concept of payoff where the aim is to capture the expected payoff from the actions that follow from the discoveries (deviations).

- *External Resources:* in approaches to mining patterns from text data (Basu et al., 2001a), the quality of the discoveries (e.g., rules) is measured according to the existing lexical and semantic information in a general-purpose linguistic resource (WordNet). The evaluation involves assessing criteria including the coverage of the rule (i.e., number of items covered by the rule in a training set), and the semantic interestingness. For this interestingness, items of the rule's antecedent and consequent are evaluated according to the semantic distance between them in WordNet. The longer the semantic distance is, the more intriguing (interesting) the relation. Finally, a small set of filtered best discovered rules which were evaluated as interesting by the system, are provided to human evaluators. Resulting experiments show that the system evaluation correlates well with the human judgement. However, it is noted that the dependence on this kind of resource, where not all the items contained in the rules are present, is a constraint.

A different view of the process of finding interesting nuggets in which the user feedback is taken into consideration is accomplished by (Williams, 1999). He proposes a methodology to evolve (via Genetic Algorithms) the patterns of interest in an exploratory way in which users after exploring the nuggets, provide ranking of their degree of interestingness. This ranking is analysed and then converted into a measure of interestingness of the rule which is used for the system to refine the search space and so to look for better nuggets. As the system may generate too many rules in each generation, just a small subset is chosen from each each generation which then are presented to the user for ranking.

Although the approach contributes with a user-centered view for assessing discovered patterns, its usefulness has not yet been proved. In addition, the method to capture the interestingness so as to be used by the system (i.e., function that convert the ranking into system's feedback), still requires considerable research.

Despite the effort of the different approaches to measure interestingness, there is no evidence proving that this actually leads to valuable patterns from a KDD perspective. In most cases, there seems to be a confusion between the real interestingness, novelty and usefulness of the discoveries. Instead, they are all perceived as assessing the same criteria (novelty and interestingness), hence no clear distinction is commonly made. In many real applications, this distinction would have strong consequences on the outcome of the KDD process: the user might find the discovered patterns novel but not interesting, very interesting but not useful, interesting but not novel, and so on.

2.2 Text Mining and Knowledge Discovery from Texts

The problem with text mining is that unlike tabular records in databases, documents are not structured and normalized so that they can be easily interpreted by computers. The lack of structure raises the difficulty of uncovering the implicit knowledge inside the documents.

For some TM approaches, this lack has been dealt with by assuming that, following the IR tradition, the documents are represented as "bags of words" (i.e., keywords) or terms from which some classification or association is computed in a high dimensional space.

In other approaches, deeper analysis and high-level representations (other than keywords) are proposed. These are characterized by using IE techniques or more tailored NLP techniques so as to come up with more abstract discovered knowledge in terms of more meaningful and deeper relationships, concepts, or extending existing conceptual resources, where the discovery is seen as unseen links to be found between concepts.

2.2.1 Approaches to TM and KDT

We can view the different approaches to TM and KDT by first looking at how these approaches deal with the mining problem in terms of representations used and the analysis carried out, and then, the common kinds of supporting tools used. Accordingly, the rest of the chapter is organized as follows: sections 2.2.1.1 and 2.2.1.2 describe the main representative bag-of-words and higher-level approaches to TM/KDT along with issues concerning evaluation in some of them. Section 2.2.2 outlines some tools used or appropriate for TM purposes so far that are worth taking into consideration. Finally, in section 2.3, the main issues, problems and challenges arising from these approaches and the tools are highlighted.

2.2.1.1 Bag-of-Words based Approaches

Part of the initial work on Text Mining came from the IR community, hence the assumption of having texts represented as just *Bags of Words* (BOW), and then proceeding via classical data mining methods (Feldman and Dagan, 1995; Feldman, 1998a; Rajman and Besancon, 1998; Reinartz, 1998). Usually, two types of information can be handled from text documents: *Keywords* or *Terms* (word sequences that are likely to have meaning in the domain). Once these have been extracted, traditional KDD operations are carried out to discover associations.

Although the architecture of each TM system varies depending on the application and task in mind, common underlying working principles can be seen in the general framework in figure 2.1.

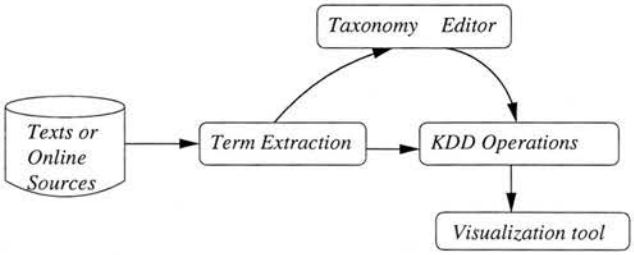

Figure 2.1: A Typical Architecture of a BOW-based TM System

The system is provided with collections of raw texts. Text documents are first labeled with terms extracted directly from them by using indexing techniques (Jacquemin, 2001; Muller, 1997), which are often assisted by external resources such as thesauri. Once the terms are extracted, these are used to support a range of KDD operations. As the system is provided with a domain resource (e.g., thesaurus), and most of the terms are directly (or indirectly) related to those included in the resource, a taxonomy-based organization along with tools (taxonomy editor) to handle and update the resource (according to the user's feedback) are usually included so as to assist the KDD process. A visualization tool is optionally provided to have the terms and associations displayed.

The TM tasks carried out have involved using their own methods to discover an underlying conceptual structuring of text data, going beyond pure keyword-based techniques. The most important approaches can be summarized as follows:

1. *Regular Associations:* here is assumed that the textual data is indexed on the terms contained on these (manually or automatically) with the help of morphosyntactical and morphological analysis techniques (Muller, 1997; Jacquemin, 2001).

   Next, this information is used to discover association rules between indexed terms that typically look like: `IF ("Arab", "Egypt","Iran") THEN "Oil"`, meaning that the occurrence of keywords "Arab", "Egypt" and "Iran" is associated with "Oil" (the concepts here can be either terms or keywords). In order to generate these association rules, adapted measures of support (a measure that indicates the portion of items that satisfy both the antecedent and the consequent of a rule) and confidence (indicates the frequency with which the consequent is satisfied when the antecedent is also satisfied) are applied (Feldman, 1998a). Once this has been computed, a common interest is in the groups of items that have transaction support (i.e., pair of items numerically related) bigger than a given minimum support, usually referred to as *itemsets*.

2. *Concept Hierarchies:* each document is tagged with terms extracted from a hierarchy of concepts (i.e., thesaurus). Then, a system analyses content distribution of these terms according to metrics of joint distributions so as to extract new

relationships between them (Feldman, 1998a). Approaches such as Feldman's FACT (Feldman, 1998b) use this kind of relationship for filtering and summarizing news articles, utilizing concept distributions. These can later be marked as interesting by some when browsing.

3. *Prototypical Text Mining:* unlike regular associations which exclusively operate on the document indexes, this technique takes advantage of the textual content of the documents. In this context, "prototypical" is defined as information that occurs in a repetitive fashion in the document collection (Rajman and Besancon, 1998). The working hypothesis is that repetitive document structures provide significant information about the textual base that is processed. Basically, the method relies on the identification of frequent sequences of terms in the documents, and uses NLP techniques such as POS tagging and term extraction (Jacquemin, 2001) to preprocess the textual data.

   Note however that, this kind of extraction only deals with textually consecutive patterns which represents a limitation on the knowledge acquired, and restricts the search of the discovered patterns.

4. *Concept Clustering:* The representative concepts (e.g., keywords) of the documents are extracted and then clusters that represent the groups of highly related keywords are built. The results of the clustering method can be used in one of two ways. First, for summarizing the contents of the target database by considering the characteristics of each created cluster rather than those of each item of the text. Secondly, as an input to other methods, e.g., supervised induction. Two of the most popular clustering methods used in TM include **Axial K-means** and **Self-Organizing Maps** (SOM). In particular, SOM have become very popular as the use of their underlying strategy (Kohonen's neural network) has proved to be very effective for classifying and/or clustering terms of the documents analysed, that is, the task of ordering high-dimensional statistical data.

5. *Hybrid discovery:* so far, numeric association techniques and symbolic methods have worked separately, depending on the TM task in mind. However, some promising approaches have explored the benefits of hybrid systems for

TM which combine both kind of methods. Specifically, Toussaint and colleagues (Toussaint and Simon, 2000; Lamirel and Toussaint, 2000) prove that the mixture of these approaches, has mutual advantages so as to produce significant improvements on traditional term-based tasks such as classification. On the numerical side, they use SOM to produce classes of documents which are related according to the similarities of the concepts (keywords) contained in them.

The symbolic side involves a conceptual model based on Galois lattices (Kourie and Oosthuizen, 1998; Ganter and Wille, 1999; Wille, 2001) which is used to build a concept hierarchy. Each class (concept) is a set of keywords that represents the original document. The method is then used to build the lattice from these classes. The outcome is the hierarchy of classes that are related in terms of shared features. In addition, new classes are added as a result of specific graph-based operations (Ganter and Wille, 1999).

Once the lattice is built, the task aims at extending the knowledge provided by the SOM with the lattice so to find (or to confirm) more specific connections (this assumes that one method can find associations that the other method can not). Specifically, the groups that have been clustered by the SOM can be extended with the generalization rules obtained from the lattice as they involve the same starting classes, but the lattice contains knowledge about new classes (enabled by the lattice operations) and a kind of relationship (e.g., generalization) that SOM does not. In other words, the description of the SOM is completed using rules extracted from that lattice so as to compliment or to augment the associations.

Issues and Problems

The specific approaches above provide tailored or specific techniques for mining textual data. A common limitation is that these are mostly concerned with finding some association between concepts, whether it is hierarchical or regular, and no evidence is provided to help one to understand the nature of the discovered relationships. For example, imagine that one of the methods discovers an association between the terms "weather" and "sales". One may wonder, what is the specific relationship of interest

between them. There are many possibilities that certainly depend on the context and the semantics of the patterns found. If the text database involves a set of summarized reports of an umbrella factory, a candidate for that relation might be: `increases`, meaning that *"weather increases sales"*, assuming that "weather" really means bad weather. Whereas, if the textual information involves reports from an ice cream factory, a candidate is likely to be a relation such as `decreases`. In this example, there are some important issues that current TM approaches fail to handle and so which need to be addressed:

- The usual representation based on BOW has some important limitations in patterns it can represent. For instance where a text includes numerical relationships these will usually be lost in a BOW model. In the example above, two different candidate relationships are even opposite, and since no deep knowledge from the text is provided, this could not be detected by the system.

- As most discovered associations are numerical in nature, no linguistic information beyond the terms is used, and so rich textual knowledge may be lost (i.e., predicate-level information along with the terms). Hence deeper representations are needed.

- Even though an explanation or further information may be provided, the complete responsibility for validating and assessing the usefulness and novelty of the mined patterns is with the user. The majority of these approaches do not provide any evaluation of what they are producing. As a consequence, the unseen produced knowledge has not been proved to be useful in most cases (Pazzani et al., 1997; Pazzani, 2000).

- In general, if the discovered patterns and the information contained in the documents are strategical and highly critical for some decision maker, then it is his/her responsibility to look through the thousands of documents and to try to find out the missing relationships and/or concepts involved. Because of this, the associations are just a clue of concepts that may be worth searching for in the text database. Hence the TM tool turns into a kind of augmented searching tool with keyword analysis capabilities rather than a proper discovery system.

2.2.1.2 Deeper and Structured Approaches

Some promising high-level approaches dealing with most of the previous issues have appeared. These are characterized by using some level of IE processing which allows them to have deeper representations, deal with these using more suitable techniques, in specific or general-purpose domains, and to use external resources to assist the discovery. The discovered knowledge takes many diverse forms, ranging from unseen relationships between known concepts or rules relating concepts, to inferred and unseen semantic relations that allow for the augmentation of a thesaurus or some lexical database. The main approaches are discussed as follows:

- **Swanson's Causal Relations Discovery**

  An early and influential piece of research by (Swanson, 1988), was a very promising approach to exploratory text data analysis. His main aim was to discover causal relations of interest (i.e., for which the answer was not currently known) from the title of the articles stored in the MEDLINE medical database (Ding et al., 2002). He showed that chains of causal implication within the medical literature can lead to hypotheses for causes of rare diseases, some of which may receive scientific supporting evidence.

  The underlying discovery method is based on the following principle: some links between two complementary passages of natural language text can be largely a matter of form "A causes B" (association AB), and "B causes C" (association BC). From this, it can be seen that they are linked by B irrespective of the meaning of A, B or C. However, perhaps nothing at all has been published concerning a possible connection between A and C, even though such link if validated would be of scientific interest.

  In the biological world of multiple causation, the above construction is not transitive, and so any conclusion would in general depend on understanding the biological meaning of the two premises. However, the two premises suggest that the hypothesis "A causes C" might be worth testing as a novel pattern. In other words, AB and BC are complementary if useful information can be inferred by considering them together that cannot be inferred from either one alone.

Swanson designed a system which using these assumptions was able to find new associations such as *"stress is associated with migraines"*, *"magnesium is a natural calcium channel blocker"*, etc. Although the hypotheses had to be tested by humans, these proved that it is potentially plausible to derive new patterns from a combination of text fragments and the explorer's medical expertise.

- **Hearst's Lexiconsyntactical Patterns**

  Despite the promise of Swanson's approach for simple hypothesis discovery, its scope is restricted as this only applies to a domain in which small number of patterns arise. Besides, Swanson makes the strong assumption that the knowledge lies in the titles of the articles, something potentially useful for searching purposes but not sufficient for text analysis in which key linguistic knowledge may be obtained from the full text.

  Hearst (Hearst, 1997; Hearst, 1999) tries to deal with some of these issues by proposing a domain-independent method for the automatic discovery of WordNet-style lexicosemantic relations. The method searches for corresponding lexicosyntactical patterns in unrestricted text collections, and then finds unseen links which relate the new concepts to those existing in the WordNet hierarchy (Fellbaum, 1998).

  The starting point is the extraction of information from the texts using specific extraction patterns called *Lexico-Syntactical Patterns* (LSP) to capture basic linguistic relations such as hyponyms and hypernyms. For example, a typical handcrafted LSP would look like: "`NP0 such as NP1 {,NP2,...(and/or) NPi}`", where NP_i represents a Noun Phrase, and which $NP_i, i \geq 1$ is hypothesized to be a candidate concept in a hyponym relation with NP_0 (i.e., *hyponym*(NP_0, NP_i)).

  According to the proposed discovery method, this relation is regarded as "new" (unseen) in WordNet, and suggested to be added only if both NP_0 and NP_1 are present in WordNet, and the relation is not. Similarly, the NPs are regarded as "new" if one or both of these are not present in WordNet in whose case these are suggested to be added to the existing relation.

Despite this approach's simplicity for augmenting the structure (and contents) of WordNet in terms of simple semantic relations, the new knowledge is perceived as unseen relations that can be incorporated into WordNet. Because of this, the discovery is restricted to the idiosyncractic organization of WordNet and so the patterns have to be specifically designed with this in mind.

This suggests that the method has a low coverage in terms of a wider range of semantic relations beyond those captured in the patterns. However, even if further patterns are added, the dependence on the WordNet organization is still a constraint, which also applies to the concepts: if these do not appear as they are in WordNet, the method fails to make a proper decision.

- **Jacquemin's Augmentation of Thesaurus**

  Hearst's method uses a general-purpose lexical database (WordNet) which fails to cope with specific terminology in different domains. The fact that concepts do not appear in the ontology as it is, does not mean they do not exist. In some cases, the ontology may contain a variation of the concepts (e.g., "crop in vitro of enbryon" versus "crop of enbryon", "high and low pressure" versus "high pressure", etc), which can recognized with more sophisticated techniques.

  In order to exploit these underlying terminological variations, Jacquemin (Morin and Jacquemin, 1999) proposes a method which combines classical term acquisition techniques and the automated construction of a specific thesaurus aimed at augmenting the semantic links contained in this thesaurus (Jacquemin and Tzoukermann, 1999). Like Hearst, Jacquemin also aims at discovering new semantic links. However, term variations are explored to enable a more robust way to compare new and existing terms.

  Unlike Hearst's IE patterns, the relevant patterns to be used in the extraction can automatically be generated from a set of initial candidates (primitive IE patterns). However, like Hearst, the aim is also to find hypernym relations of interest from the information matched by these selected patterns in the corpus.

  The process for generating the IE patterns to be used in the extraction, is based on a hypothesis called *Syntactic Isomorphy* (this and other specific-purpose hy-

potheses can be seen in (Jacquemin, 1999)). The basic working assumption is that there should be common characteristics between the primitive extraction patterns so that only some of them should be selected for the information extraction task itself, for example, the most representative ones.

Specifically, this states that if there is a pattern with an expression involving a pair of terms A_j and A_k (linked by a relationship $hypernymy(A_j, A_k)$), and another pattern with an expression involving the terms B_j and B_k (linked by $hypernymy(B_j, B_k)$), then their items (A_j, B_j) and (A_k, B_k) have the same syntactic function. Next, one of these patterns (i.e., the simplest one) is selected to participate in the extraction process.

For example, consider that there are two similar extraction patterns:

```
NP find in NP such as NP1,NP2,..
and
NP such as NP1,NP2,..
```

The hypothesis above states that both patterns have the same syntactical function, therefore, only one of them should be chosen. In this case, the simplest one "`NP such as NP1,NP2,..`" is selected as a representative one for the further task of IE. Once all the patterns are selected and the extraction is performed, the term acquisition techniques (Jacquemin and Tzoukermann, 1999) are applied. These aim at detecting term variations so as to have the terms and then matched against those in the thesaurus.

One of the advantages of this approach is on to simultaneously exploit morphological, syntactic and semantic links to detect the variations and to use this information to automatically extend the thesaurus with unseen concepts and links.

Although the use of a thesaurus allows the method to overcome the lack of specific terminology, the kind of relation is still fixed and strongly based on the terms included in the handcoded extraction rules.

- **Harabagiu's Commonsense Knowledge Discovery**

None of the previous approaches offers any reasoning capability in order to discover richer concept connectivity beyond basic fixed semantic relations. In fact, in these cases the simple "inference" is restricted to the specific kind of relation enabled (e.g., hyponyms, hypernyms, etc).

This issue is partially overcome by (Harabagiu and Moldovan, 1998; Moldovan, 2000) who propose a reasoning method to exploit the underlying knowledge in WordNet's semantic relations so as to discover unseen facts.

Their underlying mechanism assumes that input data consisting of pairs of predicative sentences (actions represented as a verb and arguments) are provided. For this, the main argument's glosses (the description of each of the argument's terms according to WordNet) are looked for. Next, semantic paths between the concepts are followed in order to achieve common links. For example, in the following two sentences: *"Jim was hungry"*, and *"He opened the refrigerator"*, a human may infer that the intention is eating, that the most likely event takes place at home, and that the text is coherent since being hungry is a cause for opening the refrigerator where the food is stored.

In order to infer this kind of fact, it is suggested that there will be common meeting (collision) points for the concepts' semantic paths which enable this kind of reasoning. Once these "micro-contexts" have been represented, some inference rules (Harabagiu and Moldovan, 1998) are applied to combine the semantic relations found in the chain so to find unseen derived semantic relations.

The inferences may be made by collecting the concepts retrieved from these paths and the other semantic paths established between both sentences by using different marker propagation rules.

Although this represents a plausible commonsense reasoning strategy, in specific domains, where specific terminology is involved, the method fails to draw any inference as new relations or terms appear. The method will not be able to deal with relations beyond those established in WordNet, and therefore, any further complex inference will be restricted to the small number of inference rules defined.

In addition, the plausibility of the knowledge discovered cannot be quantified so as to indicate whether this produces any novel or interesting knowledge beyond commonsense discovered facts.

- **Mooney's Discovery of Novel Patterns**

  By looking at the essence of the previous methods, it can be observed that, apart for Swanson's approach, all of them are focused on using WordNet either for creating/enhancing its contents or for enabling further commonsense knowledge inference mechanisms.

  A different view of mining patterns using WordNet has been taken by Mooney and colleagues (Nahm and Mooney, 2000a; Nahm and Mooney, 2000b) in which the resource was used as a knowledge base that allows for the assessment of the mined associations (rules).

  In order to capture the initial data, a text mining system (DiscoTEX) is used to discover prediction rules from natural language corpora by using a combination of principles of information extraction and data mining.

  In the experiments, they used rules mined by DiscoTEX from book descriptions extracted from `Amazon.com` in several categories including "literature", "romance", etc. DiscoTEX first extracts a structured template from the Amazon book description pages. It constructs a template for each book description, with pre-defined slots (e.g., title, author, subject, etc.) that are filled with words extracted from the text. The system then uses a rule mining technique to extract prediction rules from this template database. A typical mined rule in which different sorts of slots are filled may look like:

```
IF   <title>  daring, love
     <synopses> woman
     <subject> romance, historical, fiction
THEN <comments> story, read, wonderful
```

  In order to automatically evaluate the novelty of this kind of rule, it is claimed that the rule requires to be compared to an existing body of knowledge the user

is assumed to already possess (Basu et al., 2001a). In the context of traditional text mining, in which rules consist of words in natural language, a relevant body of common knowledge is basic lexical semantics, i.e., the meaning of words and the semantic relationships between them. For this, a method is proposed for measuring the novelty of text-mined rules using WordNet. Here, a measure of the semantic distance, $d(w_i, w_j)$, is defined between two words based on the length of the shortest path connecting w_i and w_j in WordNet. The novelty of a rule is then determined as the average value of $d(w_i, w_j)$ across all pairs of words (w_i, w_j), where w_i is in the antecedent and w_j is in the consequent. Intuitively, the semantic dissimilarity of the terms of a rule's antecedent and in its consequent is an indication of the rule's novelty. For example, "beer $\rightarrow$ diapers" would be considered more novel than "beer $\rightarrow$ pretzels" since "beer" and "pretzels" are both food products and therefore closer in WordNet.

Formally, this semantic distance measure between the two words w_i and w_j is defined as follows:

$$d(w_i, w_j) = \overbrace{Dist(P(w_i, w_j))}^{Distance} + K * \overbrace{Dir(P(w_i, w_j))}^{Direction}$$

where $P(w_i, w_j)$ is a path between w_i and w_j, `Dist(p)` is the distance along path `p` according to a weighting scheme, `Dir(p)` is the number of direction changes of relations along path `p`, and K is a suitable chosen constant.

The direction component is based on the direction classes for the relations of WordNet defined by Hirst and St-Onge (1998): "up", "down" and "horizontal", depending on how the two words in the relation are lexically related. The direction information for the relation types used in the evaluation can be seen in table 2.1. The more direction changes in the path from one word to another, the greater the semantic distance between the words because changes of direction along the path reflect changes in semantic context.

The distance component is defined as the shortest weighted path between w_i and w_j, where every edge in the path is weighted according to the weight of the WordNet relation corresponding to that edge, and is normalized by the depth in

Relation	Direction	Weight
Synonym, Attribute, Pertainym, Similar	HOR	0.5
Antonym	HOR	2.5
Hypernym, (Member|Part|Substance) Meronym	UP	1.5
Hyponym, (Member|Part|Substance) Holonym, Cause, Entailment	DOWN	1.5

Table 2.1: Direction and weight information used to measure semantic similarity

the WordNet hierarchy where the edge occurs. In this method, 15 different relations between words in WordNet were used, and different weights were assigned to different link types (e.g., hypernym represents a larger semantic change than synonym, so hypernym has a higher weight than synonym). The weights are given in column **Weight** of table 2.1.

Keeping the distance and direction issues in mind, a scoring algorithm to compute the novelty of a rule containing an antecedent and a consequent (i.e., set of words) is designed. The algorithm basically calculates the above distance ($d(w_i, w_j)$). However, note that there are several problems to finding the path between words in WordNet: some sub-hierarchies are disconnected (e.g., there are 11 trees with distinct root nodes for nouns, and 15 for verbs), some words might be not valid ones in WordNet, etc. In order to deal with these issues, the method connects trees so that a path can always be found between words. For example, the 11 nodes of the noun hierarchy are connected to a single node R_{noun}. Furthermore, R_{noun} and R_{verb} are connected to a top-level root node, R_{top}. In this composite hierarchy, the weighted shortest path between two words is obtained by performing a branch and bound search. However, given the large number of combinations of paths within the hierarchy while performing the search, the method provides a user-defined time-limit within which the shortest path between the words w_i and w_j is looked for. If the path cannot be found within

this limit, the algorithm finds a default path between w_i and w_j by going up the hierarchy from both w_i and w_j, using hypernym links, until a common root node is reached.

The algorithm used to compute the score of each rule based on the distance with the previous constraints is highlighted in figure 2.2.

For each rule R in the rule set (mined rules) DO
  Let A be the set of antecedent words of R
  Let C be the set of consequent words of R
  For each word $w_i \in$ A and $w_j \in$ C DO
    IF w_i and w_j are not valid words in WordNet THEN
      $Score(w_i, w_j) \leftarrow PathViaRoot(d_{avg}, d_{avg})$
    ELSEIF w_j is not a valid word in WordNet THEN
      $Score(w_i, w_j) \leftarrow PathViaRoot(w_i, d_{avg})$
    ELSEIF w_i is not a valid word in WordNet THEN
      $Score(w_i, w_j) \leftarrow PathViaRoot(d_{avg}, w_j)$
    ELSEIF path not found between w_i and w_j (in time-limit) THEN
      $Score(w_i, w_j) \leftarrow PathViaRoot(w_i, w_j)$
    Else
    $Score(w_i, w_j) \leftarrow d(w_i, w_j)$
  End-For
  Score of rule $\leftarrow$ average of all (w_i, w_j) scores
End-For
Sort scored rules in descending order

Figure 2.2: The Rule Scoring Algorithm

The function `PathViaRoot` in figure 2.2 computes the distance of the default path. For nouns and verbs, `PathViaRoot` calculates the distance of the path between the two words as the sum of the path distances of each word to its root.

If one of the words is an adjective or an adverb and the shortest path method does not terminate within the time limit, then the algorithm finds the path from the

adjective or adverb to the nearest noun, through relations like "pertainym", "attribute", etc. It then finds the default path up the noun hierarchy, and `PathViaRoot` incorporates the distance of the path from the adjective or adverb to the noun form into the path distance measurement. For words that are not valid in WordNet, the algorithm assigns the average depth of a word (d_{avg}) to those words.

In order to check if the automatic ratings of novelty of the algorithm correlate with human judgments, an experiment was carried out in which the method took rules generated by DiscoTEX from 900 book descriptions taken from Amazon. Then, they were filtered so that the set was not too large for human evaluators. A portion of these rules was used for training purposes, and the others were used as test sets for the experiment. Next, two kind of measurements were performed: one considers the average correlation between human subjects in the judgment of the novelty of the rules, and the other considers the correlation between the novelty scores of the algorithm and those of the human evaluators.

The overall results show well correlated ratings between human evaluators (i.e., from a correlation (Spearson) r=0.337 to r=0.412). The correlation between the groups of humans and the algorithm ranged from r=0.137 to r=0.386, showing that the correlation between the humans and the algorithm is on the average comparable to that between the human subjects.

However, this correlation is not very high which suggests that some other factors may influence the system's evaluation or the human judgement. Indeed, it is recognised that some measurements might be misleading as the set of mined rules contains many words unknown to WordNet (e.g., personal names) from which default processing is assumed. In these cases, the unseen information has been scored highly interesting by the system but is often uninteresting to humans. In other cases, it is noticed that rule such as "sea $\rightarrow$ oceanography" is rated high by the algorithm, while most human subjects rated that rule as uninteresting. This usually happens because there is no short path between "sea" and "oceanography" in WordNet.

This evidence shows again that a discovery process depending on WordNet may produce misleading results because of the lack of domain-specific knowledge.

In addition, having a specific resource (WordNet) fails to capture all semantic relationships between words. Hence the method could benefit from other approaches to lexical semantic similarity such as co-occurrence, Latent Semantic Analysis, etc.

Issues and Problems with Structured Approaches

Most of the structured approaches to KDT discussed rely on a specific organization of conceptual resources, whether they are domain-independent (e.g., WordNet) or domain-specific (e.g., a thesaurus). However, it can be seen that the effectiveness of the methods, from a KDT perspective, is affected in terms of robustness as the discovered knowledge is highly dependent on the existing information, and the particular semantic acquisition task in mind. However, it is unclear whether the results are novel, interesting, etc because they have not been evaluated by humans.

This dependence also restricts the scope of the IE patterns to extract key information (i.e., the patterns rely on specific-purpose semantic information to be extracted from a linguistic resource), and so makes it difficult to search for global novel knowledge. Instead, the search seems to be focused on what the semantics of the links states and on the operations enabled to move within an external resource so to infer some facts.

Interestingly, this shows that the ability for global search which is mostly performed in BOW approaches, has been lost in more structured approaches because of their focus on specific semantic information. For example, in clustering tasks, the system must consider the "influence" of all the items on other individual items so as to make similarity predictions which allow for their grouping. In other tasks that involve finding relevant semantic similarity between items, the whole corpus must be considered to obtain the representation for the items so as to make further judgments. Note that in most of the structured approaches, the judgments are only made from the information which is extracted by the IE patterns and so from the direct "inferences" drawn from them (along with the resource). Presumably, one explanation for not going beyond the information that the patterns convey is that, the methods would be unable to prove whether the possible inferences are valid.

While the use of these IE patterns for extracting initial knowledge is a common good feature explored in the different approaches, the scope for these linguistically-motivated patterns is then an aspect which should be addressed. In order to deal with a more resource-independent analysis and to allow higher coverage for the IE patterns, more flexible ways to exploit the patterns by using other kinds of knowledge should be considered.

As most of the tasks have concentrated in enhancing the conceptual resources rather than discovering proved novel knowledge, the assessment of the discovered knowledge has generally been neglected. Although Mooney and colleagues (Nahm and Mooney, 2002) address this issue, the fact that the evaluation depends on the resource organization restricts the real effectiveness of the discovered patterns from a user viewpoint. This is because both the concepts and relationships need to be included in the resources, a situation that in real applications can not be afforded (in specific domains, the general lexical database would not be very useful except in a few generic cases). However, this kind of method could be improved if other alternative ways of measuring semantic distance were provided. In this context, assessment based on semantic information provided by corpus-based techniques has not yet been exploited.

As regards the representation, note that because of the resource-driven nature of the discovery, the different techniques have essentially aimed at defining patterns to extract the information of interest. Little effort has been put on first representing the documents in a proper way for then carrying out more focused discovery tasks.

2.2.2 Tools for Text Mining

Despite the large number of techniques and tools for text mining, only a few of them have, strictly speaking, effectively been used for or as a part of major KDT tasks. The rest of them have just concentrated on typical mining activities as previously highlighted. In order to focus our scope, these tools can be divided into three major groups: Tools which have been used for performing tasks involving similarity measurements (e.g., clustering, retrieval, etc) such as *Latent Semantic Analysis*, those that have been used for capturing and extracting relevant information from text documents such as *Information Extraction*, and finally, machine learning techniques which are promising

in searching and optimization problems such as *Genetic Algorithms*. We concentrate on these three specific tools because they have shown fair performance and effectiveness in other contexts and so we believe their potential can be further explored for the purpose of knowledge discovery.

2.2.2.1 Latent Semantic Analysis

Most of the low level and structured approaches to TM use some sort of similarity judgment to carry out the task itself (i.e., clustering) or to produce some kind of relevant association between different items. Some of them rely in simple statistical correlations measures whereas others are based on more sophisticated forms of learning tasks via self-organization (e.g., SOM).

We concentrate on one of these tools: *Latent Semantic Analysis* (LSA) which look promising as this has long been successfully used in the context of IR, and, as discussed later, has been cognitively validated so as to handle complex human language tasks.

LSA is a mathematical technique that generates a high-dimensional semantic space from the analysis of a huge text corpus. It was originally developed in the context of IR (Berry and Browne, 2001) and adapted by psycholinguistics for Natural-Language Processing tasks (Landauer et al., 1998a).

LSA differs from some statistical approaches for textual data analysis in two significant aspects. First, the input data "associations" from which LSA induces are extracted from unitary expressions of meaningful words and the complete meaningful utterances in which they occur, rather than between successive words (i.e., mutual information, co-occurrence). Secondly, it has been proposed that LSA constitutes a fundamental computational theory of the acquisition and representation of knowledge as its underlying mechanism can account for a long-standing and important mystery, the inductive property of learning by which people acquire much more knowledge than appears to be available in experience (Landauer et al., 1998b).

By keeping track of the patterns of occurrences of words in their corresponding contexts, one might be able to recover the latent structure of the meaning space, this is, the relationship between meanings of words: the larger and the more consistent their overlap, the closer the meanings.

In order to produce meaning vectors, LSA must be trained with a huge corpus of text documents. The initial data are meaningful passages from these texts and the set of words that each contains. Then, a matrix is constructed whose rows represent the terms (i.e., keywords) and the columns represent the documents where these terms occur. The cells of this matrix are the frequencies with which the word occurred in the documents. In order to reduce the effect of the words which occur across a wide variety of contexts, these cells are usually multiplied by a global frequency of the term in the collection of documents (i.e., logarithmic entropy term weight) (Berry and Browne, 2001).

These normalized frequencies are the input to LSA which transforms them into a high-dimensional semantic space by using a type of principal components analysis called *Singular Vector Decomposition* (SVD) which compresses a large amount of co-occurrence information into a much smaller space. This compression step is somewhat similar to the common feature of neural networks where a large number of inputs are connected to a fairly small number of hidden layer nodes. If there are too many nodes, a network will "memorize" the training set, miss the generality of the data, and consequently perform poorly on a test set. Otherwise, this will tend to "capture" the underlying features of the input data representation.

In order for SVD to perform this compression, an approximation of the initial `T x D` terms-by-document matrix of an arbitrary rank K is computed. For this, the original matrix is decomposed into three matrices: a `U x D` documents matrix, a `K x K` singular values matrix, and a `K x T` terms matrix. Multiplying these matrices together results in an approximation to the original matrix. As we wish to compute the meaning vectors with a lower dimension so as to reduce the effects of missing and noisy data, and carry out semantic judgements in a space with smaller dimensionality, this multiplication must be made using a desired value of dimensionality K (the reduction of the number of vector's components of the observed data from the number of initial contexts to a much smaller but still large number).

Common values of K are between 200 and 500, and are thus considerably smaller than the usual number of terms or documents in the used corpus. These values have been chosen from different experiments using LSA. The results of using these ex-

perimental values in similarity judgements have proved to correlate well with human predictions.

This compression is said to capture the semantic information which is latent in the corpus itself, that is, the captured regularities in the patterns of co-occurrence across terms and across documents are related to the semantic structure of the terms and documents (Wiemer-Hastings and Zipitria, 2001). In other words, this approximation to reduce to a rank K means that LSA will discard all of the excess information and focus only on the essential semantic information in the corpus.

This means that if two terms have similar patterns of occurrences across documents, the decomposition would reveal this fact by giving high similarity values when comparing the semantic vectors of both terms. Note that unlike other decomposition methods (i.e., Principal Component Analysis, QR), this relatedness between vectors also applies to the semantic space of the documents as these can also be represented in the semantic space.

For traditional IR applications, user-built queries are compared to documents from the text database (vectors corresponding to the columns of the matrix V) so as to retrieve relevant documents. A usual comparison measure to calculate the relatedness is the cosine of the angle between the two vectors, so the higher the cosine, the more similar the query is to the document.

LSA has also been successfully applied to an important number of natural-language tasks in which the correlations with human judgements have proved to be promising, including the treatment of *Synonymy* (Landauer et al., 1998b), Tutorial dialog management (Graesser et al., 1999; Wiemer-Hastings, 1999), *Anaphora resolution* (Klebanov, 2001), and text coherence measurement (Foltz et al., 1998).

In terms of measuring text coherence, the results have showed that the predictions of coherence performed by LSA are significantly correlated with other comprehension measures (Dijk and Kintsch, 1983) showing that LSA appears to provide an accurate measure of the comprehension of the texts. In this case, LSA made automatic coherence judgements by computing the similarity between the vector corresponding to consecutive passages of a text. LSA predicted comprehension scores of human subjects extremely well and so it provided a characterization of the degree of semantic

relatedness between the segments.

Despite this success in building LSA-intensive applications and dealing with a wide range of complex tasks, there is still an open issue: LSA considers only patterns of word usage, syntactic or rhetorical structure are not taken into account.

In the context of BOW approaches to TM, this issue is specially relevant. If LSA is to be used in finding interesting associations, one would expect as discussed in section 2.2, that the method would use deeper knowledge (i.e., predicate-level information) so as to provide meaningful relationships or explanations about what it has discovered.

To this end, Wiemer-Hastings and Zipitria (2001) propose a simple method to add syntactic knowledge to LSA by using part-of-speech (POS) labels and simple syntactic rules, in the context of dialog management in the AUTOTUTOR tutorial system (Graesser et al., 1999).

In order to process a student's answer in these dialogs, his/her answer is semantically compared via LSA to those expected and previously stored by the system. It might be the case that an answer is highly semantically related to the expected one but when one looks at the structure of this, the sentence has a different meaning. Since LSA is not affected by the ordering of the words in the sentences, this kind of distinction can not be made. In order to deal with this, the sentences are represented by using the initial words, POS labels and some basic syntactic rules. The resulting semantic vector for a sentence is simply the sum of the usual elements (terms). However, in making similarity judgements the POS labels are considered in such a way that the answer and the expected answer are matched only if their individual elements (i.e., nouns, verbs, adjectives) are roughly the same.

A feature of this kind of representation for LSA is that the semantic vector for a sentence is obtained by just computing the sum (or centroid, in some cases) of the elements' vectors. Computing the semantic representation of the sentence does not depend on the context on which this occurs.

Kintsch (Kintsch, 2001) proposes a method for representing predicate-level information of a sentence in which, unlike Wiemer-Hastings and Zipitria (2001), the meaning vectors take into account the different contexts (semantic spaces) where this predicate information is used. For example, the term "run" may have different senses

depending on the context where they are used, such as "the paper ran out", "the program runs", "the robber ran away", and so on. However, there are contexts where the terms are used in a sense which is close to the item of interest (predicate). The method suggests that only those terms which are close to the predicate sense should be used for representing the predicate. This set of terms (usually *K items*) is referred to as the predicate's semantic neighbourhood and are obtained through a net-based activation mechanism called the *Construction-Integration* (CI) model (Kintsch, 1998).

An algorithm to determine the semantic vector according to its semantic neighborhood is proposed. The algorithm known as *Predication*, basically performs two key activities. First, the semantic neighbours of the predicate are obtained through a process of activation of the link that connect the items in which connections between the predicate and and its semantic space are inhibited or activated depending on the strength they have on the predicate. Then, these relevant items are weighted and used to compute the predicate's final vector. In order to capture context dependence, the algorithm strengthens features of the predicate that are also appropriate for the argument of the predication. Items of the semantic space of the predicate that are relevant to the argument are combined with the predicate vector in proportion to their relevance through a spreading activation process carried out by the CI model.

This has the effect of highlighting only those aspects of the predicate that are relevant to the argument of the predicate. Hence a different sense of the predicate emerges every time it is used in a different context.

Specifically, let $P(A)$ be a given proposition of interest, where P (predicate) and A (argument) are items of the LSA semantic space represented by term vectors. Let S be the set of all items in the semantic space except for P and A. The terms I (nouns, for the purpose of Predication) in S can be arranged in a space around P and their similarity to P determines how close or far they are.

The similarity $sim(P,I)$ will represent the LSA closeness between predicate P and item I. In the same way, $sim(A,I)$ will be the similarity between A and item I. Next, a network involving the nodes P, A, and all I in S can be constructed, where one set of links connects A with all other nodes.

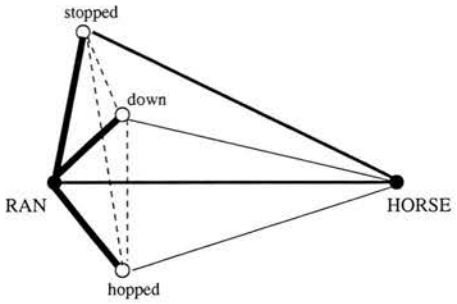

Figure 2.3: Integration network for predicate RAN(HORSE) (taken from (Kintsch, 2001))

The strength of these links is codetermined by how closely related they are to both A and P, that is:

$$strength(A,I) = f(sim(A,I), sim(P,I)) \tag{2.1}$$

where a function f must be chosen in such a way that $strength(A,I) > 0$ only if I is close to both P and A. A second set of links connects all items I in S with each other. Next, the CI algorithm is applied and links in the net are inhibited/activated (initially, the activation values correspond to strength values between items but they are modified as the CI algorithm goes through the net). From this, the items with the most strongly "activated" links in the net will be those from the neighbourhood of P that are in some way related to A. These items are then used in the construction of the vector of $P(A)$ by performing the weighted average of the k most activated items, including P and A, where the weights are the activation values of the nodes.

An example of this integration net for a predicate `RAN(HORSE)` whose representation is to be obtained can be seen in figure 2.3. The integration process will select items that are close to `ran`, but also relevant to the argument `horse`: in this case, `ran` and `stopped` will be most strongly activated, `down` will be somewhat activated, and `hopped` will receive no activation.

For the purpose of the meaning representation, the experimental tests and analyses carried out indicate that in some cases, *Predication* gives intuitively more adequate results than using context-independent measures (i.e., centroid, sum). However, these details may not weigh very much when it comes to the meaning of longer passages than a sentence, such as an essay.

2.2.2.2 Information Extraction

With the exception of BOW-based methods to TM, the rest of the approaches involve some sort of information extraction task, whether it is traditional in which some patterns are designed to extract linguistically motivated information, or some sophisticated system involving specific tasks focused on diverse levels of analyses (i.e., syntactic, semantic, etc).

In general, *Information Extraction* (IE) refers to the activity of automatically extracting pre-specified sorts of information from natural language texts (Gaizauskas and Wilks, 1997) with a homogeneous representation so to be summarized and presented in a uniform way.

IE systems do not attempt to understand the text in the input documents, but analyse those portions of each document that contain relevant information. Relevance is determined by predefined domain guidelines which specify what types of information the system is expected to find.

There are two main approaches to the design of IE systems (Appelt and Israel, 1999): the *Knowledge Engineering Approach* (KEA) and the *Automatic Training Approach* (ATP). In KEA a grammar expressing rules for the system are constructed by hand using knowledge of the application domain. The skill of the knowledge engineer plays a large role in the level of performance of the system. However, the development process can be very laborious, and sometimes the required expertise may not be available. For ATP, there is not the same need for system expertise when customizing the system for a new domain. Instead, someone with sufficient knowledge of the domain and the task at hand annotates a set of training documents. Once a training corpus has been annotated, a training algorithm is run, training the system for analysing new texts. This approach is faster than the KEA, but requires that a sufficient volume of training data is available.

A key element of IE systems is a set of text extraction rules or *extraction patterns* that identify the relevant information to be extracted. In order for these patterns to be effective, IE must pay attention to the structural or syntagmatic properties of texts to be able to deal with sophisticated entities (e.g., NP, role players, events, etc), grammatical variation, lexical variation, and cross-sentence phenomena such as anaphora.

According to (Grishman, 1997), two major steps can be distinguished in an IE process. First, the system extracts individual "facts" from the text of a document through local text analysis. Secondly, it integrates these facts, producing larger facts or new facts (through inference) which are then translated into the required output format (*templates*).

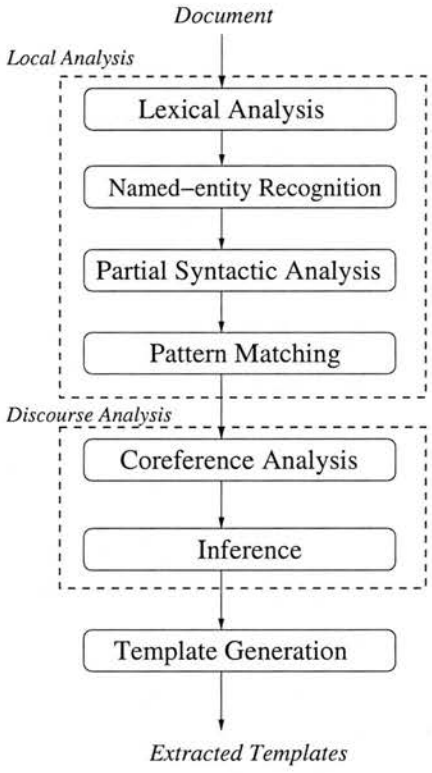

Figure 2.4: Structure of an IE system (adapted from (Grishman, 1997))

Although the different components vary from one system to another, typically IE systems consists of a sequence of modules (figure 2.4) as follows:

- **Lexical Analysis:** here the different components of the input text are tokenised, and then assigned their possible Parts-Of-Speech (POS) and features. Some systems (Gaizauskas and Wilks, 1997) additionally use facilities of gazette lookup to facilitate the process of recognizing and classifying named entities (e.g., organization names, location names, etc).

- **Named-Entity recognition:** this phase identifies types of proper names and other special forms such as dates, names, etc. commonly identified by a set of simple patterns (regular expressions) which are stated in terms of POS, syntactic features, orthographic features, etc. For example, names might be identified by a preceding title ("Mr. James King"), by a common first name ("James King"), and so on.

- **Partial Syntactic Analysis (Parsing):**

  Some form of partial parsing is performed to identify noun groups, verb groups, etc, which aims to simplify the subsequent phase of fact extraction. Some systems do not have any separate phase of syntactic analysis. Others attempt to build a complete parse of a sentence. However, most systems fall somewhere in between, and build a series of parse fragments. In some cases, a semantic interpretation is built up during parsing (Humphreys and Gaizauskas, 2000). For example, the sentence "*Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm*" can be parsed and assigned a top level structure (e.g., BP, NP, etc). From the identified structural relations, a logical form may be assigned which may take the form:

```
person(e21)  name(e21,'Don Wright')  name(e22), lobj2(e22,e23)
title(e23,'executive president')     firm(e24), det(e24,this)
```

  which is later used for discourse processing tasks.

- **Pattern Matching:** the goal of the patterns is to extract the events or relationships relevant to the domain/scenario. While IE patterns are usually defined and coded manually, there are also a significant number of efforts aimed at learning these patterns automatically from a training corpus (Riloff and Lorenzen, 1999; Califf, 1998). Whether these are handcrafted or automatically acquired, the patterns usually try to capture the information of interest such as "<person> is killed by <person>", "<person> is killed by <organisation>", where <person> and <organisation> are pattern elements which match NPs with the associated type, and other kinds of linguistic features.

- **Coreference Analysis:** aims at resolving anaphoric references by pronouns and definite noun phrases, and it is usually seen as a part of a major discourse interpretation activity. To this end, some approaches have a world model consisting of an ontology plus an associated attribute knowledge base on which the analysed information can be classified, matched, etc.

- **Inferencing and Event Merging:** in many situations, partial information about an event may be spread over several sentences; this information needs to be combined before a template can be generated. In other cases, some of the information is only implicit, and needs to be made explicit through inference mechanisms.

- **Template Generation:** as a result, the set of extracted and/or inferred templates are generated, such that these represent the facts of interest. For example, a typical event template may look like:

```
EVENT: leave job                    PERSON: James King
POSITION: vice president            COMPANY: BT plc.
```

Despite successful results in current IE systems, practical research over the last years has also raised important issues which should be taken into account for achieving more adaptive IE systems and user-driven IE. According to (Wilks and Catizone, 1999) the main issues concern the fact that the quality of the system depends partly on the quality of the training data it is provided with. This makes the provision of tools to involve users in this process as part of their normal workflow important. On the other hand, the type of learned data structures impact the maintainability of the system. Stochastic models, for example, perform well in certain cases, but cannot be hand-tailored to squeeze out a little extra performance, or to eliminate an obvious error.

From the IE patterns viewpoint, recent advances make possible, in most cases, an unsupervised notion of template learning in which statistically significant words are located in a corpus, and used to locate the sentences in which they occur as key sentences. This has been the basis of a range of summarisation algorithms. Although the implementations cannot be considered to have proved that such learning is effective, some promising prototype results have been obtained.

2.2.2.3 Genetic Algorithms

The discussed approaches to TM/KDT use a variety of different "learning" techniques. Except for cases using Machine Learning techniques such as Neural Networks (e.g., SOM), decision trees, and so on, which have also been used in traditional DM, the real role of "learning" in the systems is not clear. There is no learning which enables the discovery but instead a set of primitive search strategies which do not necessarily explore the whole search space due to their dependence on the kind of semantic information previously extracted.

Although DM tasks have been commonly tackled as learning problems, the nature of DM suggests that the problem of DM (i.e., finding unseen, novel and interesting patterns) should be seen as involving search (i.e., different hypotheses are explored) and optimization (i.e., hypotheses which maximize quality criteria should be preferred) instead.

Despite there being a significant and successful number of practical search and optimization techniques (Mitchell, 1996; Deb, 2001), there are some features that make some techniques more appealing to perform this kind of task than others, in terms of representation required, training sets required, supervision, hypothesis assessment, robustness in the search, etc.

In particular, the kind of evolutionary computation technique known as *Genetic Algorithms* (GA) has proved to be promising for search and optimization purposes. Compared with classical search and optimization algorithms, GAs are much less susceptible to getting stuck to local suboptimal regions of the search space as they perform global search by exploring solutions in parallel. GAs are robust and able to cope with noisy and missing data, they can search spaces of hypotheses containing complex interacting parts, where the impact of each part on overall hypothesis fitness may be difficult to model (Goldberg, 1989).

In order to use GAs to find optimal values of decision variables, we first need to represent the hypotheses in binary strings (the typical pseudo-chromosomal representation of a hypothesis in traditional GAs). After creating an initial population of strings at random, genetic operations are applied with some probability in order to improve the population. Once a new string is created by the operators, the solution is evaluated

in terms of its measure of individual goodness referred to as *fitness*.

Individuals for the next generation are selected according to their fitness values, which will determine those to be chosen for reproduction. If a termination condition is not satisfied, the population is modified by the operators and a new (and hopefully better) population is created. Each interaction in this process is called a *generation* and the entire set of generations is called a *run*. At the end of a run there is often one or more highly fit chromosomes in the population.

In a simple GA, three basic components can be distinguished: a Population update mechanism which is resposible for selection of fit individuals for reproduction, the genetic operators (Crossover and Mutation), which are responsible for producing new offspring, and the fitness evaluation (see figure 2.5):

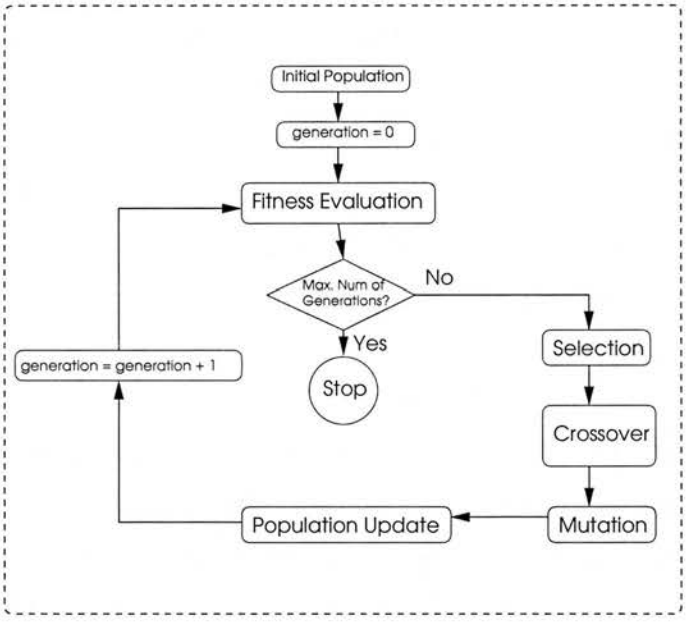

Figure 2.5: Structure of a Single GA

- *Population Management:* this component aims at selecting duplicates of good solutions for reproduction via the genetic operators, and at updating the population once the offspring have been produced.

  Selection mechanisms basically make duplicates of good solutions, while keeping the population size constant (by eliminating bad solutions in a population)

(Deb, 2001). This is usually achieved by identifying good solutions in a population, and then making multiple copies of good solutions.

The selection of individuals may be implemented in a number of ways. Some common methods include *tournament selection*, *proportional selection*, and *ranking selection*. In tournament selection, tournaments are played between two solutions and the better solution is chosen and placed in the mating pool. Two other solutions are picked again and another slot in the mating pool is filled with the better solution. If carried out systematically, each solution can be made to participate in exactly two tournaments. The best solution in a population will win both times, thereby making two copies of it in the new population. Using a similar argument, the worst solution will lose in both tournaments and will be eliminated from the population. In this way, any solution in a population will have zero, one or two copies in the new population.

In proportional selection (e.g., *roulette-wheel* mechanism), solutions are assigned copies, the number of which is proportional to their fitness values. Using the fitness value of all solutions, the probability of each solution can be calculated by adding the individual probabilities from the top of the list. Next, the bottom-most solution in the population has a cumulative probability equal to 1. Thus, a string corresponding to the chosen random number in the cumulative probability range for the solution is copied to the mating pool.

In ranking selection, the solutions are first sorted according to their fitness, from the worst (rank 1) to the best (rank N). Each member in this sorted list is assigned a fitness equal to the rank of the solution in the list. Thereafter, the proportional selection operator is applied with the ranked fitness values, and N solutions are chosen for the mating pool.

The update of the population is performed by eliminating bad solutions from the population so that multiple copies of good solutions can be replaced in the population.

Most methods used for replacement have been "generational", that is, at each generation the new population consists entirely of offspring formed by parents

in the previous generation. However, in some schemes, successive generations overlap to some degree– some portion of the previous generation is retained in the new population (Mitchell, 1996) and so only part of the new population needs to be updated. In this non-generational scheme, the fraction of new individuals at each generation has been called the "generation gap".

Some popular non-generational replacement strategies have included *steady-state* and *elitism*. In steady-state replacement, a small number of the least fit individuals are replaced by offspring resulting from the fittest parents (Whitley, 1989; Whitley, 1994; Mitchell, 1996). This is specially useful in evolving rule-based systems in which incremental learning and remembering what has already been learned is important. Elitism is a method that forces the GA to retain some number of the best individuals at each generation (Mitchell, 1996; Deb, 2001). Such individuals can be lost if they are not selected to reproduce or if they are destroyed by crossover or mutation.

- *Crossover:* since selection cannot create new solutions in the population, a recombination operator, *Crossover*, is introduced. In the simple case (*single-point crossover*), two strings (hypotheses) are picked from the mating pool (population) based on their fitnesses, and then some portions of these strings are exchanged to create two new strings Specifically, a single crossover position is chosen at random and the parts of two parents after the crossover position are exchanged to form two offspring (other variations of simple crossover are discussed in (Goldberg, 1989; Michalewicz and Fogel, 1999)).

- *Mutation:* Mutation is needed because even though crossover effectively searches and recombine hypotheses, occasionally it may become overzealous and lose some potentially useful genetic material.

  By itself, mutation is a random walk through the string space. When used sparingly with crossover, it is an insurance policy against premature loss of important information (Goldberg, 1989; Mitchell, 1996). In a simple GA, mutation is the occasional (with small probability) random alteration of the value of a string position. In binary coding, this means changing the value of a 1 bit of a string to 0

or vice versa.

An improvement is not guaranteed during a GA generation. However, it is expected that if bad strings are created, they will be eliminated by the selection method in subsequent generations and if good strings are created, they will be emphasized.

GAs in Data Mining and Text Mining

One of the major contributions of evolutionary algorithms (e.g., GAs) for an important number of DM tasks (e.g., rule discovery, etc) is that they tend to cope well with attribute interactions. This is in contrast to the local, greedy search performed by often-used rule induction and decision-tree algorithms (Berthold and Hand, 2000; Han and Kamber, 2001). Most rule induction algorithms generate (prune) a rule by selecting (removing) one rule condition at a time, whereas evolutionary algorithms usually evaluate a rule as a whole via the fitness function rather than evaluating the impact of adding/removing one condition to/from a rule. In addition, operations such as crossover usually swap several rule conditions at a time between two individuals.

Typical tasks for GAs in DM have included (Freitas, 2001a; Williams, 1999): *Classification*; in which the goal is to predict the value (the class) of a user-defined goal attribute based on the values of other attributes; *Discovery of Association rules*; where binary attributes (items) contained in data instances (i.e., records) are used to discover associations of the form IF-THEN, *Rule discovery/prediction*; in which the system can produce many different combinations of attributes. (even if the original attributes do not have much predictive power by themselves, the system can effectively create "derived attributes" with greater predictive power) to come up with new rules.

A common representation used for this kind of task encodes attributes and values of a rule in a binary string of rule conditions and rule consequent. Suppose that an individual represents a rule antecedent with a single attribute-value condition, where the attribute *Marital_status* and its values can be "single", "married", "divorced" and "widow". A possible representation would be a condition involving this attribute encoded by four bits, so the string "0110" (i.e., the second and third values of the attribute are present) would represent the antecedent *IF marital_status=married OR divorced)* using internal disjunctions (i.e., logical OR).

Conjunctions can be represented by just including an appropriate number of bits to represent each attribute-value condition. Other schemes for representing categorial data can easily be extended. For this, the bits are used to represent the value of the attribute in binary notation, e.g., "00001101" can represent the value 13 for an integer-valued attribute. Similar representations can also be used to represent more complex attribute values. Note however that as we are looking for comprehensible rules, the specific binary representation used has often to be traded off with the size of the rule.

Although this kind of approach constitutes an efficient way to represent the hypotheses in traditional GAs, it is also suggested (Freitas, 1997) that a high-level symbolic representation for a rule such as:

```
IF (marital_status="married OR "divorced") AND (Age>21)
THEN approval="yes"
```

would be more desirable in terms of comprehensibility, uniform treatment of attributes, and ease to apply adapted genetic operators. This is the case, for example, for the representation for data and programs used in the context of Genetic Programming (Koza, 1992).

There are several proposals for genetic operators which deal with this representation for tasks such as rule discovery (Berthold and Hand, 2000). In some of them, *Selection* is performed by electing individuals (prediction/classification rules) of the population to be mated according to the training examples, that is, the more examples they cover (aka. votes), the better the individual (rules covering the same examples compete with each other). More precisely, the probability of voting for a given individual is proportional to the fitness of that rule. Hence, this procedure effectively implements a form of niching, encouraging the evolution of several different rules, each of them covering a different part of the data space.

Crossover and mutation operations are designed such that rule generalization and specialization are enabled. In binary representations, mutation can be accomplished by turning bits off or on, depending on whether one wishes to generalize (cover more examples) or to specialize. In symbolic representations, the generalization operation can be dealt with by simply deleting one of the conditions, or by adding/subtracting a

small randomly-generated value to an attribute's value. For example, if a small value is subtracted from the first condition in: `(Age>25) AND (Marital_status="single")`, then, the following condition is produced: `(Age>21) AND (Marital_status="single")`, which tends to cover more examples in the database than the former.

Generalization and specialization for crossover can be implemented as the logical OR and the logical AND, respectively. For example, the following parents: "`0|10|1`" and "`1|01|0`" can be used in a two-point crossover with the points denoted by the symbol "`|`". The children produced by the operator would look like:

```
Generalizing Crossover:          Specializing Crossover:
0|11|1      1|11|0               0|00|1      1|00|0
```

From the evaluation viewpoint, the quality of the discovered rules is directly assessed through the fitness function in terms of metrics such as the rule predictive accuracy, its comprehensibility, and its interestingness (Jaroszeqicz and Simovici, 2001; Hilderman, 1999; Noda et al., 1999). Accuracy (see section 2.1.1) usually is defined as a confidence factor which is proportional to the number of training examples satisfying the conditions and consequent of the rule (Freitas, 2001a; Radcliffe and Surry, 1994), provided that these examples are available.

One general aspect worth noting in applying GAs for DM tasks is that both the representation used for the discovery and the evaluation carried out assume that the source data are properly represented in a structured form (i.e., database) in which the attributes and values are easily handled.

When dealing with text data, these working assumptions are not always plausible because of the complexity of text information. In particular, mining text data using evolutionary algorithms requires a certain level of representation which captures knowledge beyond discrete data (i.e., semantics). Thus there arises the need for new operations to create knowledge from text databases. In addition, fitness evaluation also imposes important challenges in terms of measuring novel and interesting knowledge which might be implicit in the texts or be embedded in the underlying semantics of the extracted data.

Applying evolutionary methods to TM/KDT is a very recent research topic. With the exception of the work of (Bergstron et al., 2000) on the discovery of semantic

relations no other research effort is under way as far as we know as the most promising KDT techniques have been tackled with more traditional search/learning methods (see section 2.2).

Bergstron (Bergstron et al., 2000) proposes a system to automatically evolve patterns for extracting semantic relations from Web text using a variation of GAs known as *Genetic Programming* (Koza, 1992; Banzhaf et al., 2000). Unlike (Hearst, 1998), the main aim is to extend the WordNet hierarchy by matching IE patterns that can be automatically be learned. The proposed method is capable of finding patterns expressing relations of interest such as meronyms and hyponyms (Fellbaum, 1998), as in "`X is a part of Y`", "`X such as Y`", etc.

Genetic Programming (GP) is used for automatically learning feasible IE patterns so as to capture semantic relations of interest. Initially, each individual randomly makes a relationship by picking a pair of words and the name of the relation, expressed by a couple of words (e.g., `<first> "such as" <last>`). Next, randomly constructed trees of the evolved patterns are generated through GP. The fitness of these patterns is then measured in terms of their coverage in the WordNet hierarchy: if a pair of concepts matches an instance of a hyponym relation in WordNet, it is assigned a positive fixed score. If the pair matches an instance of another relation in Wordnet, it is scored a fixed negative value. If no match is found, the pair is given no points. At the end, all the scores for an individual are added and used as its fitness for competition. The obtained fittest individuals (evolved patterns) are selected to extract information of interest from the texts.

The advantage over a similar approach for discovery of unseen relation as in (Hearst, 1998), is that this approach provides more robust results in a way that exploits a wider number of possible hypotheses in the search space. In addition, the IE patterns finally used for the extraction are automatically learned, whereas for (Hearst, 1998), these need to be handcrafted. Although the obtained relations have been evaluated in terms of their coverage in WordNet, the subjective quality of this unseen knowledge has not been assessed from a KDD viewpoint as no user has been involved in the process.

Note that whether the problem involves TM or traditional DM using some form of evolutionary algorithm, there is a common issue: all of the optimization and search

strategies reduce the measure of quality (e.g., interestingness, novelty, simplicity, accuracy) to a single criterion. This has been transformed to a fitness function form, and researchers have thereafter proceeded with the genetic operations. For instance, in the context of prediction rules, (Freitas, 2001a) measures the interestingness of a rule by using its Confidence Factor seen as the predictive accuracy of the rule (CF), the Completeness (Comp) of the rule, and the simplicity of the rule. The fitness value is then some linear combination of these criteria as in *"w_1*(CF * Comp) + w_2*Simp"*, where w_1 and w_2 represent the used-defined weights, usually determined according to their degree of importance. Others such as (Radcliffe and Surry, 1994) have coped with this aggregation of objective by using nonlinear combination (e.g., logarithmic). However, beyond merely statistical justifications, the use of this kind of aggregation and its influence in objectively measuring the quality of the hypotheses from a KDD perspective, is not clear. In fact, the evaluation has not proved to correlate in any way with human judgement.

Although this kind of approach may seem to work well in many problems, there are times when several criteria are present simultaneously and it is not possible (or wise) to combine these into a single number (Goldberg, 1989; Deb, 2001). When this is the case, the problem is said to be a *multi-objective* or *multi-criteria* optimization problem.

Evolutionary Multi-Objective Optimization

Most interesting and real-life problems are multi-objective in nature, in that they have several (possible conflicting) objectives that must be satisfied at the same time. In this context, the notion of "optimum" has to be re-defined and instead of aiming to find a single solution, we will try to produce a set of good compromises or "trade-offs" from which the decision maker will select one.

Multi-objective optimization (Coello, 2000; Deb, 2001), can then be defined as the problem of finding a vector of decision variables which satisfies constraints and optimizes a vector function whose elements represent the objective functions, that is, $F(x) = [f_1(x), f_2(x), ..f_k(x)]$. As it is rarely the case that there a single point that simultaneously optimizes all the objective functions, we normally look for trade-offs rather

than single solutions. The notion of optimum is therefore different and usually referred to as a *"Pareto Optimum"* (Stadler, 1988).

A vector of decision variables x^* is said to be a Pareto *optimal* iff:

$$\nexists x \neq x^* \mid f_i(x) \textbf{ is worse than or equal to } f_i(x^*)\ \forall i = 1,...k$$
$$\wedge\ f_j(x) \textbf{ is worse than } f_j(x^*) \text{ for at least one j}$$

In words, assuming a minimization problem (i.e., "worse" involves smaller values), x^* is a Pareto optimal if there exists no feasible vector x which would increase some criterion (objective) without causing a simultaneous decrease in at least one other criterion. Unfortunately, this concept almost always gives not a single solution, but rather a set of solutions called the *Pareto Optimal set.* The vectors x^* corresponding to the solutions included in the Pareto optimal set are called non-dominated. The space of the objective functions whose nondominated vectors are in the Pareto optimal set is called the *Pareto front* (Coello, 2001; Deb, 2001; Fonseca and Fleming, 1995).

The use of this concept in evolutionary algorithms to solve multi-objective optimization problems (also known as *Evolutionary Multi-objective Optimization*) has been mainly motivated by the fact that GAs simultaneously deal with a set of possible solutions. This allows us to find several members of the Pareto optimal set in a single run of the algorithm, instead of performing a series of separate runs.

Evolutionary Multi-Objective Optimization (EMOO) techniques range from simple aggregation functions, in which the objectives are combined in a single function by determining the relative weight of each objective in the fitness function, to those measuring fitness in terms of portions of regions of the search space that are dominated (Coello, 2001; Coello, 2000; Deb, 2001).

Note that because of the nature of the textual data used in our model, we are interested in possible optimization approaches that are independent on the chromosome representation (i.e., the current approaches rely on a direct binary representation of hypotheses) in terms of fitness assignment, and therefore adaptable to more complex kind of data and operations. In KDD, it is desirable to provide not just a unique solution but a set of hypotheses which convey novel knowledge worth exploring. As simple GAs

tend to provide mostly single solutions, we also have an interest in additional strategies that can promote the formation of groups of solutions (e.g., niches).

Traditionally, GAs do not encourage the formation of niches as they rely on the fact that, in general, only one optimum solution is searched for. For many real-life applications though, multiple near optimum solutions are often required. This is the case, for example, of DM in which the system must come up with a set of solutions which meet the quality criteria[1].

Basically, common niching strategies include *"Sharing"* and *"Crowding"*. *Sharing* aims to share the fitness of the fittest individual with other individuals with a lower fitness. In this scheme, promoting several "peaks" of the fitness in the population is performed by updating the fitness of the close individuals as a function that depends on the best individual fitness, and some accumulated function of share which involves the distances between the best individual and the others (Goldberg, 1989). Whereas *Crowding* involves an overlapping population in which individuals replace a group of existing strings according to their similarity which is measured on the basis of a bit-by-bit similarity count.

Note that in both strategies, the "close" individuals are chosen according to a similarity measure which uses the underlying individual binary representation. However, if a problem uses a non-binary representation, the strategy can not be applied as it is, hence new (or adapted) mechanisms to deal with the individuals are needed.

An EMOO technique which deals with the diversity of the solutions (i.e., niche formation) and the fitness assignment as a whole in a representation-independent way is known as the *Strength Pareto Evolutionary Algorithm* (SPEA) (Zitzler and Thiele, 1998a). An attractive feature of SPEA in this regard is that in order to create niches, this does not define a neighborhood by means of a distance metric on the genotypic or phenotypic space. Instead, the classes of solutions are grouped according to the results of a clustering method which uses the vector of objective functions of the individuals, and not the individuals themselves.

[1]The need for niches comes mainly from evolutionary methods based on the Pittsburg approach which corresponds to the described way of functioning of a GA, that is, the solutions are represented by individuals that fight each other. In the Michigan approach, each individual represents a set of solutions instead, and therefore new ways of assessing the individuals are needed (Mitchell, 1996; Goldberg, 1989).

SPEA proposes an elitist evolutionary strategy in which the best individuals of the current population (elites) are combined with a previously computed group of non-dominated solutions so as to maintain a consistent set of good solutions. This external set is referred to as the "external population" (strictly speaking, *"external" Pareto set*) and stores a fixed number of the non-dominated solutions that have been found up to the beginning of a run of the current population (at the first time, this will only contain the non-dominated individuals of the initial population).

At every generation, newly found non-dominated solutions in the current population are compared with the existing external set[2] the non-dominated individuals (potentially new elites) and the resulting non-dominated solutions are preserved. This combination means that old individuals might be removed and new ones (i.e., elites) might be added so as to keep this external population consistent in terms of dominance conditions.

As this updated external Pareto set can be overcrowded with non-dominated individuals, the size of this is bounded by a limit. For this, a clustering method is proposed to maintain a fixed size while keeping good individuals in the set. The clusters are computed according to the distance between the objective vectors of the solutions. Once these are formed, one solution from each cluster is chosen and the others are removed from the clusters to join the current population. The chosen solution in every cluster is the one having the minimum average distance from other solutions in its cluster.

Note that the overall aim of the algorithm is to carry out the fitness assignment providing that the current population and the external population (external Pareto set) have been updated (details of the whole evolutionary strategy that considers this procedure and the working of the genetic operations are discussed in (Zitzler and Thiele, 1998a)). For simplicity's sake, the algorithm has been fragmented into two sections: one highlighting the updating step, and the other describing the steps needed to have the fitness values computed. First, the details of the algorithm for updating the sets is shown:

[2]The Pareto set is considered "external" so to make clear the difference between the set of non-dominated individuals (elites) obtained so far, and that containing non-dominated individuals from the whole current population. Nevertheless, both non-dominated groups contain individuals which are part of the population.

```
ALGORITHM FitnessAssignment (Part 1 of 3)
INPUT: CurrentPopulation, ExternalParetoSet
OUTPUT: Updated ExternalParetoSet, fitness

A ← CollectNonDominatedSolution(CurrentPopulation)
B ← CombineParetoSet(ExternalParetoSet,A)
IF | B |>Maximum number of Pareto points THEN
        ExternalParetoSet ← ReduceParetoSet(B)
ELSE    ExternalParetoSet ← B
END-IF
```

Where *CollectNonDominatedSolution(CurrentPopulation)* collects and returns the non-dominated individuals of the current population. *CombineParetoSet(ExternalParetoSet,A)* basically combines the external Pareto Set and the group of non-dominated individuals of the current population. Finally, *ReduceParetoSet(B)* performs the clustering and then updates the current population.

Next, since that so far there is no single fitness value for an element of the population, an approximation of the fitness in terms of dominance relations is computed by using the current population and the external population so as to enable further selection of fittest solutions. This will enable individuals to be selected for breeding. As a Pareto set represents the group of fit solutions so far, the fitness for the individuals of the current population and those from the external population are calculated differently.

As (external) Pareto members should be preferred as solutions, these are assigned better fitness values, where "better" means low values in a range between 0 and 1. This individual fitness is based on Holland's *strength* (Holland, 1992) in which the fitness (strength) is proportional to the number of members of the current population that the current individual dominates. The details of the assignment are as follows:

```
ALGORITHM FitnessAssignment (Part 2 of 3)
/* Compute fitness of Pareto members (strength) */
FOR ParetoInd in ParetoSet DO
    count ← 0
    FOR Popind IN Population DO
        IF covers(ParetoInd,PopInd) THEN
            count ← count + 1
        END-IF
    END-FOR
    strength ← count/(| Population | +1)
    fitness(ParetoInd) ← Strength
END-FOR
```

where $covers(A,B)$ is a boolean function which determines whether an individual A dominates another individual B. The resulting fitness (strength) for this kind of individual is a real value in the range $[0,1)$. Given this, the individuals of the current population ("bad" solutions) should be assigned a fitness value equal or larger than 1 so as not to be preferred as desired solutions.

Note also that individuals of the current population that are less dominated by Pareto members are good candidates to be improved as they are close to the Pareto front. Indeed, the quality of the Pareto set can be improved by dragging "less bad" individuals from the current population into the Pareto set (if the dominance conditions permit). Keeping this in mind, the fitness for each of these individuals is calculated as the sum of all the strengths of all Pareto solutions by which it is dominated plus 1 to guarantee that Pareto solutions are most likely to be reproduced (Pareto members will always have lower fitness values). To this end, the final segment of assignment looks as follows:

```
ALGORITHM FitnessAssignment (Part 3 of 3)
/* Calculate fitness for members of the population */
FOR PopInd in Population-ParetoSet DO
    Sum ← 0
    FOR ParetoInd In ParetoSet DO
        IF covers(ParetoInd,PopInd) THEN
            Sum ← Sum + GetFitness(ParetoInd)
    END-FOR
    fitness(PopInd) ← Sum+1
END
```

Overall, it is important to highlight the effect that this assignment has on the selection of individuals. As the GA goes on, the optimization aims at improving the solutions (Pareto set). This can be accomplished by bringing individuals of the current population into the Pareto set. But which "candidates" of the current population have more posibilities to be improved? (i.e., either by improving within the current population or by moving into the Pareto set).

Because of the way the fitness is computed, the fitness of the Pareto individuals are affected by that of the current population and viceversa. Specifically, if Pareto individuals have low fitness (strength) values, then they will contribute positively to the individuals (of the current population) dominated. That is, the sum of the strengths for the covered individual will be low and therefore it will receive a low fitness value (i.e., "better" individual within the current population). On the contrary, if pareto individuals have high strength values, then they will contribute negatively to the individuals (of the current population) dominated, and therefore these individuals will be "bad" dominated individuals (higher fitness values). Consequently, individuals from the current population with lower fitness value (but still dominated) are closer to be improved. Finally, as the individuals for reproduction are usually selected from the Pareto set, those with lower strength values will be preferred.

2.3 Summary

The problems arising from the current approaches to TM/KDT suggest that three essential challenges should seriously be addressed so as to design effective and practical models. These are mostly centered on representation, mining and evaluation:

- *Representation:* new domain-independent and meaningful schemas which capture key information from the text documents should be designed with a practical perspective (i.e., intended to be understood by a normal user) rather than a specific semantic task in mind.

- *Mining:* the process of discovery should be able to deal with information drawn from across the texts to come up with globally learned novel hypotheses. When possible, learned hypotheses should be independent of any external concept models, that is, relying only on the information directly extracted from the texts. In addition, with some exceptions, search techniques in the context of KDT have been underused. The ability of some these to search for global hypotheses, to consider the evaluation as a part of the discovery process, and so on, are key issues that must be addressed in order to come up with effective knowledge.

- *Assessment:* for practical purposes, the novelty and interestingness of the discovered knowledge should effectively be assessed in terms of KDD. Note that the effectiveness of the methods involves putting together two kind of evaluations. First, the system evaluation in which objective criteria are used to measure the quality of the mined knowledge. Secondly, as the usefulness of the discovered knowledge is up to users, the human assessment of the real quality of this knowledge must be considered.

In order to deal with these issues, it is worth exploiting the different tools previously described. LSA may constitute a plausible way to represent lexical-level knowledge. However further work needs to be done to incorporate more structured knowledge. To this end, IE techniques look promising in terms of extracting key contextual information from texts so as to compliment the basic information provided by LSA.

From the search point of view, GAs represent successful methods which may be used for KDT purposes. GA also may benefit from knowledge supplied by LSA and IE techniques in order to search for plausible novel knowledge. However, proper genetic representations need to be designed. Next, new genetic operators should take into account contextual information so as to explore a semantically-rich search space of possible solutions.

Chapter 3

Evolutionary Knowledge Discovery from Texts

In this chapter, we describe a semantically-guided model for evolutionary KDT which does not use any external resource, is genre-based and potentially domain-independent. Unlike previous approaches to KDT, our approach does not rely on external resources or descriptions. Instead, it performs the discovery only using information from the original corpus of text documents and from the training data generated from them. In addition, a number of strategies have been developed for objectively evaluating the quality of the hypotheses.

The aims of our model is to prove that it is plausible to conceive an effective KDT approach independent of domain resources and to make use of the underlying rhetorical information so as to represent text documents for text mining purposes. The specific objectives can be stated as follows:

1. To achieve a plausible search of novel/interesting knowledge based on an evolutionary knowledge discovery approach which makes effective use of the structure and genre of texts.

2. To establish evaluation strategies which allows for the measuring of the effectiveness in terms of the quality of the outcome produced by the model and which can correlate with human judgments.

3. To produce novel knowledge which contributes additional information to help one better understand the nature of the discovered knowledge, compared to BOW approaches.

Generally speaking, Data Mining is reminiscent of the kind of problem in which large databases may contain valuable implicit regularities that can be discovered automatically or of poorly understood domains where humans might not have the knowledge needed to develop effective algorithms. Hence search/optimization techniques (Mitchell, 1997; Knight, 1999; Kodratoff, 2000) may be appropriate for dealing with most of the data mining issues, provided that the representation is handled in a way that it makes it easy to learn new explanatory knowledge and does not restrict the scope of the discovered patterns.

In particular, we have adopted GAs as central to our approach for KDT as they exhibit important advantages from a representation and search viewpoint in knowledge discovery (Freitas, 2001a), including the following:

- As GAs perform global search, they are able to evaluate solutions in parallel so as to avoid local minima.

- GAs allow us to explicitly incorporate problem information to measure the quality of the individuals being generated in terms of their goodness which has an effect on the quality of the solutions finally produced.

- In their core part, GAs constitute a learning technique in which, unlike other approaches (Bratko and Muggleton, 1995; Muggleton, 1999), no training positive/negative examples are required (i.e., they support unsupervised learning). This is highly critical as in general KDT we have no idea what we will discover.

- Although the traditional GA representation is based on binary codings, it has been proved in many applications that GAs also allow more natural representations to the problem, as in prediction rules and others (Freitas, 2001b). Their ease to represent non-numerical data may be especially suitable when dealing with complex data such as natural-language texts.

- GAs are robust methods which can efficiently deal with noisy and missing data. This is specially important for KDT purposes because the initial information extracted from real documents may have missing or misleading information.

Despite these advantages and features, GAs have mainly been proved to be effective techniques for knowledge discovery on structured databases (Freitas, 2001a). Using GAs for mining text databases has received little attention mainly due to the complexity in handling natural-language text and the lack of proper criteria to measure the goodness of the novel patterns. Thus, three main key issues for proper evolutionary KDT need to be addressed:

- *Individual Representation:*

  Using GAs for knowledge discovery purposes is based on the idea of symbolically representing the data in binary and/or discrete forms, so that it is relatively easy to numerically assess the individuals. However, in KDT, there is unrestricted and linguistically rich information (e.g., semantic knowledge, syntactic patterns, etc) which needs to be coded. Even if this can be represented in discrete terms, it would be hard to determine from the representation, for example, whether different concepts are semantically related to each other as this requires contextual knowledge. For this, new representation schemas that capture this kind of knowledge are needed.

- *Guided Genetic Operations:*

  Because of the nature of the data and the individuals being produced, the GA operations cannot be applied in a random way as these may produce incoherent information, so it is necessary to provide guided mechanisms in order to ensure that the offspring being produced are semantically consistent before being evaluated.

- *Fitness Evaluation:*

  The majority of the traditional approaches to TM/KDT have not used any explicit method to assess the quality of the discovered patterns, whereas in a GA context this is a key issue as the evolution is guided by the outcome of the evaluation

of the individuals. This therefore needs to be part of the process itself in a way that captures domain and problem knowledge. For the purpose of dealing with knowledge extracted from texts, new evaluation metrics need to be developed to assess both the plausibility of the hypotheses and their quality from a KDD viewpoint.

Our overall approach integrates *Information Extraction* (IE) technology and a multi-objective GA so as to extract key underlying linguistic knowledge from text documents (i.e., rhetorical and semantic information) and then hypothesize and assess interesting and unseen explanatory knowledge.

In order to deal with the issues discussed in chapter 1 and to achieve an effective KDT process, we developed a two-level model whose architecture can be seen in figure 3.1. This is based on the general process of scientific discovery (Langley, 1987) involving four key steps: collecting/selecting data, finding appropriate data descriptions, formulating explanatory hypotheses, and testing these hypotheses.

The first level is a preprocessing phase aimed at extracting, representing, and computing relevant information from the text corpus which will be later used for feeding the discovery process and evaluating its outcome. The second level constitutes the knowledge discovery itself which takes the form of an evolutionary learning system (e.g., GA) aimed at producing and evaluating the discovered knowledge (i.e., explanatory unseen hypotheses).

The whole processing starts by performing an IE task which applies extraction patterns and then generates a rule-like representation for each document of the corpus. That is, after processing n documents the extraction stage will produce n rules, each one representing the document's content in terms of its antecedents and consequents. Once generated, these rules, along with training data, become the "model" which will guide the GA-based discovery.

In order to generate an initial population to feed the discovery, a set of hypotheses (small size $p << n$) is created by building random hypotheses from the model, that is, hypotheses containing semantic and rhetorical information from the rules are constructed. The GA then runs for a number of generations, and at the end, a small set of the best **K** hypotheses is obtained ($1 < K < p$).

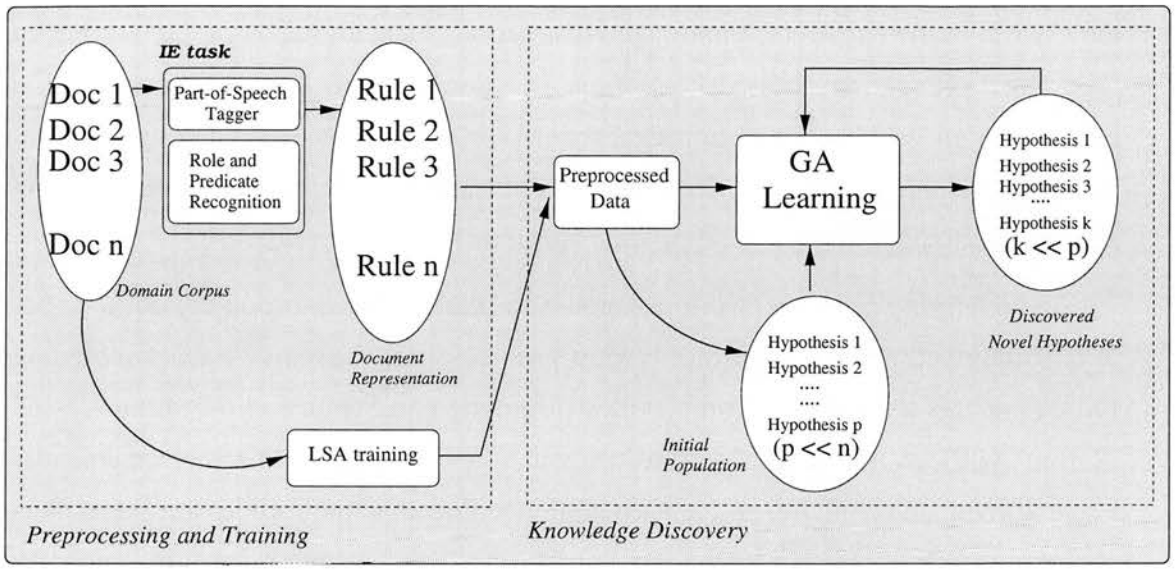

Figure 3.1: Architecture of the GA-based KDT

The description of the model is organized as follows: section 3.1 presents the main features of the preprocessing and training tasks in order to generate the initial knowledge. Then, section 3.2 describes how the discovery is carried out by the GA and section 3.3 proposes some strategies to deal with the automatic evaluation of the discovered hypotheses.

3.1 Text Preprocessing and Training

The preprocessing phase has two main goals: to extract important information from the texts and then, to use this to generate both training data and the initial population for the GA-based discovery.

3.1.1 Extracting Information from the Corpus

Before designing the IE task, we need to address two basic questions: what kind of representation is to be used? And how can the extracted information fit into it?

Because of the intended general purposeness and the statistics-based analysis of-

ten used in current KDT technology, no distinction is usually made in the techniques used when processing different text domains, genres or structures. An exception is some cases in which only technical abstracts are mined due to their simplicity and length (Polanco and Francois, 1998; Ding et al., 2002). The majority of the existing approaches assume that any kind of text can usefully be processed.

An underlying principle in our research is to be able to make good use of the structure of the documents for the discovering process. It is well-known that processing full documents has inherent complexities (Manning and Schutze, 1999; Jurafsky and Martin, 2000), so it is necessary to restrict our scope somewhat. For this, we consider a scientific/technical genre, and in particular, documents in the form of abstracts. This choice is mainly motivated by the fact that a (rough) well-defined structure is used by an abstract's author to describe the ideas and contents in a concise format which is intended to "summarize" what the author states in the full document. In addition, scientific/technical genres avoid many concept-level ambiguities owing to the restricted use of key concepts in a specific contexts. Recent studies by (Ding et al., 2002) on using different kinds of information organization (e.g., abstract, sentences, phrases) for text mining purposes, report that increasing the sophistication from phrases to the full abstract can improve the effectiveness of text mining tasks and related applications (e.g., IR), at least in medical contexts where the applications can benefit from combining different levels of information so as to (automatically) extract interactions of interest in the domain.

Experimental evidence shown by (Manzur et al., 1998) in analysing scientific abstracts by statistical means also suggests that studying the different connections and associations between components of abstracts (e.g., goals, results, etc) can provide interesting insights. In particular, aspects such as misleading results, deficiencies in the results according to the proposed goals can be detected, as well as good features such as having conclusions that answer what is claimed in the goals, stating clear goals which make it easy to find associations with the rest of the components, and therefore allow us to relate other research, to find common methodological issues, etc.

In general, it is suggested that an abstract in some given domain follows a "macrostructure" (i.e., genre-dependent rhetorical structure) which is used by its author to

state the background information, methods, achievements, conclusions, etc. There is also practical evidence that this structure is even modular, (sometimes) hierarchical (Hartley and Benjamin, 1998; Van der Tol, 1998; Kando, 1999) and allows for the organization of the contents of the abstract in an efficient way such that applications such as information retrieval may make more effective use of the documents.

Whether the target application is information retrieval or summarisation there is evidence this it is possible to deal with that structure and the semantic information contained in it, in an effective way. In this context, research by (Teufel and Moens, 1998; Teufel, 1998) shows interesting results in extracting knowledge about the rhetorical structure of a full text so as to automatically produce the abstract. For this, a method is proposed to extract relevant information from the text (e.g., sentences) and then to identify the correct rhetorical roles based on a classification task whose output is the rhetorical annotated structure (Teufel and Moens, 1998).

Unlike information extracted for usual text mining purposes, this macro-structure and its underlying rhetorical information (i.e., roles), are domain-independent but genre-based, so it is relatively easy to translate it into different contexts.

As an example of how the extracted information and the representation is put together in our approach, let's suppose we are given a technical abstract as shown in figure 3.2.

Some IE patterns (along with regular expression which capture specific components) can be designed to capture the underlying rhetorical roles and their semantic information. In the figure, the patterns can extract the key information related to the goal, the object, the method used, and the conclusion. Then, we speculate that this information can be connected as a structure of antecedents and consequents (rule) as shown as the outcome in figure 3.2. Accordingly, important constituents can be identified as follows:

- **Rhetorical Roles:** these indicate important places where author makes some "assertions" about his/her work (i.e., he is stating the goals, methods/means used, achieved conclusions, and so on) or states the scientific evidence (Teufel and Moens, 1998). In the above example, the roles are `goal`, `object`, `method` and `conclusion`.

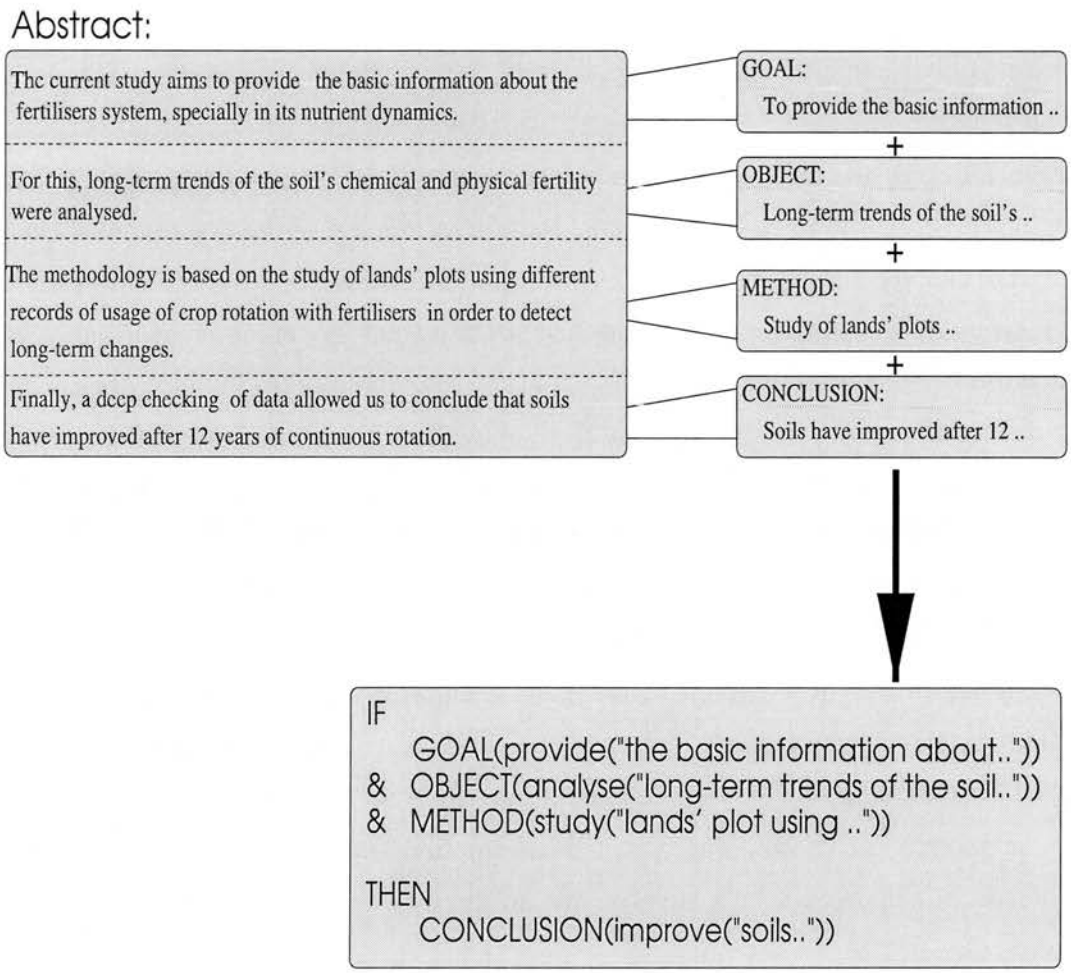

Figure 3.2: Rule Representation from the semantic and rhetorical information extracted from a document's abstract

For this work, we have defined just four roles as follows (the set of patterns which recognize these roles can be seen in appendix A):

- `goal`: denotes the goals or aims of the work/research.
- `object`: denotes the object of the research itself.
- `method`: denotes the method, procedure or analysis carried out by the research discussed in the document.
- `conclusion`: denotes the results produced, outcome obtained, or the kind

of conclusion drawn from the work.

- **Predicate Relations:** these are represented by semantic actions (predicate and arguments) which are directly connected to the role being identified and state the kind of specific relation which holds between a set of terms (words which are part of a sentence), its predicate and the role which they are linked to. Thus, for the example, they have been processed as:

```
provide('the basic information about ...')
analyse('long-term trends ...')
study('lands plot using different ...')
improve('soil ..improved after 12...')
```

- **Causal Relation(s):** Although there are no explicit causal relations in the above example, we may get high-level evidence which is usually true and states a strong connection among the relations being represented. So in general, we can hypothesize rules which have the following form:

```
IF    the current  goals are G1,G2, ..
  and the means/methods used M1,M2, ..
  and any other constraint/feature
THEN it is true that we can achieve conclusions C1,C2, ..
```

In order to extract and to represent this kind of initial knowledge from the texts, a basic IE module was built. It takes a set of input documents and produces a template-like intermediate representation for every document.

The IE task is essentially composed of two phases: Part-of-Speech (POS) tagging and role and predicate recognition. *Tagging* is carried out by using the Brill Tagger (Brill, 1997; Manning and Schutze, 1999). First, a small set of training samples (100 out of 1000 documents in Spanish) was tagged by hand. Then, the tagger was trained to learn the tagging rules for the texts. Finally, the tagger and its output configuration was applied to the whole corpus to have it automatically tagged.

Once the set of documents is properly tagged, a set of hand-crafted domain-independent extraction patterns are used to extract the key information from the tagged texts.

To this end, a set of 30 matching patterns was designed and coded (Appendix A). Each pattern constructs an output representation which involves two-level linguistic knowledge: the rhetorical role and the semantic output represented by the predicate relation, and its arguments (partial sentences).

In general, a pattern is defined as a pair containing the compounds to match and the output representation to be produced (Ciravegna and Cancedda, 1995; Moens and de Busser, 2001; Appelt and Israel, 1999). For example, in the following simple extraction pattern:

```
results show that X  : conclusion(show(X))
```

the left-hand side expression denotes the matching component, and the right-hand side (following ":") denotes the corresponding action to be instantiated (role and predicate action). In addition, variables (uppercase names) and further constraints can be specified if needed.

For the example above, there is no constraint but there is a variable `X` whose role is twofold. Firstly, this may be used to instantiate the actions with the corresponding current values, therefore, this is used as a reference to some part of the matching input. Secondly, this may provide further information about the semantics which is being taken into account along with its category. For example, if we change `X` to `EFFECT`, the semantics of the results becomes more clear. However, the practical outcome, if no constraint is provided, remains the same (i.e., either X or EFFECT is instantiated with the words that occur between "that" and the end of the sentence). The patterns are also capable of recognizing different sorts of phrases or simple constrained compounds (e.g., verbs, articles, etc) through regular expression processing.

The rhetorical role to be produced in the output action (`conclusion`, for the example above) must be specified in advance, whereas the associated semantic action can be either specified in advance (if known) or postponed to match any relation or compound coming from the texts and matching the pattern.

To clarify the use of roles within the patterns, consider the following pattern:

```
it is|was  concluded|shown that  NP1 @ACTION/v NP2:
                                  conclusion(&ACTION[NP1,NP2])
```

where some options and/or constraints are noted:

- There are options to match in the input text. In this case, these are represented by the triggered verbs: `concluded` or `shown`.
- Components specified as `@OBJECT/TAG` represent a constraint on the input for the unknown object `OBJECT` (optional) which must be tagged under a specific category `TAG`. For example, `@ACTION/v` constraints the element in the current position to be a verb ("v" is a previously defined tag used across the corpus) and named `ACTION`. If the element is known but this may take different grammar categories, one can specify it as e.g., `@fly/v`. Finally, if what matters is only the category (e.g., for prepositions), then this can be specified as `@p`.
- The two objects of the unknown semantic relation (probable noun phrases NP1 and NP2) are unconstrained. However, the relation to hold between them is restricted to be a *verb* (present or past)
- On the action side, once the elements are recognized and instantiated, these are properly filled in the action slot. In the example above, `&ACTION[NP1,NP2]` will implicitly define a predicate `ACTION` with arguments `NP1` and `NP2`. In other words, the symbol `&` forces the action (and arguments) to be converted into a more standard form: `ACTION(NP1,NP2)`.

In order to highlight how the processing is performed, suppose that we have the following fragment of a tagged input text:

```
In/p the/art  experiment/n, it/pro was/aux shown/v that/pror the/art
dose/n of/p vaccine/n AX-1/nn affected/v the/art allergical/adj
reactions/n in/p ....
```

Where `/TAG` denote the corresponding tags given to every element of the sentence: `p` for prepositions, `art` for articles, `n` for nouns, `nn` for proper names, `adj` for adjectives, `v` for verbs, `aux` for auxiliary verbs, `pror` for relative pronoun, and `pro` for pronoun.

As this input text comes in, it is matched by the previous pattern provided by the example, and a tag-free representation is produced as follows:

```
conclusion(
   affected('the dose of vaccine AX-1','the allergical reactions in..')
)
```

Once the patterns are matched and the corresponding slots are instantiated, we have the basic units of knowledge of the document (abstract) represented as an intermediate template. For instance, the following template is automatically instantiated (NB., there are no pre-existing slots to be filled and a role can occur any number of times) with the information extracted from the document in figure 3.2:

```
template('document 1',
 [
  goal(provide('basic information about the fertilisers system..')),
  object(analyse('long-term trends of the soil...')),
  method(study('lands plots...'))
  conclusion(improve('soils have improved after 12..'))
 ]
)
```

Since we claim that in the scientific/technical genre, the abstracts follow a rule-like form of evidence containing an antecedent and consequent, we can easily convert the template above into a rule representation. This is carried out by hypothesising that all the units related to the rhetorical role `conclusion` will be part of the consequent, and the rest will be part of the antecedents (conditions). Thus, the finally produced rule which graphically is shown in figure 3.2 will look like:

```
rule('Document 1',
 [   % Antecedent List
  goal(provide('basic information about the fertilisers system..')),
  object(analyse('long-term trends of the soil...')),
  method(study('land's plots...'))
```

```
 ],
 [  % Consequent List
  conclusion(improve('soils have improved after 12..'))
 ]
)
```

Once the rules are produced from the tagged documents, these become a rich source of data from which training information can be obtained for discovery purposes.

Note that unlike many IE systems previously discussed in chapter 2, the IE component of our model deals with only some IE tasks, and others are partially treated as needed. We do not have a named-entity recognition module neither a full parser. Instead, we carry out a POS tagging task, pattern matching so to identify key elements from the texts, and generate template-like information (i.e., rules). Although using a fully implemented IE system may produce improvements in the accuracy of the information extracted, we think that as far as this research is concerned, a simple IE component can yet be useful enough to produce some sensible results.

3.1.2 Training Information from the Rules and Raw Documents

As far as IE is concerned, all the information obtained can be directly extracted from the documents by the techniques of the last section, with no further processing. However, at this stage both the rules just generated and the original set of documents represent a rich source of underlying knowledge which is worth exploring beyond extraction patterns.

Specifically, the raw set of documents conveys data at the word semantics level which can effectively be computed by training analysis methods such as LSA on the corpus in order to produce the corresponding meaning representation so as to enable similarity judgements between words, sentences, or predicate actions.

On the other side, our prototypical structure for representing the abstracts is aware of rhetorical information extracted from the documents and its association with semantic information. Unfortunately, the organisation (i.e., the sequence in which the roles are stated) of this is not fixed and may vary from one abstract to another, or even among

domains. Consequently, there will be structures which are better organized than others which may give the discovery process a clue of the optimal organization or association of the rhetorical and semantic information in creating new hypotheses. This kind of key training information can be computed from the rules produced in the previous stage.

Both the training information computed from the documents and that computed from the rules will aim to feed the evolutionary discovery process in producing plausible hypotheses.

- **Getting training information at the lexical semantics level:**

  It has been suggested that huge amounts of text represent a valuable source of semantic knowledge (see chapter 2). In particular, in LSA that sort of knowledge is at the word level and considers patterns of word usage in different contexts. In our model, LSA is trained on the raw corpus to generate the vectors representing the meaning of every relevant word across the documents. Due to the high dimensionality of this semantic space, decomposition techniques (e.g., Singular Vector Decomposition) are later applied to reduce those dimensions. The final information is said to capture the semantic information which is latent in the corpus and which will focus on the essential semantic features.

  The vector representation is then converted into Prolog clauses. For example, consider two terms frequently found in an agricultural context: `soil` and `horticultural`. Their representation can be computed as the 6-dimensional vectors[1]:

```
termVector(soil,
[0.137536,-0.075237,0.121556,-0.429545,-0.376794,0.004476])

termVector(horticutural,
[0.000683,-0.000414,0.001436,-0.003579,-0.003333,-0.000154])
```

[1]The choice of the dimension is due to default working settings for the LSA program to work efficiently (Landauer et al., 1998a)

It is important to note that at this stage, no distinction is made in computing vectors for noun terms and verb terms. However, we later extend the LSA model to take into account syntactic and rhetorical information in such a way that the predicates and their arguments will only consider verb terms, and lists of terms, respectively. Furthermore, similarity judgements between predicates will use this kind of syntactic constraint and the predicate-level information extracted in the IE task to make comparisons, so preserving some grammar consistency, which is not considered in the early LSA approach.

- **Getting training information from the rules:**

  One of the problems with using the information provided by LSA is that this only takes account of word knowledge and therefore, syntactic and rhetorical information beyond this is not considered. Although the information extracted in the IE phase partially aims at overcoming this lack of structure, we claim that there is still some further analysis which can be performed in order to obtain information which guides the production and evaluation of hypotheses. To this end, we address two key questions: Which information can provide cues about the best ways to organise the units contained in the rules? And How do rhetorical and semantic information co-occur in these rules?

  The information so obtained will not provide evidence of quality in terms of novelty of the new knowledge, but will help to ensure that the new hypotheses produced in the discovery are sound in the way that they inherit some features of the good rules.

  In order to address the first issue, we obtain data concerning how the rhetorical roles are related to each other (structure) in the rules. To deal with the second issue, we compute data regarding the association between the rhetorical and the semantic information in all the elements of the rules:

  - *Computing training information about the structure of the rhetorical information:*

    By looking at typical technical abstracts, it can be seen that using rhetorical information is a key way to understand the evidence which is being stated

and to attempt to capture the essence of what is implied there. However, one can also note that the way this information is interconnected is important as the whole sequence of rhetorical information pieces represents the coherent order in which the work was carried out and the way this draws its conclusions. Taking this into consideration a question comes up: to what extent is the order of the rhetorical information in an hypothesis important?

One can think of an abstract as paragraphs which are semantically related to each other in terms of the normal measures of text coherence, in which case clearly the order matters[2]. This suggests that, in generating valid hypotheses there will be rule orderings which are more or less desirable than others. For instance, if every rule contains a "goal" as the first rhetorical role, and the GA has generated a hypothesis starting with a "method", it will be somewhat penalised and therefore, it will be very unlikely for that to survive in the discovery phase.

Since the organisation matters, a *Mutual Information* measure (Manning and Schutze, 1999; Klavans and Resnik, 1996) or any co-occurrence based metric is not sufficient as this does not consider the underlying structure even at a surface level.

Instead, we developed an order-aware approach based on the following principle: consider the p rhetorical roles of a rule as a sequence of tags: $< r_1, r_2, .. r_p >$ such that r_i precedes r_{i+1}, irrespective of the predicate actions as these are considered in a further stage. Then, the rules are used to compute the training probabilities $Prob(r_p \mid r_q)$, that is, the conditional probability that the role r_q precedes r_p.

In order to empirically verify the existence of this kind of ordering, we trained the model with the initial rules extracted from the sample corpus (1000 documents) used for further evaluations, and then these probabilities were generated. The simple Markov chain that represents the probabilities of the roles can be seen in figure 3.3, in which the arcs denote the

[2]The assessment of the hypotheses discussed in the experiments involves giving the experts an abstract-like text that represents the internal hypothesis. Thus even if the order does not change the meaning, it may affect the readibility.

(transition) probabilities and the nodes represent the corresponding roles ($< START >$ denotes the starting node). For visualization purposes, zero probability transitions are not shown.

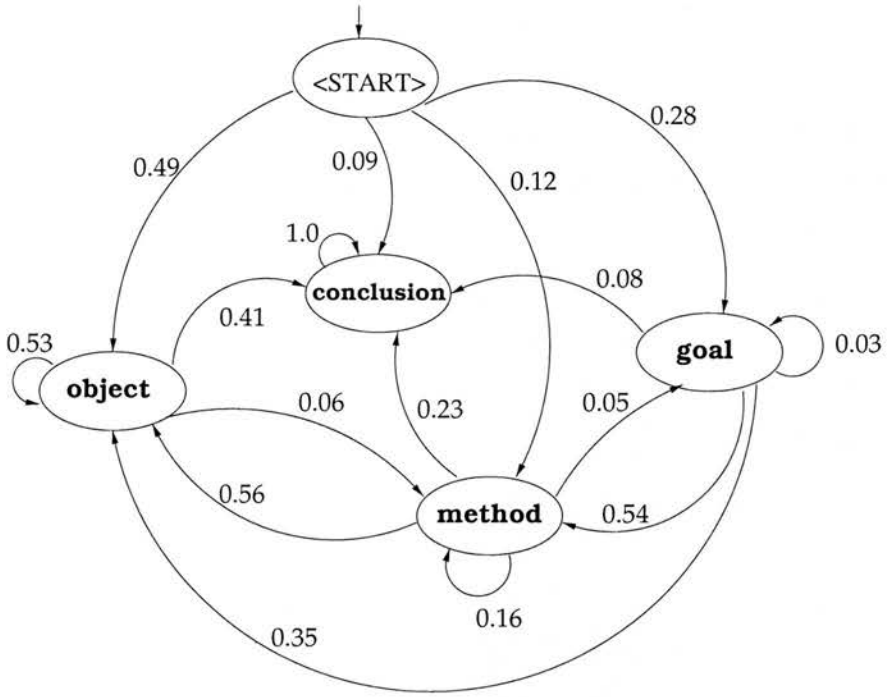

Figure 3.3: Markov Chain for Sequence of Roles

It can be observed that the information in the chain is consistent with the intuitions and captures the underlying order in a realistic form. For example:

* The description of a method is much more likely to follow a goal (0.54) than the way around (0.05).
* The `object` is the most likely piece of information to start describing the work in the text (0.49). In second place, there is the goal (0.28).
* There is no role that can follow the conclusion. Therefore, the only possible role to follow is another conclusion (1.0).
* It is very unlikely that the conclusions follow the goal (0.08).

This suggests that there are regularities in the order of the roles which a

human evaluator will prefer more than others. In addition, the information obtained and the strategy used seem to provide more information about these associations than a traditional MI-based metric.

- *Computing training information about the associations between rhetorical roles and predicate relations:*

  We suggest that a proper connection between rhetorical information and the predicate action performed constitutes key information for producing coherent units in the hypotheses. For example, the goal of some hypothesis might be associated with the *construction* of some component. In a "health" context, that connection would be less likely than having "*finding* a new medicine for .." as a goal.

  In order to get training information which takes account of this, we address the question: how likely is it for some semantic relation to be associated to some rhetorical information? For this, we adopted a Bayesian approach in which we obtain the conditional probability of some predicate `p` given some attached rhetorical role `r`, namely $Prob(p \mid r)$.

  For example, using the same previous sample corpus[3], some obtained conditional probabilities include the following:

```
Prob(analyse|goal)=0.0014        Prob(design|goal)=0.000159
Prob(study|conclusion)=0
```

  from which it can be observed that it is more probable for the goal of the research to be the analysis of some activity/process/component than the design of this. In addition, it is unlikely for the conclusion to be the study of some element/activity, etc, and so on.

- **Generating an initial population of hypotheses:**

  In order for the GA-based discovery to search and explore novel hypotheses, it is necessary to provide information which guides the learning. To this end,

[3]The examples shown come from the corpus in Spanish. Translations have been provided just to make it easier to understand what they look like.

the rules and the training information computed from the corpus are used as a plausible way to guide the search for good solutions. However, the discovery also needs an initial seed (initial guesses) to start off which is accomplished by randomly creating a set of hypotheses, the initial population.

Each hypothesis is built by combining random units from the rule set which have been extracted and have become separate databases of predicates and roles. This randomness allows for the exploration of different parts of the unseen search space to look for worthwhile hypotheses.

As an example of this combination, assume that we have the following training rules:

```
rule(1,
  [goal(describe("a")),method(burn("b"))],
  [conclusion(take("c"))])
rule(2,
  [method(anayse("x")),goal(produce("y")),
    goal(describe("z")),method(obtain("w"))],
  [])
..
```

Then, the role database will contain:

```
rrole(1, goal).
rrole(2, method).
rrole(3, conclusion).
rrole(4, method).
rrole(5, goal).
rrole(6, goal).
rrole(7, method).
..
```

and the predicate database will include:

```
relation(1,describe("a")).
relation(2,burn("b")).
relation(3,take("c")).
relation(4,anayse("x")).
relation(5,produce("y")).
relation(6,describe("z")).
relation(7,obtain("c")).
..
```

Next, the hypotheses can be built by randomly picking pairs of units from both databases to constitute its antecedent and consequent. Though, the consequent is restricted to have a `conclusion` role. For example, the hypotheses:

```
hypothesis(1,
  [method(produce("y")),goal(burn("b"))],
  [conclusion(analyse("x"))])
hypothesis(2,
  [goal(describe("a")),goal(take("c")),method(obtain("w"))],
  [conclusion(produce("y")).
```

are one possible outcome of this process. Bear in mind that at this stage we are not making any judgement on whether the hypotheses are plausible or not as this is up to the discovery strategy. However, in producing random hypotheses this way, we impose some basic requirements in terms of the following three matters:

1. *Missing data:*

   As can be seen in the sample rules, rule 2 does not contain any conclusion. This is due to the fact that either the role was not recognised by the IE task at all, or the source abstract did not contain any explicit reference to it. In either case, the rule is not complete and it may lead to unwanted hypotheses which have nothing important to say (the same holds for the rules with no antecedents). In order to avoid this, we force the initial hypotheses

to contain one conclusion drawn from any predicate relation, and at least two roles in their antecedent to make sure these have some material worth exploring (e.g., bi-grams).

2. *Encouraging frequent roles:*

   Having roles and predicates duplicated in the databases, according to the way they appear in the rules, stresses the fact that there will be roles which are more likely to appear than others, and therefore these will influence the structure of the hypotheses. In the example, `goal` and `method` have the same probabilities to be picked (3 out of 7).

3. *Keeping semantic consistency:*

   As we have no additional information on semantic types or related constraints, we are unable to combine arguments from different predicates as the semantic information they convey is different. Hence there is certain information which can not be produced. For example, from predicate relations `produce("y")` and `describe("a")`, it is not possible to have a predicate `describe("y")` in an hypothesis as the semantic type of `"y"` may not match the one which the argument of `describe` is supposed to have. The exception is when dealing with the same relation where the arguments are changeable because we assume they should have the same semantics in the current context. For the above example, this holds for `describe("a")` and `describe("z")` as long as the number of arguments is the same.

3.2 Hypothesis Discovery

Because of the nature of the text data and the contents of the hypotheses being produced, our approach to evolutionary KDT is strongly guided by semantic and rhetorical information. Consequently, operations for searching and exploring new solutions have been designed in such a way that these take into account semantic constraints before producing the offspring so to keep them coherent. By no means does this aim at immediately producing interesting hypotheses from a KDD viewpoint as it is up to the evaluation strategy to do that by selection.

Accordingly, a key issue which influences the search, production and evaluation of hypotheses is the semantic relatedness between the units of these hypotheses and the underlying semantic and rhetorical information contained in the extracted rules. For search and production purposes, semantic relatedness determines whether some combinations and connections are plausible. For the purpose of the evaluation, the semantic measure establishes whether these plausible hypotheses are good or not according to high-level quality criteria. In order to deal with the semantic similarity, we have first to determine how a unit is semantically represented, how this is put together with the representation of another, and finally, how this is used to make further judgements.

We have described in section 3.1.1 our method to represent the basic units of knowledge, whether they are predicates or arguments, which in a LSA sense are treated as terms. However, we need a more structured way to deal with this information in order for each hypothesis to be represented as a whole. First, we propose a simple strategy for representing the meaning of the predicates with arguments. Then, a simple method is developed to measure the similarity between these units.

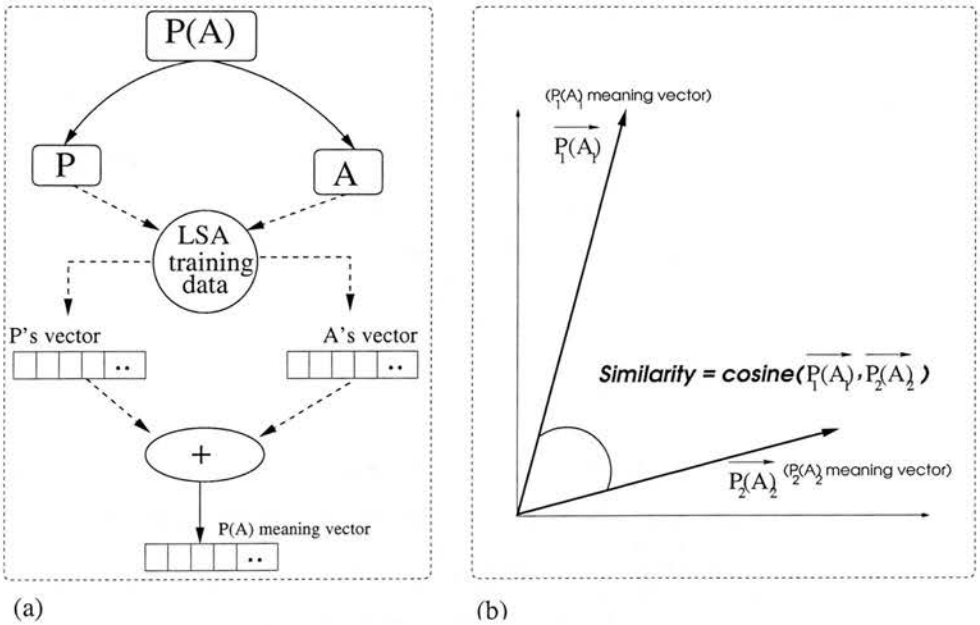

Figure 3.4: LSA levels of Processing:(a) Computing the vector representation for predicate and arguments. (b) Computing semantic similarity between the predicates' vectors

Given a predicate P and its argument A as shown in figure 3.4(a), the vectors representing the meaning for both of them can be directly extracted from the training information provided by the LSA analysis. Representing the argument involves averaging all the vectors representing the terms of the argument, as is usually performed in semi-structured LSA (Wiemer-Hastings and Zipitria, 2001; Klebanov, 2001; Bellegarda, 2000) which has proved to be effective in a wide range of problems. Once this is done, the meaning vector of the predicate and the argument is obtained by computing the sum of the two vectors as used in (Wiemer-Hastings, 2000). If there is more than one argument, then the final vector of the argument is just the sum of the individual arguments' vectors.

Next, in making further semantic similarity judgements between two predicates $P_1(A_1)$ and $P_2(A_2)$ as seen in figure 3.4(b), we take their corresponding previously calculated meaning vectors and then the similarity is determined by how close these two vectors are. We can evaluate this by computing the *cosine*[4] between these vectors which gives us a closeness measure between -1 (complete unrelatedness) and 1 (complete relatedness).

With this in mind, the procedure to compute the LSA representation for a whole hypothesis H is straightforward: the compound vector (resulting vector representing the predicate and arguments together) of all the predicates of H is obtained by using the following procedure:

PROCEDURE Compute CompoundVector
IN: *List of predicates with arguments (P(A)) of hypothesis H*
OUT: $\vec{V}_h$ (Resulting vector)

$\vec{V}_h \leftarrow \vec{0}$
For all $P_i(A_i) \in$ H, $P_i, A_i \in$ terms and sequence of terms, respectively :
$\vec{P}_i \leftarrow$ LSA vector for P_i
$\vec{A}_i \leftarrow$ LSA vector for A_i
$\vec{V}_h \leftarrow \vec{V}_h + (\vec{P}_i + \vec{A}_i)$

[4]The standard cosine between two vectors is used as a plausible and efficient method to determine the angle between vectors and therefore, the closeness between their directions in a multi-dimensional space.

RETURN $\vec{V}_h$

Note that rhetorical roles are not taken into account in getting the meaning vectors or the similarity measures because the LSA training information may not contain the terms representing the roles. The roles are considered in other ways as the hypotheses are evaluated. Then, the semantic similarity between two hypotheses H_a and H_b, can be effectively calculated with the simple procedure:

PROCEDURE SemanticSimilarity
IN: *List of predicates with arguments of* H_a *and* H_b *(Pa,Pb)*
OUT: *SemanticSimilarity* (between H_a and H_b)

$$\vec{V}_a \leftarrow ComputeCompoundVector(Pa)$$
$$\vec{V}_b \leftarrow ComputeCompoundVector(Pb)$$
$$SemanticSimilarity \leftarrow \frac{\vec{V}_a.\vec{V}_b}{\|\vec{V}_a\|*\|\vec{V}_b\|} \qquad \text{/* Cosine */}$$

RETURN SemanticSimilarity

Note however, that if instead of predicates with arguments, we provide just terms (whether they are arguments or predicates), the procedure simply performs the calculation by extracting the LSA vectors for these terms, and then the cosine is calculated as above.

The designed search operations which explore new solutions will be semantically-driven either by restricting the content of the offspring to be produced or using the hypotheses' underlying semantics or relatedness in changing some of their elements.

Accordingly, our proposal for GA-based discovery can be seen as a two-stage process which involves:

1. Searching for and exploring new solutions by making certain changes in the hypotheses and evolving to a new population through a number of learning steps (i.e., generations).

2. And evaluating, in terms of several criteria, the offspring being produced in order to establish which solutions fit best, and therefore, which ones should remain as candidates from one learning step to another.

Next, we discuss the details of the operations for the discovery itself. The evaluation stage is left to section 3.3.

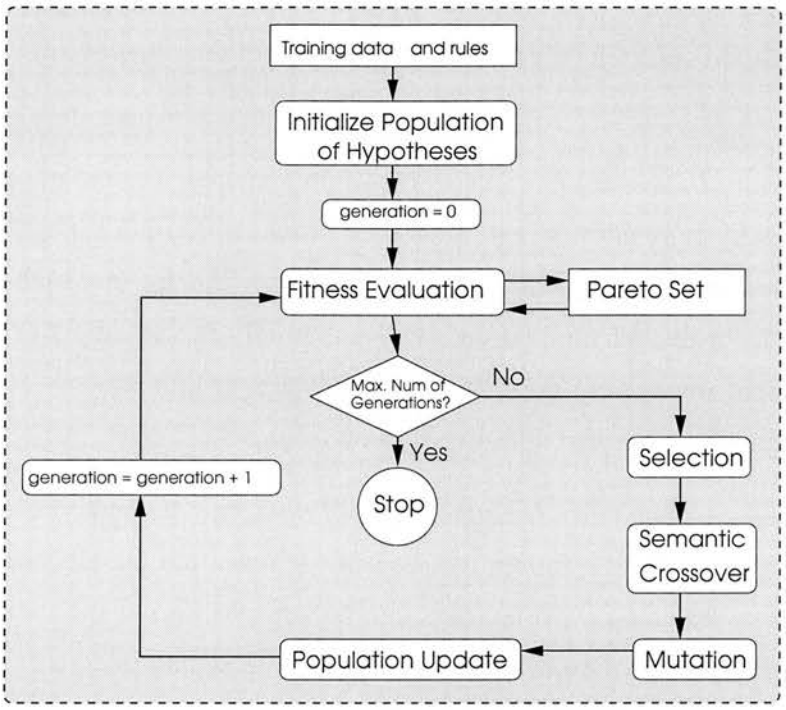

Figure 3.5: The Structure of the Semantically-Constrained and Multi-Criteria GA

The general working for the GA-based discovery is highlighted in the algorithm in figure 3.5. The GA starts off from an initial population of hypotheses which are assessed and assigned a "goodness" value that determines how well they do according to certain criteria. The role of the Pareto set in this case will be to keep the current not-worst solutions which have been traded-off so far. Then, constrained GA operations are applied to create a hopefully better population in the next generation and a new learning step is performed. The complete process stops when a fixed number of generations (e.g., 1000) is achieved.

We have designed semantics-aware GA operations which both inherit basic properties of binary GA representations (Goldberg, 1989; Deb, 2001) and consider structural knowledge and constraints.

In particular, we have developed two basic genetic operators: crossover and mu-

tation, and a steady-state based population management strategy. For distinction purposes, the individuals/hypotheses (chromosomes, in GA terminology) which are handled by the operators are referred to as *parents*, and the outcome produced by the operators are called *children*, which will become part of the *offspring* in every generation.

- *Population Management:* picks a small number (generational gap) of best parents in every generation according to their fitnesses. Furthermore, the population is updated by using a steady-state replacement strategy in which the children created from the selected parents will potentially replace the worst parent hypotheses of the current population. The details of this strategy for population replacement can be seen in figure 3.14 of section 3.3.2.

- *Crossover:* takes a pair of selected hypotheses (parents) and performs a recombination with a fixed probability p_c, in which the individuals swap their components to produce new offspring in a random position of the individuals. Unlike traditional GA-based crossover, we use the semantic constraints to define two kind of recombinations: Swanson-based crossover (figure 3.6) and default crossover (figure 3.7).

  1. *Swanson's Crossover:* based on Swanson's hypothesis (Swanson, 1988; Finn, 1998; Swanson, 2001) we propose a recombination operator which allows two selected hypotheses AB and BC to be swapped (see figure 3.6) as long as semantic constraints are met.

     Swanson's early proposal for novelty of extracted patterns from the titles of the documents looked like:

     If a document (AB) contains two concepts (and relations) such that "A implies B", and another document (BC) contains concepts such that "B implies C", then a new interesting connection "A implies C" can be inferred and is worth exploring (as long as it is not already in either document).

     Although this has proven to be plausible in certain simple cases, we need to make some adaptations as the original approach only considers single terms in the title of the documents. In addition, since our model does not have

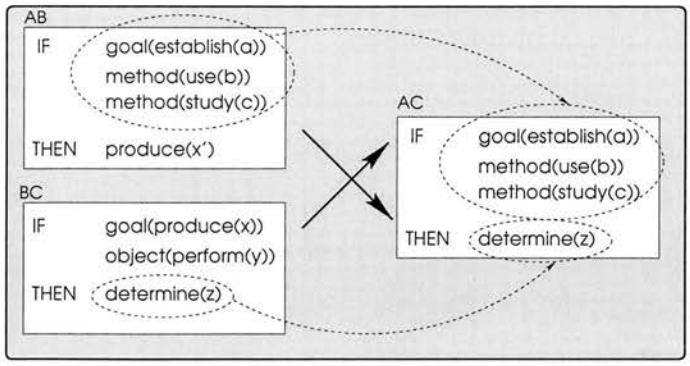

Figure 3.6: Swanson's Crossover

domain-specific information, it is not possible to know which relationships should be considered, and consequently, the inferences to be enabled cannot be defined in advance. Considering this scenario, we develop a more open method to draw general inference from the relationships extracted from the texts.

We have then approached this kind of crossover by using a transitivity-like hypothesis in a more flexible way:

If there is a hypothesis (AB) "IF A THEN B" and another one (B'C) "IF B' THEN C", (B' being something highly similar to B) then a new interesting hypothesis "IF A THEN C" which is worth exploring can be inferred, only if the conclusions (B) of AB have high semantic similarity (e.g., via LSA) with the conditions (B') of hypothesis B'C.

The similarity is computed from the predicate relations and their arguments by using LSA, irrespective of the rhetorical roles as the ones of "B" and "B'" are disjoint (i.e., one contains conditions whereas the other contains conclusions). For this, the similarity is calculated by calculating *Semantic-Similarity(B',B)* (similarity applied to sets of predicates and arguments in this case) as described in section 3.2.

In practical terms, this means that if the relevant parts of two hypotheses are highly similar according to a fixed threshold in the LSA-based measure, they will always swap their material at the same point (i.e., the point

between antecedent and consequent).

2. *Default Semantic Crossover:* if the previous transitivity does not apply then a default recombination is performed only if both hypotheses as a whole have high semantic similarity which is defined in advance by providing minimum thresholds (otherwise, no crossover is performed). The recombination is performed by swapping the conditions of both hypotheses in a random location as shown in figure 3.7.

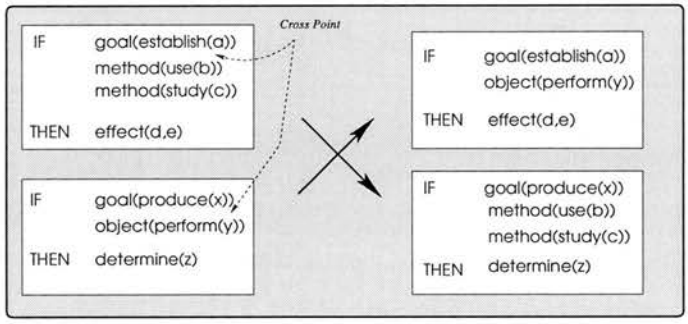

Figure 3.7: Default Semantic Crossover

Note that irrespective of the kind of crossover, no roles are allowed to be swapped if the hypotheses do not meet the semantic constraints. Even though the offspring may not be so interesting, the constraint ensures that the hypotheses will at least have some minimum semantic coherence.

- *Mutation:* aims to make small random changes to hypotheses to explore new possibilities in the search space. The probability for an individual to be mutated is given by a fixed probability p_m. As with crossover, we have dealt with this operation in a constrained way, so we propose three kinds of mutations which are randomly chosen to deal with the hypotheses' different objects:

  1. *Role Mutation:* one rhetorical role (including its contents: predicate and arguments) is selected randomly and replaced by a random but a legal role and predicate-argument from the database as shown in figure 3.8.

     As an effect, it is hoped that in creating new hypotheses, this mutation modifies the structure of rhetorical roles in a good way.

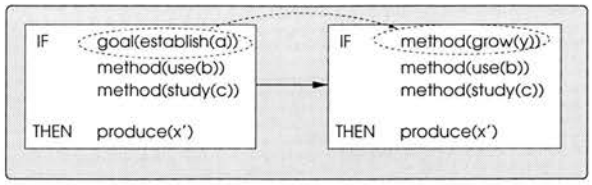

Figure 3.8: Role Mutation

2. *Predicate Mutation:* one inner predicate with its arguments is selected and replaced by another one from the database (see figure 3.9).

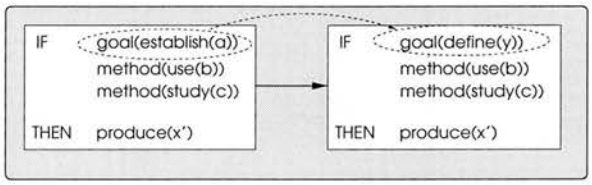

Figure 3.9: Predicate Mutation

This kind of operator should have a direct effect on the semantic cohesion between the current role and the new predicate relation as described later in the evaluation process.

3. *Argument Mutation:* since we have no information about semantic types of predicate arguments, we choose a new argument to a predicate by following a guided procedure:

   Let S_p be the set of possible predicate relations (predicates with arguments) in the database produced by pre-processing, P_{curr} and A_{curr} the current predicate and argument, respectively, and $SemanticSimilarity(P_{curr}(A_i), P_{curr}(A_j))$, the semantic similarity between $P_{curr}(A_i)$ and $P_{curr}(A_j)$, then

   (a) Select candidate predicate relations:

   $$CandPreds = \{P_i(A_i) \in S_p \mid P_i = P_{curr}\}$$

   (b) Select the list of candidate arguments:

$$CandArgs = \{A_i \mid SemanticSimilarity(P_{curr}(A_{curr}), P_{curr}(A_i)) > threshold \\ \wedge P_i(A_i) \in CandPreds\}$$

(c) Select random argument:

$$BestArg = random(CandArgs)$$

Figure 3.10 shows an example of this operation in which the argument `a` has been changed to `a'`. The latter is supposed to have a high similarity with the former in order to preserve a close semantic meaning.

Note that the operation does not alter the structure of the roles. However, the overall meaning representation of the hypothesis will be altered as the new argument of the predicate `establish` will have a new vector representation.

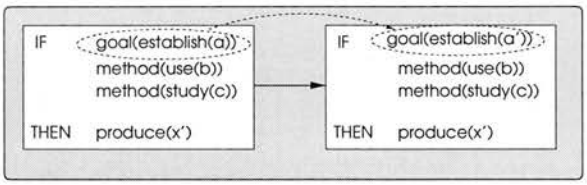

Figure 3.10: Argument Mutation

As can be seen in the algorithm highlighted in figure 3.5, once the operations are applied and new individuals are created, the population needs to be updated to go on with the next learning generation (see update algorithm in figure 3.14). To this end, we use a non-generational GA in which some of the worst parents are replaced by the new offspring in order to preserve the hypotheses' good material from one generation to other, and so to encourage the improvement of the population's quality.

3.3 Automatic Evaluation

The role of the genetic operators is only to produce new hypotheses which may become part of the new generation. However, the goodness of these individuals must be

evaluated in order to establish whether they provide good solutions or not. As a consequence, the fittest individuals will survive to the next learning generation and others will be eliminated. To this end, two aspects need to be tackled: developing appropriate criteria in order to assess the hypotheses' quality, and designing an optimisation strategy which trades off between these criteria to incrementally create a pool of good solutions.

3.3.1 Model Metrics

In developing evaluation criteria (metrics), we have taken into account different issues such as scope (Is the information of the hypothesis itself sufficient to have it evaluated?, etc), plausibility (Is the hypothesis semantically sound?, Is the current hypothesis coherent?, etc), and quality itself (How is the hypothesis supported from the real documents?, How interesting is it?, etc). Accordingly, we propose a set of eight evaluation metrics to assess the hypotheses: **Relevance, Structure, Cohesion, Coherence, Coverage, Simplicity, Interestingness**, and **Plausibility of Origin**. It is worth noting that all the criteria must be maximised within a scale between 0 and 1, hence higher values of each are searched for in every hypothesis.

In order to have a general view of which aspects of the hypotheses are being evaluated by the criteria, one can look at the general schema of a hypothesis H in figure 3.11. In this, R_i denotes the $i-th$ rhetorical role, and P_i and A_i represent the $i-th$ predicate and arguments, respectively.

For example, the metric *structure* will involve an evaluation of consecutive roles, coverage denotes an evaluation involving the whole hypothesis and the set of extracted rules, and so on.

Next, for describing the underlying working of these criteria, suppose we have been given the following intermediate hypotheses H_1 and H_2 produced by the GA at some point:

```
H1:
IF goal(determine('the optimum doses of Furadan that ..'))
 method(establish('robust grevillea either by its own or associated..'))
```

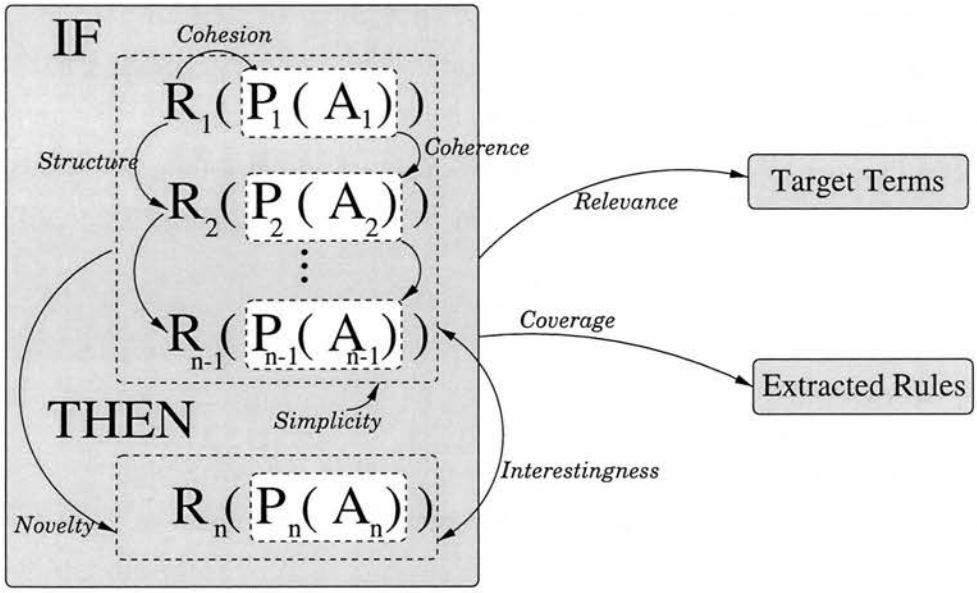

Figure 3.11: Scope of Evaluation

```
THEN
 conclusion(produce('aspects of the biology and behavior of the ..'))

H2:
IF
 method(use('white quebracho wood...'))
 goal(carry_out('fitosociological study of the vegetation...'))
THEN
 conclusion(produce('high coefficients of correlation..'))
```

Furthermore, suppose that we are provided with some data obtained from the training information, such as bi-gram probabilities (see figure 3.3):

```
/* Prob(<current role>|<previous role>) */
/* <start> indicates that there is no previous role */

  Prob(goal|<start>)   = 0.28
  Prob(goal|method)    = 0.05
```

```
Prob(method|<start>)= 0.12
Prob(method|goal)   = 0.54
..
```

And the conditional probabilities:

```
/* Prob(<relation>|<role>) */

Prob(determine|goal)   = 0.005
Prob(establish|method) = 0
Prob(use|method)       = 0.084
Prob(carry_out|goal)   = 0.002
..
```

The criteria used by the model will be described in terms of the questions they try to address and the methods developed to deal with them.

1. **Relevance**

   Relevance addresses the issue of how important the hypothesis is to the target concepts. This involves two concepts (i.e., terms), as previously described, related to the question:

   What is the best set of hypotheses that explain the relation between $< term1 >$ *and* $< term2 >$*?*

   Considering the current hypothesis, it turns into a specific question: how good is the hypothesis in explaining this relation?

   This can be estimated by determining the semantic closeness between the hypothesis' predicates (and arguments) and the target concepts[5] by using the meaning vectors obtained from the LSA analysis for both terms and predicates.

   Although this may look straightforward to do, there are three aspects worth discussing:

[5]Target concepts are relevant nouns in our experiment. However, in a general case, these might be either nouns or verbs.

(a) As we use a domain corpus, an important number of terms tend to be highly similar (via LSA) to each other because of similar contexts. In this case, a simple measure of semantic relatedness between the predicate (and arguments) and target concepts can be misleading as this would eventually reveal the fact that most of the pairs predicate-terms have high similarity when in reality, they do not. Hence we need to "filter" the pairs that are really relevant from those that are not.

(b) As we are looking for predicates (and arguments) relevant to both terms, simple aggregation measures would not highlight the fact that the relevance is considered for both of them. For example, assume that the similarity between a predicate (with arguments) $P_1(A_1)$ and each one of the target concepts is 75%. Next, the similarity between a second predicate (with arguments) $P_2(A_2)$ and the target concepts is 100% and 50%, respectively. If we calculate, for instance, the average closeness for both predicates, this gives us the same value, that is, 75%, which might be good. However, note that one of the concepts for $P_2(A_2)$ may not be that relevant as this shows a relatively low similarity (50%). For this, the computation of the relevance should take into account the details of the similarities.

Our method for assessing relevance takes these issues into account along with some ideas of Kintsch's Predication. Specifically, we use the concept of *Strength* (equation 2.1: strength(A,I) = f(sim(A,I),sim(P,I))) between a predicate with arguments and surrounding concepts (target terms in our case) as a part of the relevance measure, which basically decides whether the predicate (and argument) is relevant to the target concepts in terms of the similarity between both predicate and argument, and the concepts.

We define the function f as proposed by (Kintsch, 2001) to give a relatedness measure such that high values are obtained only if both the similarity term-argument (α) and term-predicate (β) exceed some threshold. Next, we highlight the closeness by determining the square difference between each similarity value and the desired value (1.0). If we take the average square difference, we obtain an error metric which is a *Mean Square Error* (MSE). As we want to get low

error values so to encourage high closeness, we subtract MSE from 1. Formally, $f(\alpha,\beta)$ is therefore computed as the function:

$$f(\alpha,\beta) = \begin{cases} 1 - MSE(\{\alpha,\beta\}) & \text{if both } \alpha \text{ and } \beta > threshold \\ 0 & \text{Otherwise} \end{cases}$$

where the MSE is the *Mean Square Error* between the similarities and the desired value ($Vd = 1.0$), is calculated as:

$$\text{MSE}(\{\text{list of n values } v_i\}) = \frac{1}{n}\sum_{i=1}^{n}(v_i - Vd)^2$$

In order to account for both target terms, we just take the average of **strength** for both terms. So, the overall relevance becomes:

$$relevance(H) = \frac{\frac{1}{2}\sum_{i=1}^{|H|} strength(P_i,A_i,<term1>) + strength(P_i,A_i,<term2>)}{|H|}$$

in which $|\ H\ |$ denotes the length of the hypothesis H, that is, the number of predicates.

Assume that the strength of each predicate with arguments is computed for the target terms `(glycocide, inhibitor)`, and the length of the hypotheses is 3, then the final relevance values are calculated as follows[6]:

```
Relevance(H1)=
 (1/2)*
 (strength(determine,'the optimum doses ..', <target>)
 +strength(establish,'robust grevillea ..'), <target>)
 +strength(produce,'aspects of the biology ..', <target>)
 ) /3  = 0.218

Relevance(H2)=
 (1/2)*
```

[6]For visualization reasons, the application of *strength* is shown for $< target >$, meaning that the sum of strengths for the two targets is shown.

```
  (strength(use,'white quebracho wood ..',<target>)
  +strength(establish,'fitosociological study ..', <target>)
  +strength(produce,'high coefficients..', <target>)
 )/3  = 0.185
```

Although hypothesis H1 is more relevant than hypothesis H2, this is not necessarily a determining factor as H2 may have rhetorical information which makes it more plausible, therefore, other criteria must be considered.

2. **Structure**

   This metric addresses the question of how good the structure of the rhetorical roles is, which is approached by determining how much of the extracted rules' structure is exhibited in the current hypothesis.

   Since we have previous pre-processed information for bi-grams of roles, the structure can be computed by following a Markov chain (Manning and Schutze, 1999; Klavans and Resnik, 1996) as follows:

$$Structure(H) = Prob(r_1) * \prod_{i=2}^{|H|} Prob(r_i \mid r_{i-1})$$

   where r_i represents the $i-th$ role of the hypothesis H, $Prob(r_i \mid r_{i-1})$ denotes the conditional probability that role r_{i-1} immediately precedes r_i. $Prob(r_i)$ denotes the probability that no role precedes r_i, that is, it is at the beginning of the structure (i.e., $Prob(r_i \mid < start >)$).

   One hypothesis could be more relevant than another but it may not have an appropriate structure to represent its knowledge. So, the structure measure for our sample hypotheses is:

```
structure(H1)= Prob(goal|<start>)*
               Prob(method|goal)*Prob(conclusion|method)=0.0336
structure(H2)= Prob(method|<start>)*
               Prob(goal|method)*Prob(conclusion|goal)=0.0065
```

Despite the fact that the role structure for H1 is better than for H2, the rest of the criteria and the trade off with other hypotheses need to be taken into account. For this, we need to go beyond this structural information and to come up with some criteria which can tell us more about the rest of the semantic information that the hypothesis conveys, for example, examining the association between the rhetorical information and the predicates.

3. **Cohesion**

Cohesion addresses the question of how likely a predicate action is to be associated with some specific rhetorical role. In other words, it should measure the degree of association between rhetorical information and predicate actions. The underlying issue here is that there will be some predicate relations P_i which are more likely than others to be associated with some role r_i. Consequently, the best ones should be "rewarded" in the optimisation phase.

Using the conditional probabilities provided by the training data, *cohesion* for hypothesis H can be expressed as:

$$\text{cohesion(H)} = \sum_{r_i, P_i \in H} \frac{Prob(P_i|r_i)}{|H|}$$

where $Prob(P_i \mid r_i)$ states the conditional probability of the predicate P_i given the rhetorical role r_i, and $\mid H \mid$ is the length of hypothesis H (i.e., number of predicate actions).

For instance, the cohesion for the sample hypotheses is calculated as follows:

```
cohesion(H1)=
 (Prob(determine|goal)+ Prob(establish|method)+
      Prob(produce|conclusion))/3
 = 0.045

cohesion(H2)=
 (Prob(use|method)+ Prob(carry_out|goal)+
      Prob(produce|conclusion))/3
 = 0.050
```

The degree of cohesion for H2 is slightly higher than for H1. However, the criteria so far are conflicting, so it is necessary to distinguish them based on deeper factors, as cohesion only involves a superficial association between the rhetorical information and the predicate actions. Another way to explore the deep semantic knowledge stated in the hypothesis is to look at the organisation of the hypothesis in terms of the coherence between its elements, which is described next.

4. **Coherence**

Coherence deals with the question whether the elements of the hypothesis relate to each other in a semantically coherent way.

Unlike rules produced by evolutionary methods or other machine learning techniques in which the order of the conditions is not an issue, the hypotheses produced in our model rely on pairs of adjacent elements which should be semantically sound, a property which has long been dealt with in the linguistic domain, in the context of *text coherence* (Dijk and Kintsch, 1983) for instance using LSA (Landauer et al., 1998b; Foltz et al., 1998).

As we have semantic information provided by the LSA analysis, we developed a simple metric following the work by (Foltz et al., 1998) on measuring text coherence.

Specifically, the coherence metric is calculated by considering the average semantic similarity between consecutive elements of the hypothesis. However, note that this closeness is only computed on the semantic information that the predicates and their arguments convey (i.e., not the roles) as the role structure has been considered in a previous criterion.

Keeping in mind the organisation of predicates seen in figure 3.11, the criterion can be expressed as follows:

$$\text{Coherence(H)} = \sum_{i=1}^{(|H|-1)} \frac{SemanticSimilarity(P_i(A_i), P_{i+1}(A_{i+1}))}{(|H|-1)}$$

where $|H|$ is the length of the hypothesis, and $(|H|-1)$ denotes the number of adjacent pairs.

For the example, the coherence is given by:

```
Coherence(H1)=
 (SemanticSimilarity(
   determine('optimum doses..'),establish('robust gevillea..'))
   +
 SemanticSimilarity(
   establish('robust gevillea..'),produce('aspects of the biology..'))
 )/2
  = 0.43
Coherence(H2)=
 (SemanticSimilarity(
   use('white quebracho wood ..'),
   carry_out('fitosociological study ..'))
   +
 SemanticSimilarity(
   carry_out('fitosociological study ..'),
   produce('high coefficients..'))
 )/2
  = 0.48
```

Although H2 is better than H1 in terms of the semantic information this conveys (e.g., coherence, cohesion), it is unclear whether the new knowledge stated by H2 is well supported based on the knowledge exhibited by the original rules.

5. **Coverage**

The coverage metric tries to address the question of how much the hypothesis is supported by the rules.

Some researchers in KDD have also measured the coverage of the hypothesis along with other statistical-based metrics as an indication of the quality of the hypothesis. Because of the discrete nature of the attributes of the hypotheses discovered, some assume that the coverage and the (predictive) accuracy of the

hypotheses can be calculated from the number of examples satisfying the antecedent and/or the consequent (Freitas, 2001a; Lee, 2000; Radcliffe and Surry, 1994). However, the examples (positive and negative) are not always available. Besides, in the context of TM/KDT, there are no structured attributes permitting this kind of computation.

A good approximation to this, in the context of BOW-based text mining has been carried out by (Basu et al., 2001a). They measure the coverage of a discovered rule by computing the portion of items in a sample set covered by this rule. Here, one attribute of a rule covers an attribute of a sample only if the former is a more general case than the latter. As the method is just using keywords and additional semantic information is provided via WordNet, it is plausible for them to see whether one concept is more general than another (i.e., via hypernyms). Note however that the method relies on the provision of an external general-purpose resource which unfortunately does not contain all the information of interest.

Recent approaches to measuring coverage in DM/KDD, such as Lattice-based methods (Kourie and Oosthuizen, 1998; Valtchev and Missaoui, 2001; Mugnier, 2000; Godin et al., 1995) can partially deal with the above issues. From a search space composed of hypotheses and a set of graph-oriented operations, the methods can build a generalization space which contains new instances of hypotheses and explicit information which makes it possible to uncover some interesting associations, for example, hypotheses that share features (objects) with others, etc. By using this kind of shared knowledge, it is possible to measure the cover of the nodes (i.e., groups of objects that are contained either in the original hypotheses or in the new ones) in terms of downward closure operations (Sahami, 1995).

Despite the benefit in building a lattice-based model independent of external resources (or examples), efficiency reasons make it almost impossible to apply any lattice construction algorithm that has a fair performance (at least, within hours). While the lattices' ability to "discover" knowledge is highly relevant in the context of DM/KDD, this may not be the critical factor in our model as the discovery itself is accomplished by the GA. Accordingly, only some ideas about measuring "cover" in terms of sharing features between hypotheses can be

adapted. Indeed, none of these methods to measure coverage have been used in the context of TM/KDT yet.

Considering the different issues arising from the plausible methods to compute some sort of coverage, we approached this criterion by assuming that one hypothesis is supported by (covers) a rule from the training set if the elements of the hypothesis are exactly or roughly contained in the rule.

Based on this, it is sensible to think of the hypothesis' antecedent as a set of elements, no matter in which order they occur because this is being assessed by other criteria, and therefore, part of the final decision of establishing whether the hypothesis is good or not will rely on the trade-off between the instances that occur and the way they occur (i.e., `structure`). For an effective design of a cover relationship, two basic questions have to be addressed: which elements should be considered, and given that the matches between instances are not necessarily exact as this involves the underlying semantics, how is the relation dealt with?

For the first issue, we claim that predicate (with arguments)-level processing is sufficient. This is supported by the fact that rhetorical roles are taken into account in other ways already, and unlike predicates, the roles do not have any semantic representation, which would make it difficult to perform any further comparison.

For the second issue, we propose an approximate computation of the criterion in which (the antecedent part of) a hypothesis H will cover the antecedent part of a rule RU_i only if the objects of H (predicates and arguments) are roughly (or exactly, in the best case) contained in RU_i. This means that if the membership of some elements of the hypothesis in the rule is not exact, they still can be accommodated as long as there is a certain similarity between the two.

On the other hand, since we have no additional semantic information to establish whether a predicate covers (or may be more general than) another one, we have restricted the similarity comparisons to be carried out with predicates of the same name and number of arguments. Thus, predicates such as `analyse("x")` and `analyse("y")` are comparable, whereas `generate(..)` and `recognise(..)`

are not (in other words, we cannot say anything about them). Once the predicate is eligible, the membership is said to be true either if it is exactly contained in the rule or if its similarity with one in the rule exceeds some threshold. Formally, computing the set of rules covered by some hypothesis H is defined as:

$$\begin{aligned}\texttt{RulesCovered(H)}=\{\ & RU_i \in RuleSet \mid \forall HP_k \in HP \quad \exists P_j \in RU_i : \\ & (\texttt{SemanticSimilarity}(HP_k, P_j) \geq \texttt{threshold} \\ & \wedge\ \texttt{predicateName}(HP_k)=\texttt{predicateName}(P_j))\}\end{aligned}$$

Where $SemanticSimilarity(HP_k, P_j)$ represents the semantic similarity between predicates with arguments HP_k and P_j, `threshold` defines a minimum fixed value, $RuleSet$ denotes the whole set of rules, HP represents the list of predicates (with arguments) of the antecedent of H, and P_j is a predicate (with arguments) contained in a rule's antecedent RU_i. Once the set is computed, the criterion can be obtained as follows:

$$Coverage(H) = \frac{|RulesCovered(H)|}{|RuleSet|}$$

Where $|\ RulesCovered\ |$ and $|\ RuleSet\ |$ denote the size of the set of rules covered by H, and the size of the whole rule set, respectively.

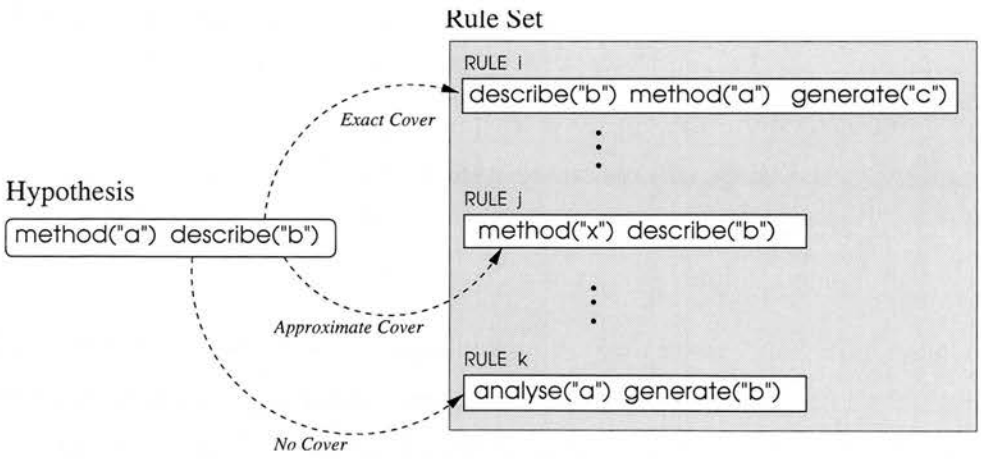

Figure 3.12: Coverage: A Worked Example

In order to see the practical outcome of the algorithm, figure 3.12 illustrates the coverage property (the example does not consider the early sample hypotheses but different hypotheses and sample rules to highlight the used features). Only the predicates (and arguments) of the antecedent part of both the hypotheses and the rules are considered in the figure. It shows three different cases:

- Hypothesis H is exactly contained in rule `i`: all the elements of the hypothesis are members of rule `i` (H covers rule `i`).
- Hypothesis H is approximately contained in rule `j`. `method("x")` is covered by H as long as `"a"` is strongly similar to `"x"`, and `describe("b")` is exactly contained in rule `j` (H covers rule `j`).
- Rule `k` is not covered by H: not all elements of H are contained in rule `k`.

Therefore, the current hypothesis covers rules $\{i, j\}$. For the sample hypotheses, what can be said if they have the same coverage? Not much as both hypotheses cover the same number of rules. Note that for the purpose of the example, both hypotheses have the same length and same coverage value so as to stress the fact that even in this case, the predicates contained in each can be different.

While this kind of computation might be a good approximation for the criterion *Coverage*, there is an important issue concerning efficiency which is worth noting. Both the number of hypotheses (N_h) and the number of rules (N_r) are relatively large numbers, and in large-scale applications, they can become even bigger than those used in our model. Computing the set of rules covered by a hypothesis means that each hypotheses must check N_r rules. As this must be performed for the whole population, $N_h * N_r$ checkings should be carried out in every generation.

However, note that because of the constraint on the predicates to be compared, not all the rules need to be checked. The system only needs to know in which rules the exact and the similar predicates are. For this, in the training step, each produced relation is linked to two lists of rule references. The first list contains the rules which contains exactly the current relation (predicate and arguments).

The second list contains the rules which contains same predicate but the argument's contents are different. Finally, when computing the set of covered rules, the criterion only needs to look at these lists.

Given that the criterion is not determining, additional quality criteria from a KDD viewpoint need also to be also taken into account.

6. **Simplicity**

   This deals with the question of how simple the hypothesis is (i.e., shorter hypotheses should be preferred). For this, the focus has concentrated on the length of the hypothesis, that is, the number of predicates contained in it.

   Since all the criteria have to be maximised, and shorter and/or easy-to-interpret hypotheses should be preferred, the evaluation is simply given by:

$$\text{Simplicity(H)= 1 - } \left(\frac{|H|}{<MaxElems>}\right)$$

   where $< MaxElems >$ denotes a fixed maximum number of elements allowed for any hypothesis. For our sample hypotheses, let $< MaxElems >= 5$, then

```
simplicity(H1)=1-3/5= 0.4
simplicity(H2)=1-3/5= 0.4
```

   Since both hypotheses have the same value no further decision can be made until they are compared to other hypotheses.

7. **Interestingness**

   The aim of objectively measuring a traditional subjective criterion such as interestingness is to estimate the degree of surprisingness and/or unexpectedness in what the current hypothesis conveys. Specifically, this assesses the hypothesis in terms of how interesting this is according to the unexpectedness of the relation between its antecedent and consequent.

   Traditionally in DM, interestingness is perceived as a measure of "surprisingness" of a rule's individual attributes. As these attributes are, in general, discrete, the surprisingness is therefore taken from a information-theory viewpoint

in which *Information Gain* (Weaver and Shannon, 1963; Schneider, 2000) is measured for each attribute, and the criterion is finally computed as inversely proportional to that gain. That is, a user would tend to be more surprised if a rule containing attributes with low information gain is observed (Freitas, 1998; Jaroszeqicz and Simovici, 2001). In the context of association rules, approaches such as (Liu et al., 2000) see interestingness in a slightly similar way, that is, as a degree of unexpectedness: rules are interesting if they are unknown to the user or contradict the user's existing previous knowledge or expectations. In terms of producing interesting prediction rules, (Radcliffe and Surry, 1994) propose a slightly different approach in which interestingness is seen as an effect of capturing some trend in the data from a statistical point of view. A common working assumption here is that previous knowledge or user-defined templates are available to indicate which combination of attributes must occur in the rule for it to be considered interesting. However, in dealing with KDT, this kind of previous knowledge is not commonly available, and the nature of the information (text) is too complex to be used in discrete terms or structured features.

In this context, some approaches to KDT such as (Nahm and Mooney, 2000a; Basu et al., 2001b) (see 2, section 2.2.1.2) try to measure "interestingness" by using an external resource (e.g., WordNet) which is seen as "previous" knowledge. However, since the approach relies on the resource's organization and specific features, this fails to capture all (or more relevant) semantic relationships between terms. Furthermore, some produced patterns are regarded as "interesting" by the system but uninteresting by human judgments due to the non-existence of some terms in the resource.

Instead, we propose a different view in which the criterion can be evaluated from the semi-structured semantic information provided by the LSA analysis without using external resources. The measure for the hypothesis H is defined as the degree of dissimilarity between the two parts of the hypothesis, and this can be expressed as follows:

```
Interestingness(H)
        = <Dissimilarity between Antecedent and Consequent>
        = 1-SemanticSimilarity(An(H),Co(H))
```

Where $An(H)$ and $Co(H)$ represent the antecedent and consequent of H, respectively.

The criterion reflects the fact that, the lower the similarity, the more interesting the hypothesis is likely to be, meaning that the hypothesis involves an association between its antecedent and consequent which is unseen and so this represents a fact worth exploring further. Otherwise, this can mean that the hypothesis exhibits a certain relation between its parts which may be commonly known and consequently not describe any interesting pattern. Note however, as similarity in terms of LSA does not necessarily mean that the connection is actually obvious between terms, the rest of the elements of either the antecedent or the consequent of the hypothesis might have a key role in determining whether the relation is actually obvious or not. Since interestingness does not look at the hypothesis' structure, it may have some limitations: if two target terms are t_1, and t_2, the antecedent is all about t_1, and the consequent all about t_2, the model may not rank the hypothesis as interesting as t_1 and t_2 might be close according to LSA. However, since the criteria (relevance and interestingness) are addressing independent questions, the interestingness for this case can still obtain a low value. The significance of this value will depend on the trade off with other hypotheses.

Note that because of the random nature of how the initial hypotheses are generated, this criterion could have them assessed as "interesting" and therefore lead the search into misleading decisions. We propose the application of "interestingness" for every hypothesis to be delayed until some conditions are met. Specifically, we would like to measure the criterion (otherwise, it is zero) only when the whole population achieves a certain degree of goodness. As the measure of goodness (fitness) is not a determining factor of the "quality" of the population, we also take into account the degree of representativeness of the produced hypotheses within the search space. Hence two conditions must be met. For the

first condition, we have set an experimental threshold in such a way that the criterion is applied when the average fitness of the whole population is below 1 (note that a fitness value below 1 is associated to a hypotheses in the Pareto set), that is, either a large portion of the hypotheses becomes non-dominated individuals or these have improved in the Pareto set. For the second condition, the average **Coverage** value of the population is considered. When this exceeds 10% (i.e., on average, every hypothesis covers 10% of the rules) and the first condition is met, the interestingness criterion is applied.

Suppose that for the sample hypotheses, we obtained the following similarity values between their antecedent and consequent:

```
For H1:
SemanticSimilarity(
 {determine('the optimum ..'),establish('robust gevillea..')}
 {produce('aspects of the  biology ..')})=0.8

For H2:
SemanticSimilarity(
 {use('white quebracho ..'),carry_out('fitosociological..')}
 {produce(''high coefficients..')}) =0.2
```

then, the interestingness can be computed as follows:

```
Interestingness(H1)=1-0.8 = 0.2
Interestingness(H2)=1-0.2 = 0.8
```

Considering that applying this measure can be delayed in the GA, the clearly high value of interestingness for H2 compared to H1 may not always be a determining factor. In fact, if the threshold for average fitness and average coverage have not both been exceeded both values are defined to be zero, giving no information at all about the quality.

8. **Plausibility of Origin**

As was previously discussed in section 3.2, the Swanson's crossover operator encourages the production of potentially novel hypotheses in terms of a transitivity-like inference. For example, given two hypotheses to be recombined:

`hypothesis(1,A,B)` `hypothesis(2,B',C)`

If Swanson's operator is applied then `hypothesis(<new>,A,C)` is produced. In addition, the computed similarity between `B` and `B'` named S_p, would indicate how accurate the inference is, so the higher the similarity the more plausible the novel produced hypothesis.

Accordingly, the criterion for a hypothesis H is simply given by:

$$Plausibility(H) = \begin{cases} S_p & \texttt{If H was created from a Swanson's crossover} \\ 0 & \texttt{If H is in the original population or is a} \\ & \texttt{result of another operation} \end{cases}$$

Note however that when the evaluation of *Plausibility* comes to play, the GA is unaware of how this hypothesis was produced, meaning that this hypothesis may have been created by other operations. The outcome is worth exploring as a plausible novel hypothesis as long as we make sure that the hypothesis was indeed created originally from a Swanson crossover operator. Although the number of possibilities that may create an offspring is large, three kinds of cases can be identified:

(a) `hypothesis(1,A,B)` and `hypothesis(2,A,C)`: if both are recombined via normal semantic crossover then `hypothesis(<new>,A,C)` may be created.

(b) `hypothesis(1,B,C)`: if the antecedent is mutated to A, then `hypothesis(<new>,A,C)` may be created.

(c) `hypothesis(1,A,B) and hypothesis(2,B',C)`: if both are recombined via Swanson's crossover then `hypothesis(<new>,A,C)` may be created.

Since there is no guarantee that the GA will capture the hypothesis which are worth exploring, *Plausibility of Origin* measures a plausibility of the current hypothesis by "remembering" the quality of the inference when this offspring was created (in the origin) as a result of Swanson's crossover (otherwise, if the hypothesis was created by other operator, its *Plausibility of Origin* is zero).

Unlike the other criteria in which some computation must be performed to measure the hypothesis, *Plausibility of Origin* does not calculate anything but uses the values of similarity already obtained as two parents are recombined via Swanson's crossover, i.e., the semantic similarity between one hypothesis' antecedent and the other hypothesis' consequent (S_p). For this, the value of plausibility of the produced hypothesis is that similarity obtained when the parents were recombined.

Suppose that the sample hypotheses meet the previous constraints so they inherit the similarity from their parents: *Plausibility(H1)=0.9* and *Plausibility(H2)= 0.8*. Therefore, in this respect it can be said that H1 is better than H2.

A hypothesis will have a zero value for plausibility if it is in the original population or arises from another operation. In this case, an early conclusion would be that this is neither plausible nor interesting. However, we assume that even in the absence of *Plausibility of Origin*, the other criteria may have a complimentary role in a way that the hypotheses still can be assessed as valuable.

3.3.2 Multi-Criteria Optimization

Once the genetic operations are performed and the criteria have been assessed, some decision on best and worst individuals must be made, in order to determine which individuals will survive and which ones will not. As a consequence, the current population will be modified to reflect the fact that hopefully better individuals are considered in the next generation (see figure 3.5).

In order to update individuals of the population, we adopted a steady-state approach for evolutionary discovery which in general, is based on the following steps:

1. Select a small number of the best parent individuals of the current generation ac-

cording to their fitnesses from which new offspring will be created. This number is usually referred to as the *Generation Gap* which is commonly suggested to be a small portion of the whole population (e.g., 5% or 10%) (Goldberg, 1989; Mitchell, 1996; Deb, 2001).

2. Reproduce the parent individuals selected in (1) via genetic operators.

3. Replace a small number of the worst (at most, the portion given by the *Generation Gap*) individuals with the offspring created in (2) only if the latter are better than the former.

In single *steady-state* evolutionary optimization (Goldberg, 1989), establishing whether one individual is better than other is straightforward as this only involves computing each hypothesis' fitness independently and then comparing them (i.e., in general, the fitness is the objective function itself). However, since we are dealing with multiple criteria, the process of combining multiple objective functions into a single value is more complex and in general, no hypothesis can be necessarily be said to be better than any other as a whole. Hence we need proper optimization strategies which handle multiple and even conflicting objectives. We specifically adopted the methods commonly referred to as *Evolutionary Multi-Objective Optimization* (EMOO) (Coello, 2000; Deb, 2001).

Our overall algorithm for the GA which performs the evolutionary optimization in which the Pareto-based fitness assignment strategy and the steady-state update are involved is highlighted as follows:

```
Get training data and rules
Initialize population of hypotheses
generation<-0
// MaxGens is the maximum number of generations
// Gg is the Generation Gap
WHILE (generation<MaxGens) DO
  Pareto-based fitness evaluation & assignment  (figure 1.13)
  For each pair of individuals of the Pareto Set:
```

```
      Select 2 hypotheses according to the best fitnesses
         (low values) from  Gg individuals
      Produce 2 offspring using recombination
         according to the   probability of crossover
      Mutate those 2 offspring according to the
         probability of mutation and   the
         kind of mutation randomly chosen
   End-For
   Steady-state based Population Update with Gg individuals
        (figure 1.14)
   generation<- generation+1
END-while
```

The genetic operators were discussed in previous sections so the rest of the chapter will describe the two strategies related to the optimization mechanism: fitness evaluation and population update.

3.3.2.1 Pareto-based Fitness Evaluation and Assignment

As described in section 2.2.2.3, Pareto-based EMOO methods trade-off between the individuals' criteria in such a way that two sets of individuals are created in each learning generation: a set of individuals whose objective functions' values are not all worse than the others' (i.e., the Pareto set), and the set of individuals all of whose objective functions' values are worse than some of the individuals'. Consequently, it cannot be said that the individuals in the Pareto set represent the best solutions but they are not worse than the non-Pareto individuals so far. The improvement through generations will therefore involve encouraging the creation of individuals which become part of the Pareto set.

We took a simple approach in which an approximation to the Pareto set is incrementally updated as the GA goes on, based on dominance relations (see chapter 2.2.2.3).

Given this, we have to face three key problems for obtaining every hypothesis' fitness which constitutes a determining factor to establish the sensible solutions:

```
ALGORITHM (SPEA-based) FitnessAssignment
IN: Population        OUT: ParetoSet, fitness

IF (ParetoSet has not been created (or empty)) THEN
        CurrentPareto ← CollectNonDominatedSolution(Population)
ELSE   CurrentPareto ← previous ParetoSet
/* Reduce size by clustering and update population if needed*/
ParetoSet ← ReduceParetoSet(CurrentPareto)
/* Compute fitness of Pareto members (strength) */
FOR ParetoInd in ParetoSet DO
      count ← 0
      FOR Popind IN Population DO
            IF dominates(ParetoInd,PopInd) THEN
                  count ← count + 1
      Strength ← count/(| Population | +1)
      fitness(ParetoInd) ← Strength
/* Calculate fitness for members of the population */
FOR PopInd in Population-ParetoSet DO
      Sum ← 0
      FOR ParetoInd In ParetoSet DO
            IF dominates(ParetoInd,PopInd) THEN
                  Sum ← Sum + fitness(ParetoInd)
      fitness(PopInd) ← Sum+1
```

Figure 3.13: Algorithm for Fitness Assignment

1. Computing the Pareto set based on the dominance relation discussed in chapter 2.2.2.3.

2. From (1), assigning the fitness values according to the range of goodness of the individuals that are in the Pareto set and those which are not.

3. Encouraging diversity (Deb, 2001; Fonseca and Fleming, 1995; Coello, 2000).

in the individuals by producing multiple groups of fit solutions.

In order to deal with these aspects, our strategy is based on the SPEA approach (see section 2.2.2.3) which uses a mixture of established methods and new techniques in order to find multiple Pareto individuals in parallel, and at the same time to keep the population as diverse as possible through clustering.

A threshold used to restrict the maximum size of the Pareto set is fixed in advance and can be seen as the number of solutions that the GA is expected to create. For example, a threshold of 5% denotes the fact that we want to obtain the best 5% of the population. The role of clustering is then to reduce the Pareto set to fit that portion of the population. Because of clustering, the actual set maintained is only an approximation to the true Pareto set (see chapter 2).

As discussed in section 2.2.2.3, the main tasks carried out by the SPEA algorithm are computing (and reducing if needed) the Pareto Set and assigning the fitness values for both Pareto members and non-Pareto members for further selection decisions.

Note however that the original SPEA algorithm uses an elitist GA (Goldberg, 1989; Mitchell, 1996), so we have adapted it to allow steady-state learning and an incremental updating of the Pareto set. For this purpose, the SPEA algorithm for fitness assignment has been slightly modified as seen in figure 3.13 in which the Pareto set is built once at the beginning. In successive generations, the set is updated using the previous one and the individuals that have to be added or removed. Note however that what is maintained is an approximation of the true Pareto set.

The procedure assumes that the Pareto set used in a previous generation has been consistently updated so that only non-dominated solutions should be part of it. In order to preserve this consistency, once the fitness is assigned and the corresponding operations to create the offspring are performed, new child hypotheses should replace some of the worst parent hypotheses and therefore the Pareto set needs to be updated for the new population. Note that this update avoids computing the whole set in every generation as not all the individuals need to be checked. Consequently, it provides an efficient way to keep the information updated specially when dealing with dominance of individuals in a huge population.

3.3.2.2 Steady-state based Population Update

In order for the individuals of a population (i.e., worst individuals) to be updated ("`Population Update`" in figure 3.5), we propose an steady-state based update strategy as shown in algorithm of figure 3.14.

It is not possible to determine whether a child is "better" than a hypothesis because we do not have independent values of fitness. Instead, the basic idea is to establish these goodness values in global terms, that is, a child is better if this is a candidate to become part of the Pareto set. Specifically, this means that *there is no hypothesis in the whole population which dominates this child.*

While this enables the child to get into the Pareto set, note that it also makes it possible for other members of the set to be dominated by the new one. In this case, the Pareto set is updated by removing any dominated element to become part of the population.

```
ALGORITHM (Steady-state based) PopulationUpdate
Let SetC be the set of Gg children produced by the Gg best parents of the population
      via the genetic operators (Gg is the Generation Gap)
Let SetW be the set of Gg worst parents (higher fitness values)
      from the non Pareto set
For i=1 to Gg Do
    Let W_i the i-th element from SetW (in descending order)
    Let C_i the i-th element from SetC (in descending order)
    IF C_i is not dominated by anything in the population THEN
          Replace W_i with C_i in the population
          Add the new child C_i to the Pareto set
          Remove elements dominated by C_i from the Pareto set to join the
                rest of the population
    END-IF
```

Figure 3.14: Steady-state based Population Update

As a consequence, the GA will behave in such a way that the Pareto set increases

or decreases in size allowing new non-dominated hypotheses to be added or new dominated ones to be removed. This is the reason why it is not possible to have the group of "not-worst" solutions monotonically increasing. In practice, this behavior before the reduction (clustering the Pareto set) can be seen in the graph shown in figure 3.15 in which, as an example, the GA was run for 200 generations with a small population of 100 hypotheses. The maximum allowed size for the Pareto set in our system is fixed as 5% of the population, but this is ignored in this example.

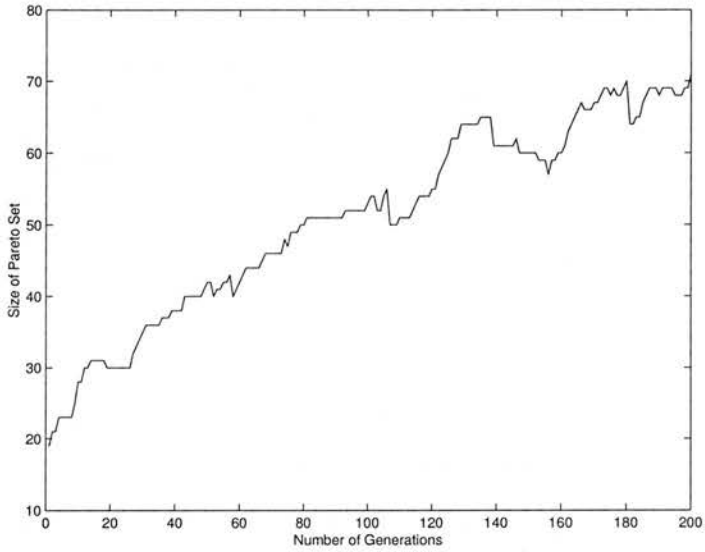

Figure 3.15: Evolution of the Pareto set before Clustering

Note that the Pareto set shows a tendency to grow. However at some times this is unstable which reflects the fact that in some generations, new fit solutions are added and so less fit individuals are removed from the set to become part of the non-Pareto solutions (see algorithm in figure 3.14). This is the case (approximately) for generation 60, 105, 155, 175, and so on.

One of the main outcomes of the `SPEA-based_FitnessAssignment` procedure is the fitness values assigned to every hypothesis. However, unlike traditional GAs, the fitness value (whether it is for Pareto or non-Pareto members) cannot be taken as the only way to compare solutions, except when selecting individuals for reproduction and for replacement.

3.3.2.3 A Worked Example

In order to clarify how the Pareto-based strategy works to assign the fitness and to update the population as highlighted in the algorithms of figure 3.13 and figure 3.14, consider the following example:

There is a sample population of 8 hypotheses

```
[70, 6, 20, 56, 1, 7, 15, 36]
```

Next, two questions will be addressed: which hypotheses will be reproduced and how will the population be updated?

For the first question, we have to distinguish the worst from the not-worse individuals. This is accomplished by computing the Pareto set the first time (later, it only needs to be updated) in terms of the hypotheses' vectors, and then by assigning the proper fitness values.

The hypotheses' criteria values (e.g., relevance, structure, cohesion, etc) are represented as components in a vector. Since that we are trying to maximize these criteria, one individual dominates another if all the criteria values of the former are greater (or equal) than for the other. For example: given the following vectors for two individuals A and B:

```
For A: [0.2, 0.2, 0.6, 0.6, 0.5, 0.4, 0.7, 0.1]
For B: [0.4, 0.5, 0.7, 0.8, 0.9, 0.5, 0.7, 0.2]
```

B is said to dominate A (or A is dominated by B). Note also that if the dominance condition is not true, then an individual is said to be non-dominated by the other, as in the following case:

```
For C: [0.2, 0.5, 0.6, 0.8, 0.9, 0.4, 0.7, 0.2]
For D: [0.4, 0.3, 0.7, 0.6, 0.5, 0.5, 0.7, 0.1]
```

where C is not dominated by D and D is not dominated by C. Keeping this in mind, assume that from the vectors of the population, five solutions are not dominated by any

individual, so the Pareto set and non-Pareto set become: `[70, 6, 20, 56, 1]` and `[7, 15, 36]`, respectively[7].

Consider that we have a Pareto size limit of 3. Since the actual size is 5, a reduction of the set needs to be performed. By clustering the elements' vectors according to how close to each other they are (average link clustering), we might obtain the following clusters for the Pareto set:

```
[ [70, 1, 20], [6], [56] ]
```

In order to obtain the new (reduced) Pareto set, we take one individual from each cluster whose average distance to the rest of the group, in terms of their vectors, is minimal (*average link clustering*). Thus, supposing that 20 is the center of the first cluster, the reduced Pareto set is `[ 20, 6, 56 ]`, and the individuals not used, 70 and 1, join the non-Pareto solutions: `[70, 1, 7, 15, 36]`. Note that because of this clustering, the algorithm only maintains an approximation of the Pareto set as some non-dominated elements are removed from the original set.

Next, the fitness for both Pareto and Non-Pareto members is calculated using the algorithm in figure 3.13. The results can be seen in table 3.1 where lower fitness values indicate better solutions. In the table, **H** denotes the hypothesis, **Hd** and **dH** are the hypotheses dominated by **H**, and the Pareto hypotheses which dominate **H** respectively, and the corresponding fitness values for the Pareto hypotheses (strength) and non-Pareto hypotheses.

Note that "fitness" is computed differently according to whether the hypothesis is a member of the Pareto set or not (see algorithm in figure 3.13).

Once fitness is computed, we can obtain a ranking of individuals according to these values. For example, among the Pareto individuals, the best can be either 6 or 56 (both have the same strength 0.22). Among the worst individuals, the worst one is 36 which has the largest fitness 1.77. Note that the outcome of this procedure is also consistent with the Pareto set before being reduced, that is, despite the solutions 70 and 1 being considered non-Pareto individuals, their fitness still reflects the fact that they are the best ones among the worst solutions as they still are non-dominated. The

[7]The whole population is composed of the combination of Pareto and non-Pareto individuals

H	Hd	dH	Fitness (Pareto)	Fitness (Non-Pareto)
20	[7,15,36]	-	$\frac{3}{(8+1)}$ = **0.33**	-
6	[15,7]	-	$\frac{2}{(8+1)}$ = **0.22**	-
56	[36,7]	-	$\frac{2}{(8+1)}$ = **0.22**	-
70	-	[]	-	0+1= **1**
1	-	[]	-	0+1= **1**
7	-	[20,6]	-	fitness(20)+fitness(6)+1=**1.55**
15	-	[56]	-	fitness(56)+1= **1.22**
36	-	[20,6,56]	-	fitness(20)+fitness(6)+fitness(56)+1= **1.77**

Table 3.1: Example for Fitness Assignment (algorithm in figure 3.13)

Pareto solutions ranked from best to worst are accordingly: `[6, 56, 20]`, and the Non-Pareto solutions, ranked from worst to best: `[36, 7, 15, 70, 1]`.

Now we are able to answer the question of which individuals should be reproduced, and which ones should be selected for updating. With a *Generation Gap* set to 2, the parents to be selected for reproduction are individuals 6 and 56. The worst parents to be updated are individuals 36 and 7.

Next, consider that two new children C_1 and C_2 are created from the selected parents via recombination and mutation, and as a consequence: *No solution dominates* C_1, C_1 *dominates solution 56, Solution 15 dominates* C_2.

The population update is performed on a pairwise basis (see update algorithm in figure 3.14) as long as the conditions are met, that is, 36 may be replaced with C_1, and 7 with C_2, respectively. The replacement proceeds as follows: C_1 meets our update condition previously described, so hypothesis 36 can be replaced with C_1. In addition, child C_1 also dominates one of the Pareto individuals (56), so the latter must be removed from the Pareto set. For the second possible update (hypothesis 7), it is provided that one solution (15) dominates C_2, therefore, the replacement does not take place. In other words, the new child C_2 is worse than some member of the population, so it cannot be included. As a result, the updated sets will look like:

```
Pareto Set=[20,56,C1] (6 removed)   Non-Pareto Set=[70,1,15,36,6]
```

and then the next generation proceeds. Finally, it is worth pointing out some features:

- The population size is preserved at all times. The Pareto set is a subset of the population.
- In the adapted SPEA algorithm, computing the whole Pareto set is only required at the beginning. In further generations, this is just updated so to make the process more efficient. Note that the update is enough to keep the set and population consistent in terms of maintaining an approximation of the Pareto set which "captures" close individuals in the neighborhood in the Pareto front.

3.3.3 Summary

In this chapter, a model for domain-independent knowledge discovery from texts has been proposed. Here, the process of search for potentially novel knowledge is performed by an multi-objective evolutionary algorithm which is guided by semantic and rhetorical information so as to create plausible hypotheses. For this end, a number of strategies (metrics) have been developed for objectively evaluating the quality of the discovered hypotheses.

Unlike other approaches, our model aims at generating novel knowledge only using information from the original corpus of text documents and from the training data generated from them.

The initial knowledge for the model involves a set of rules that represent the documents, and training information obtained from these rules and from the whole corpus. The approach assumes that only parts of a document need to be represented so as to capture key facts by making use of the underlying genre. For this, IE patterns have been designed to extract domain-independent but genre-based information (for scientific abstracts) from texts. Training information is extracted from the analysis of associations from this rule set, and from the lexical semantic knowledge provided by *Latent Semantic Analysis* (LSA) by analysing the whole corpus. The structured information contained in the rules is used to compliment the structure-free information provided by LSA so as to make further semantic similarity judgements.

The underlying nature of the model suggests that it might be possible to conceive an effective KDT approach independent of domain resources and to make use of the underlying rhetorical information so as to represent text documents for text mining purposes.

Chapter 4

Experimental Results and Analysis

In order to assess the outcome of our model, a prototype was built and further experiments were carried out on it. The IE task has been implemented as a set of modules whose main outcome is the set of rules extracted from the documents. In addition, an intermediate training module is responsible for generating information from the LSA analysis and from the rules just produced. All this information, expressed in a fact-like form, feeds a *Prolog* system in which the GA-based KDT has been implemented.

For the purpose of the experiments, the input technical corpus of documents was obtained from the **AGRIS** database collected from the *Food and Agriculture Organization* (*http://www.fao.org*) in the Spanish language. We selected this corpus because it has been properly cleaned-up, and covers a scientific area we do not have any knowledge about, which avoids any possible bias and makes the results more realistic.

A set of 1000 documents was randomly obtained from the corpus without any restriction (i.e., any abstract is useful as long as it belongs to the domain of interest in the corpus). From this set, one third were used for setting parameters, suggesting patterns for the IE and making general adjustments, and the rest was used for the GA itself in the final evaluation stage.

Using this basic information, in section 4.1 we investigate the behavior of some aspects of the model. Specifically, we discuss issues concerning the information extraction task, and the similarity judgments via LSA. The purpose of this is twofold. First, we want to provide a general flavour of what these activities look like so as to understand some issues of the search process. Secondly, we want to highlight points

where weaknesses could be in the context of the search ability of the model.

Secondly, in section 4.2, we address the research questions which allow us to evaluate the adequacy of the model according to the aims stated in chapter 3 by investigating the search ability of the whole model in terms of the evaluation criteria, how the outcome of the model is judged by human experts, and how local issues may influence the results of the prototype.

Note that the actual experiments were performed from material in Spanish but in order to provide a clearer explanation these have been translated into English when appropriate.

4.1 Investigation of Basic Properties of the Model

We investigated the model in terms of some local aspects which may influence the quality of the search for novel knowledge. Specifically, we highlight those involving how well the system does in extracting the key information from the documents (e.g., IE), and in making similarity judgments (e.g., LSA). This analysis aims to provide a general flavour of what the LSA and IE tasks look like so as to understand issues concerning semantic similarity which guide the search and information extraction which provide basic information to feed the hypotheses of the search process and to identify weaknesses. In addition, the investigations here are informal, in contrast to the formal evaluation of 4.2.

4.1.1 Information Extraction

The process of Information Extraction has been analysed using the standard evaluation metrics **Precision** and **Recall**. Given the large number of documents, doing the evaluation by hand would be a time consuming task so we only took a representative set of 20 documents and performed the evaluation as shown in table 4.1. For every document (D) the corresponding rule extracted by the system is analyzed in terms of the rhetorical roles and predicates recognized. The following aspects were captured and manually judged by two computer scientists and one linguistic:

- Number of *Answers Produced (AP)*: the number of roles (along with their predicates) produced.
- Number of *Possible Correct Answers (PCA):* number of roles (and predicates) which should be extracted from the document.
- Number of *Correct Answers (CA):* the number of roles (and predicates) extracted that are correct.

D	CA	AP	PCA
1	2	2	3
2	2	2	2
3	2	3	2
4	0	1	2
5	2	2	3
6	1	1	3
7	2	2	5
8	2	2	5
9	1	1	2
10	4	5	4

D	CA	AP	PCA
11	1	1	2
12	2	3	4
13	2	2	3
14	5	8	7
15	3	3	4
16	1	1	1
17	1	2	2
18	1	2	1
19	3	3	3
20	1	2	3

Total CA	38
Total AP	48
Total PCA	61
Precision	0.79
Recall	0.62

Table 4.1: IE Evaluation Metrics

Then, Precision and Recall were calculated. These partial results show that despite having handcrafted IE patterns that might be somewhat imprecise, most of the answers extracted are correct (nearly 80%). However, their recall is relatively low as they miss nearly 40% of the answers (roles and predicates) from the documents, which may suggest a weakness in the system as a whole.

Note that even if a large-scale handcrafted evaluation was produced, there is still a key issue that will not be captured for the IE evaluation and that needs to be dealt with. Specifically, extracting the whole information from this evaluation viewpoint does not ensure that all the information to produce good hypotheses will be available.

Although there are many informative abstracts in the corpus, the IE task can fail to produce useful information because the contents of these abstracts can be too general, provide poor information or miss explicitly mentioning key facts. For example, one abstract's fragment:

```
In this work, several conditions of the soils are
described and showed graphically...
```

can produce the following representation: `object(describe(911))`, where "911" is the argument internal representation. Although this answer is textually correct, this does not state anything important from the text. For example, the text states that "several conditions ..are described" but it does not make explicit which conditions they are.

Because of the nature of many abstracts, some constituents can not be identified either because they are absent completely or because they are too implicit. By examining the whole set of documents from the initial corpus and the rules extracted from them in detail, we indeed observed the following:

Of a sample of 336 produced rules, only 95 of them (28.2%) contain some conclusion, 88 of them state some method (26%), and 96 of them state the goal (28.5%). This will make it demanding for the learning method to achieve "cause-effect"-like hypotheses as this mostly depends on the size of the training corpus and the level of description of the abstracts.

Provided that informative abstracts (i.e., those providing some cause-effect relationship) are not a majority in a corpus (otherwise, information from the full documents should be considered), the IE task could be improved by adding anaphora resolution capabilities so to deal with more specific entities from the documents. In addition, providing some trainable classification task which allows for the recognition of rhetorical and semantic information in a more flexible way, would also be a valuable feature compared with our fixed IE patterns.

4.1.2 Simulated Similarity Judgments via LSA

In this analysis, we tried to informally assess some aspects of LSA in measuring semantic similarity at the term and at the predicate level. The aim of this is to provide an informal flavour of what the LSA task looks like so as to understand some semantic similarity issues of the search process.

The performance of the similarity evaluation through LSA can be seen at different levels from terms to the whole hypotheses. Note that the actual size of the corpus is 141,307 words spread over 1000 abstracts. Although this is not a huge corpus of data as usually suggested by (Kintsch and D. Steinhart, 2000) for the purpose of data analysis, we are also using additional linguistic information (syntactic, rhetorical), so the lack of a larger corpus may to some extent be dealt with by complementing the basic lexical information.

From a term-to-term similarity point of view, some terms drawn from the corpus and their LSA similarity are shown in table 4.2 (examples are translations from original data in Spanish). Some specific terminology seems to be highly semantically connected such as `enzyme` and `zinc`. In fact, these deserved further attention when the experiments were run (see table 4.7). For other general-purpose terms, the similarity is not this clear. For example, the high relatedness between `climatic` and `performance` is quite misleading as the terms have nothing to do with each other. The context information captured by LSA shows that they occur in similar contexts, which is also the case for `nitrogen` and `september`. This kind of situation is unclear for `dry` and `rain`, and `dry` and `humidity`. For the former, this is low to reflect the fact that they are unrelated. For the latter, it can be assumed that both terms refer to some conditions found in the "soil" being talked about. The relatedness determined for `empleado` and `tratadas` (they are left in Spanish so to stress the difference, meaning *employee* and *handled* respectively) is intriguing as they are related to treatments or use of procedures ("empleado" is treated as a verb meaning *used*). However, `empleado` is also a Spanish noun meaning "`employee`" which is semantically unconnected from `tratadas`, at least in a general sense. Considering that LSA does not make syntactic distinctions, the similarity seems to be doing a fair syntax-independent prediction.

But does the additional linguistic information contribute to resolve some of the

Term1	Term 2	Similarity
enzyme	zinc	0.9998
climatic	performance	0.9999
nitrogen	september	0.7401
dry	rain	0.3689
dry	humidity	0.8986
empleado (employee)	tratada (handled)	0.1101

Table 4.2: Term-to-term Similarity

terms above (apparently) incorrectly related? We think it does. In using the terms within their contexts and the predicate-level information, some interesting facts can be noted. Table 4.3 shows examples of similarity between predicates (and arguments) which contain the terms above.

Predicate 1	Predicate 2	Similarity
analyze('...enzyme..')	perform('...zinc..')	0.99
produce('..climatic ...')	perform('..performances..'),	0.78
use('..nitrogen..')	study('..september..')	0.96
perform('..dry..')	use('..rain..')	0.74
describe('..empleado..')	effect('..tratadas..')	0.60

Table 4.3: Predicate-Predicate Similarity

For example, although "`analyze`" and "`perform`" are likely to have some high relatedness in terms of carrying out some activity, the similarity between both predicates seems to be consistent with the term-to-term measure in a way that the predicates in this case confirm the high association. For "`climatic`", the context and the predicates say something different, that is, considering the predicate actions and the rest of the elements of the corresponding argument, the similarity is less than for the terms alone. In other words, these should not be highly related. Note that for "`nitrogen`",

this situation has an opposite effect: in considering additional information, it seems to be promoting a higher similarity presumably because the context of "september" contributes further semantic information to support the fact that they should be highly connected.

For the last cases in table 4.3, the situation seems to be not that clear. The differences in providing further information for "dry" does not seem to be significant. For "empleado", the additional information seems to be contributing to have these two terms more related, but it is still not sufficient (0.60). Given the different roles of the words (nouns and verbs), it turns out that LSA would need more training information so as to be more certain about this relatedness.

Similarity judgments are also exhibited in other aspects of the evaluation. For example, the following hypothesis from run 1:

```
hypothesis(3,
[object(perform(16321)),object(determine(25011)),method(study(18931))],
[conclusion(find(24511))]
```

whose predicates' content are about:

```
- To perform the material plantation in a experimental farm..
- To determine the forage production on raining soil..
- To study the ethylen and soluble seeds..
- To find an equation for forage production..
```

respectively, has an average coherence of 65% between its conditions (e.g., paragraphs). Intuitively, the contents of its predicates seem to be around related concepts (plants, soil, seed), so the objective value exhibits a good prediction.

Unfortunately, for other cases the role of average coherence values tends to be misleading. The following hypothesis:

```
hypothesis(67,
          [goal(determine(25011)), object(determine(25011))],
          [conclusion(study(19411))]).
```

whose predicates' content are about:

```
- To determine the size and diameter of the stake...
- To determine the size and diameter of the stake...
- To study the growth of ''canescens"  and their
  morphological components...
```

has a coherence of nearly 60% which would suggested that the semantic connections between paragraphs are not extremely good but show some relatedness. This is the result of the average of the LSA similarity between the two conditions (100%) and that between the last condition and the conclusion. However, as the conditions of the hypothesis are duplicated, the real coherence should have been determined only between the first condition and the consequent, but in evaluating the hypothesis, the GA is unaware of this fact.

In summary, LSA seems to be a plausible method to support the system's ability to measure semantic similarity at the different levels of representation (i.e., terms, predicates). However, the predictions of similarity might be affected by the size of the training corpus as LSA usually requires large amounts of input texts to be able to make accurate semantic similarity predictions. In addition, the informal experiments suggest that providing additional information (i.e., predicate-level terms) beyond keywords improves the prediction of the simulated LSA method.

4.2 Answering the Research Questions

The aims of our model is to prove that it is plausible to conceive an effective KDT approach independent on domain resources and to make use of the underlying rhetorical information so to represent text documents for text mining purposes.

In order to evaluate the adequacy of our model as stated by the specific aims in chapter 3, the different experiments discussed in this section try to address three research questions:

1. *How plausible is the search of evolutionary KDT?*

2. *Is the knowledge discovered by the model effective in terms of the quality of the outcome and its correlation with human judgments?*

3. *Does the model outcome contribute additional information to help one better understand the nature/origin of the discovered knowledge, compared to BOW approaches?*

In order to address these questions, we first tuned the GA-based KDT by using 300 rules (out of 1000 training rules) for adjustments and parameter setting. The rest of them (700) were used for running the model and generating the final hypotheses. The global parameters were set as follows: *Mutation probability* and *Crossover probability* have values 0.2 and 0.8, respectively (see sensitivity analysis in section 4.2.1). Note that the mutation rate is relatively higher than usual as the corresponding operations have additional constraints which must hold before they are applied. The *Generation Gap* is set up to 5% of the population (typical recommended gap should not exceed 10%). The size of the Pareto set has been fixed to 5% of the population so that a small number of fit hypotheses is produced. In order to enable similarity judgments which are really relevant, a similarity threshold has been set to 98%, that is, only similarity values exceeding this threshold are considered. This is because all the information in the corpus which is within the same domain, tends to be very highly related. The size of the population to be used is 100 (initial random hypotheses).

Next, we used a methodology consisting of two phases described as follows:

- *Investigation of the model's search ability:* this aimed at investigating the behavior and different search related issues of the evolutionary mode so as to answer research question 1.

- *Expert Evaluation:* this aims at effectively assessing the quality of the discovered knowledge by domain experts so as to answer research questions 2 and 3.

4.2.1 Investigation of the Model's Search Ability

The quality of the search process can be analyzed by observing the typical behavior of the GA in terms of the performance of the genetic operators in generating fit solu-

tions, its robustness (i.e., Does it always find good solutions?), and the quality of the hypotheses in terms of the objective functions.

- *Genetic Operators:* the aim here was to investigate how sensitive the GA is to different parameter values. Because of the large combination of parameter settings, we concentrated on the probabilities of crossover and mutation only, in terms of the fitness of the produced solutions. Note that because of the nature of the SPEA-based strategy, low fitness values are desired.

  Test parameter values were established as shown in table 4.4 for 20 runs of the GA, each up to 1000 generations, with a initial population of 100 hypotheses. Here, different probabilities of mutation (P_m) and crossover (P_c) are tested, and the resulting average fitness of the population, its standard deviation, and the minimum and maximum values of fitness are shown (the rest of the parameters remain the same).

  The parameters were systematically tested with steps of approximately 5% (starting from 0.025) for P_m, and 10% (starting from 0.50) for P_c. The final range for P_m is from 0.025 to 0.50, whereas for P_m, this is from 0.50 to 0.80. Thus, the table shows the different settings involved moving through the range for P_m and fixing a value for P_c. For example, the first 5 runs consider setting $P_c = 0.50$ fixed and testing with different values of P_m.

  Some aspects of the resulting values are worth highlighting:

  - Although finding good solutions is no guarantee that the search process is effective because human judgment is not considered, the GA seems to be able to find good hypotheses, that is, individuals with fitness zero or close to zero.
  - Because of the constrained genetic operators, small changes in the parameter values do not have a significant effect on the best obtained fitness. This can be clearly visualized from run 1 to 10 where the lowest fitness is unchanged. Despite this, with a higher mutation probability (and crossover constant), the average fitness of the population decreases, indicating that

Run	P_m	P_c	Avg. Fitness	Std. Dev	Min. Fitness	Max. Fitness
1	0.025	0.50	0.0911	0.0790	0.0099	0.2495
2	0.075	0.50	0.0833	0.0746	0.0099	0.2495
3	0.125	0.50	0.0934	0.0746	0.0099	0.2495
4	0.175	0.50	0.0934	0.0740	0.0099	0.2297
5	0.2	0.50	0.0799	0.0701	0.0099	0.2297
6	0.025	0.60	0.0625	0.0601	0.0099	0.2188
7	0.075	0.60	0.0725	0.0600	0.0099	0.2188
8	0.125	0.60	0.0623	0.0602	0.0099	0.2188
9	0.175	0.60	0.0625	0.0600	0.0099	0.2188
10	0.2	0.60	0.0602	0.0583	0.0099	0.2188
11	0.025	0.70	0.0323	0.0617	0	0.2495
12	0.075	0.70	0.0358	0.0622	0	0.2495
13	0.125	0.70	0.0358	0.0619	0	0.2495
14	0.175	0.70	0.0316	0.0619	0	0.2495
15	0.2	0.70	0.0301	0.0958	0	0.4950
16	0.025	0.80	0.0230	0.0556	0	0.2495
17	0.075	0.80	0.0329	0.0553	0	0.2495
18	0.125	0.80	0.0240	0.0567	0	0.2495
19	0.175	0.80	0.0221	0.0543	0	0.2495
20	0.2	0.80	0.0209	0.0470	0	0.1881

Table 4.4: Analysis of the behavior of the GA to different parameters

the mutation may be effective in improving the quality of solutions in this range of parameter values.

- Although higher values of P_m and P_c might improve the overall performance of the GA by decreasing the population fitness values, sometimes the maximum fitness values tend to increase despite the overall improvement of the population (see runs 11 to 19, compared to runs 6 to 10).

- As the parameter values increase, there is a tendency for the minimum fitness to decrease. However, note that because of the multi-objective nature of the model, having low (or zero) fitness values between runs 11 and 20 does not necessarily imply that there are no changes in individual criteria of the best solutions. Indeed, considering individual objective values, the best solutions may be those with the lowest fitness values. However, as a result of the operators, the solutions can be modified and their objectives values can slightly increase/decrease as they are still in the Pareto set.
- Sudden peaks (e.g., average fitness of run 3, 7, 12, etc) can also be explained because of decisions on dominance, e.g., some less fit solutions leaving the Pareto set.

This analysis shows that increases in both mutation and crossover can have a positive effect on the quality of the solutions. However, the role of the combined effect of both operators can not be completely visualized from the table above.

In order to investigate such a role, we tested the GA on extreme parameter values to see how the GA search proceeds. The resulting behavior of the best fitness can be seen in figure 4.1, in which the best fitness values[1] along 1000 learning generations are shown. Three basic testing cases are considered:

1. *Mutation enabled ($P_m = 0.2$), No Crossover ($P_c = 0.0$).*
2. *No Mutation ($P_m = 0.0$), Crossover enabled ($P_c = 0.8$).*
3. *Normal case using best parameters from table 4.4: Mutation ($P_m = 0.2$), Crossover ($P_c = 0.8$).*

It is not surprising that in case 2, no major improvement in the search is achieved after the few first generations. In having no mutation, the system is unable to improve beyond the few first generations. As no new changes are introduced, the quality of the fitness stabilizes at an early stage (approx. generation 20).

[1]Given that fitness values have to be minimized, low values are looked for, meaning that either dominated individuals are improved and moved to the Pareto set or Pareto set members are improving their "quality", that is, decreasing their fitness values.

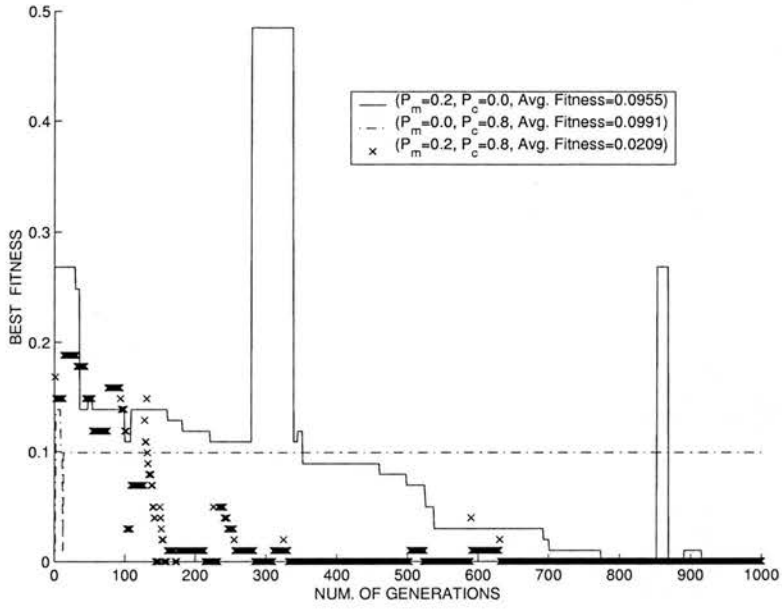

Figure 4.1: System Behavior under different Parameter Values

In the presence of mutation only (case 1), the GA starts with worse fitness values than for case 2. However, after generation 350, it is able to find some better solutions than the previous case. While the difference in the average fitness for both settings is probably not significant, the combined use of mutation and crossover in case 3 shows a marked difference in its fitness compared to the other cases (from 0.0955/0.091 down to 0.0209). In addition, having mutation and crossover enable the GA to produce good fitness values in nearly all the generations, indicating that this has a more steady improvement. Indeed, the average fitness of this case compared to the previous ones improves significantly in nearly 80%.

In summary, we show that the performance of the GA is sensitive to parameter values and so higher probability values suggest better results in terms of the overall bahavior.

- *Robustness:* we investigated some aspects of the robustness of the GA, and therefore the GA-based KDT, by doing a series of experiments in which the GA was run multiple times (e.g., 5) with identical parameters and target concepts. This

aimed at addressing two basic questions:

1. *Does it always find good fitness solutions according to its own fitness definition (human judgments are not considered here)?*

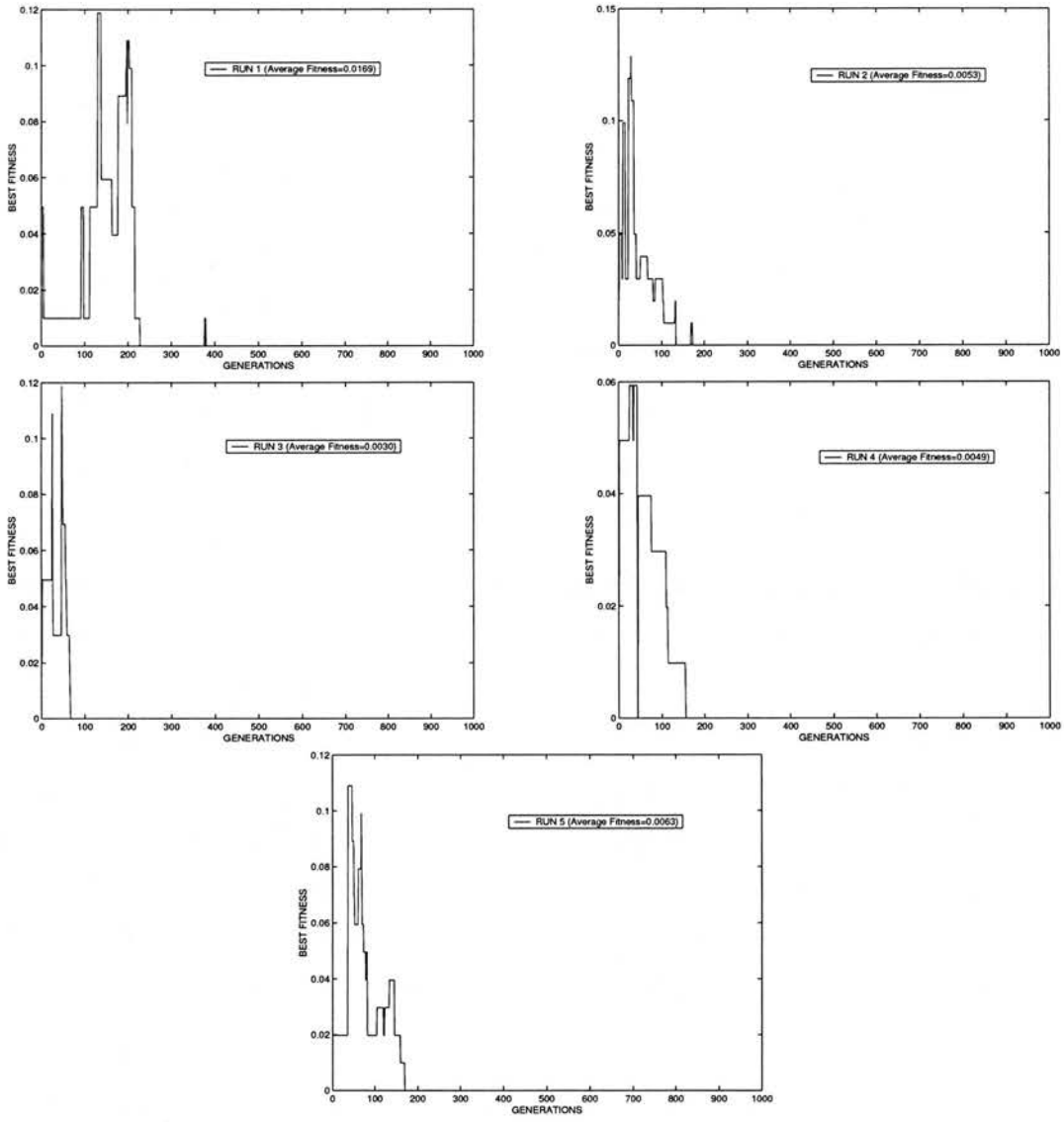

Figure 4.2: Different runs with the same parameters

As can be seen in the graphics of figure 4.2, despite some unstable periods due to dominance conditions, the GA manages to find good solutions. Solutions with low fitness (zero) are always found after generation 200. This

is consistent with the run already shown in figure 4.1 in a way that in different runs with the same parameters, the GA was able to find slightly better solutions than in case 3 in figure 4.1 (average fitness=0.0169) or much better (average fitness=0.0030). However, as previously highlighted, note that with steady zero fitness the GA may still be improving the solution's individual criteria.

2. *Does the GA always find the same solutions?:*

   Since there are no target solutions to compare with, there is no guarantee that the GA will always find exactly the same hypotheses. However, given the steady-state strategy and the way the genetic operators work, different runs may find the same material (i.e., predicates) shared by different hypotheses.

	Run				
Predicate	1	2	3	4	5
describe(9111)	√	√	√	√	√
perform(2631)	√	√	√	√	√
describe(1011)	√	√			
perform(27121)	√	√			
describe(13421)	√	√		√	√
perform(9021)	√		√		
produce(1321)		√	√	√	
describe(19511)		√	√		
perform(9711)		√	√		
produce(2661)		√	√		
was(28631)			√	√	

Table 4.5: Common Features across runs

In order to see common features across the different runs, we took the five solutions produced in every run and looked for pairs of runs where the GA found the same predicates (and arguments). The shared material is shown

in table 4.5. Indeed, the table indicates that the GA is able to find good material across the runs. For example, predicates `describe(9111)` and `perform(2631)` are contained in the best hypotheses of all runs. The GA was able to find a good hypothesis containing the same predicate `perform(27121)` for runs 1 and 2, the predicate `produce(1321)` for runs 2 and 4, and so on. This suggests that the GA does find the same good material (for some, more than twice) across the runs.

Run	Same Predicates	Num. Hypo.
1	**describe(13421)**	3
	describe(9111)	2
	produce(24321)	2
2	perform(9711)	2
	produce(16011)	2
3	describe(23711)	3
	perform(2631)	4
	describe(9111)	2
4	var(15411)	2
	establish(22821)	2
	perform(2631)	3
	describe(9111)	4
5	accumulate(19431,19432)	2
	use(7811)	2
	describe(10411)	3
	describe(9111)	4

Table 4.6: Runs containing shared material in the hypotheses

It is important to note that this kind of common feature is not a coincidence. Within a run, the GA is able to find the same material (i.e., predicates) across the best hypotheses indicating that despite not having target solutions, the system is still able to "capture" good material shared by most of the solutions. In order to show this, table 4.6 contains a summary of

the runs and the predicates that the GA finds for more than one hypothesis (from the Pareto set) in that run.

For example, three hypotheses in run 1 contain the predicate `describe(13421)`, three hypotheses in run 4 contain the predicate `perform(2631)`, etc. Note that there is a constant suggestion of the GA that good material of one run is also included in the good material of a different run. Indeed, the GA finds `describe(9111)` in two hypotheses and it also finds this material in the rest of the runs in table 4.5. For run 1 in table 4.6, the GA finds the same predicate `describe(13421)` in three hypotheses and it also finds it in the best hypotheses of runs 1, 2, 4, and 5.

- *Objective values of criteria:*

As the genetic operators prove to be performing as expected, the next step is to investigate the search ability of the whole model.

The search is analyzed in terms of the results produced for the different criteria evaluated for each hypothesis. To this end, the average fitness for the criteria across the five runs are shown in figures 4.3, 4.4, 4.5, and 4.6. Note that no absolute answers can be provided as there are no target solutions to be compared against, which is the usual case for KDD problems. For this, a more precise answer is given by complimenting the system evaluation with the answers provided by the experts later discussed in section 4.2.2.

The results are drawn from GA runs for 1000 generations for five different pairs of target concepts, and every graphic represents the average search results for a criterion across the population for each of the five runs. Keep in mind that, unlike fitness which has to be minimized because of the way it is calculated, in the current runs, maximum objective function values are looked for.

Some interesting facts can be noted. Almost all the criteria seem to stabilize after (roughly) generation 700 for all the runs, that is, no further improvement beyond this point is achieved and so this may give us an approximate indication of the limits of the objective function values. Considering that each run of the model has been performed with different target concepts, these limits appear to

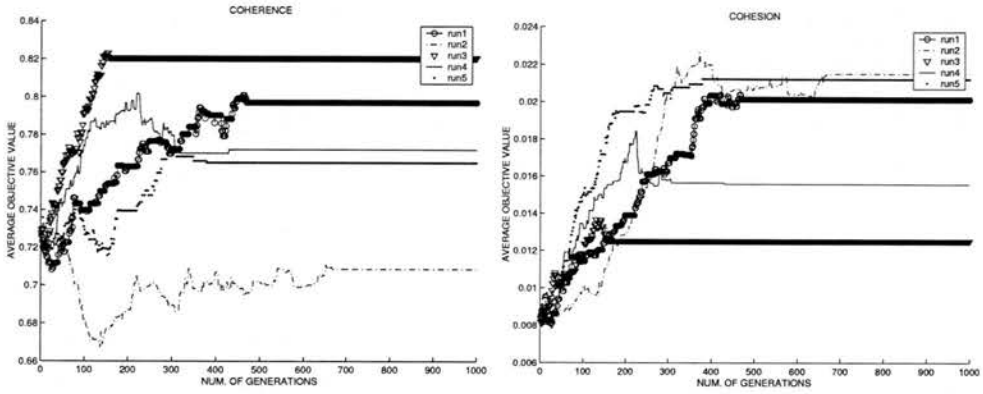

Figure 4.3: System Evaluation for Coherence and Cohesion

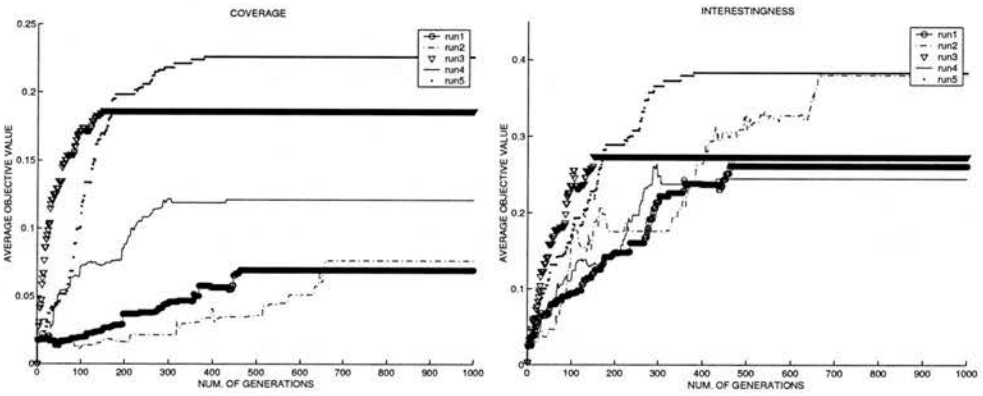

Figure 4.4: System Evaluation for Coverage and Interestingness

be a plausible factor to be taken into account for further experiments in terms of upper and lower objective values. However, having this level of stabilization does not ensure that the method achieves the best objective values as this depends on the dominance decisions made in the optimization stage.

Another aspect worth highlighting is that despite a steady-state strategy being used by the model to produce solutions, the individual evaluation criteria behave in unstable ways to accommodate solutions which have to be removed or added. As a consequence, it is not necessarily the case that all the criteria have to monotonically increase.

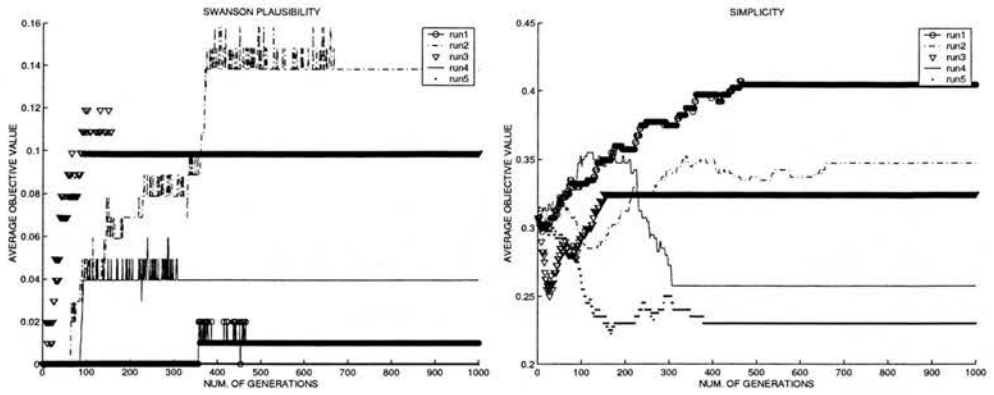

Figure 4.5: System Evaluation for "Plausibility of Origin" and Simplicity

In order to see this situation, look at the results for the different criteria for the same period of time, between generations 200 and 300 for run 4. For an average hypothesis, the qualities of *Coherence*, *Cohesion*, *Simplicity* and *Structure* get worse[2], whereas this improves for *Coverage*, *Interestingness* and *Relevance*, and has some variations for *Plausibility*.

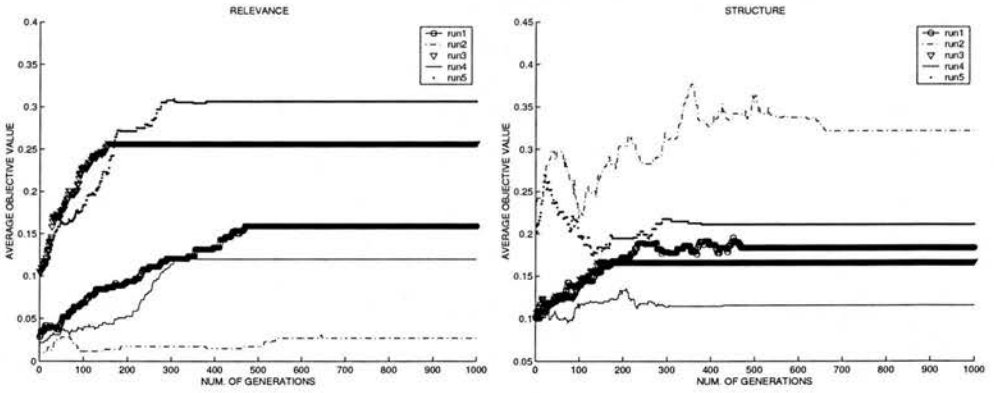

Figure 4.6: System Evaluation for Relevance and Structure

Although some of the individual criteria show overall improvement, this cannot be taken as the best possible behavior for these criteria as their results change

[2]"Worse" means the quality from the single criterion point of view. When trading off with other criteria, low objective values may not be undesirable as long as they trade off correctly with other objectives.

across runs that use different target concepts. However, for some cases, including *Coherence* (figure 4.3), the better results for run 1 above the rest could be regarded as desirable due to the high objective values achieved.

In summary, the behavior of the GA in terms of individual criteria and its robustness under different parameter values suggest that the search and the criteria computation indeed behave in a plausible way.

4.2.2 Expert Evaluation

We ran five versions of the GA with the same configuration for all of them but different pairs of target terms. The purpose of these concept pairs is twofold. First, these are used as a guide in the search process in order to look for relevant knowledge as assessed by the *Relevance* criterion (see section 3.3 in chapter 3). Secondly, these act as original terms extracted from a BOW text mining system. Accordingly, one of the aims of the model is to find hypotheses that explain the underlying relationship between them (research question 3). As we do not have an off-the shelf system to extract such input terms, we used the same data provided by LSA for terms and then applied a basic clustering analysis to come up with a candidate list of pairs of close terms.

From the previous list, we selected those pairs which are worth analyzing by a domain expert. As we are non-experts, we discarded pairs which exhibit trivial or commonsense connections (e.g., "cow" and "animal"). Next, one domain expert was asked to pick up the best pairs that were worth exploring.

With these relevant pairs, we ran five versions of the model, each one using the same global parameters but a different pair of target terms as shown in table 4.7. Given the parameters above, each run produced the best 5 hypotheses, that is, there as a total number of 25 hypotheses which were later assessed by the experts.

We then designed an experiment in which 20 domain experts (from Agriculture) were involved to assess the hypotheses generated by the model and whose personal information is summarized in table 4.8. These experts were involved in the experiment in such a way that:

- Every hypothesis was assessed by two or three randomly selected experts. The

Run	Term 1	Term 2
1	enzyme	zinc
2	glycocide	inhibitor
3	antinutritious	cyanogenics
4	degradation	erosive
5	cyanogenics	inhibitor

Table 4.7: Pairs of target terms used for each run

Age		Degree			Experience (years)		
31-40	40+	B.Eng	MSc	PhD	1-9	10-20	20+
4	16	3	5	12	4	10	6

Table 4.8: Expert Data

final assessment of every hypothesis is the average for all the experts who evaluated it.

- In order to reduce the experts' workload, each expert assessed just 5 hypotheses.
- In order to avoid bias to some groups and to ensure a normal distribution of data, the assignment of hypotheses to experts was done in a random way: one hypothesis is randomly assigned to any expert.
- In order for the experts to assess different hypotheses from different runs, each expert examined one hypothesis from each run.

Specifically, the experts were asked to assess the hypotheses in terms of four criteria *Interestingness (INT), Novelty (NOV), Usefulness (USE) and Sensibleness (SEN)*, and one criterion to assess the information contribution (ADD) as stated in research question 3, all in a range between 1 (worst) and 5 (best).

In order to run the experiment, a simple *perl* program was designed to generate web pages containing questionnaire-like forms to have every hypothesis evaluated. The information contained in the page was checked by linguists and computer scientists to

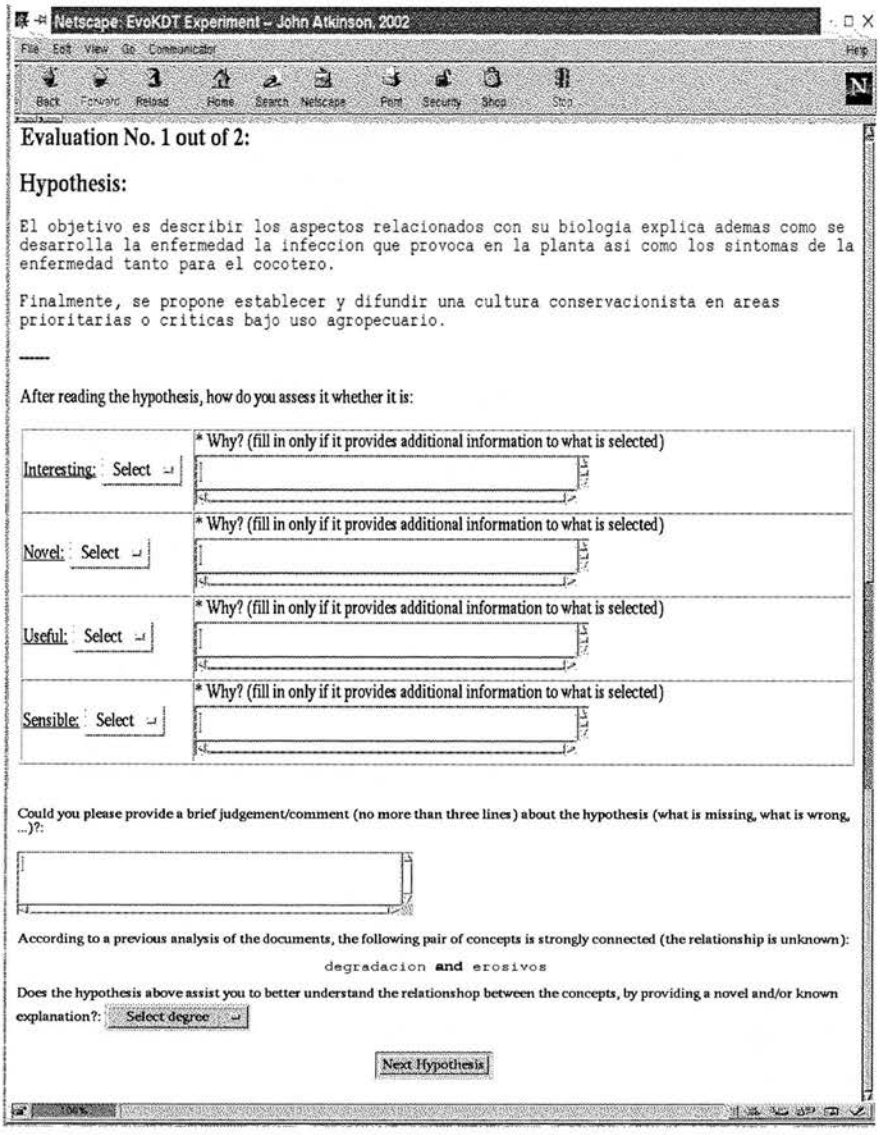

Figure 4.7: A Typical Assessment Page

make sure that the experiment is set in an appropriate way. A typical page has a layout as shown in figure 4.7 in which four sections can be distinguished:

1. The textual description of the hypothesis in an abstract-like form. This is semi-automatically produced in such a way that a hypothesis:

   `hypothesis(<Id>, [r1(p1(a1)),r2(p2(a2)), ..], [rn(pn(an))])`

where r_i, p_i, a_i are the roles, predicates, and arguments respectively can be roughly translated into a text in which the roles correspond to a paragraph like:

```
translation(r1) "is" translation(p1(a1))..
..
"Finally", translation(pn(an)) "are/is obtained..."
```

For example, the text corresponding to the hypothesis:

```
hypothesis(1,[goal(produce("a")), method(analysis("b")),..],
             [conclusion(obtain("c"))]).
```

where *"the goal of this work"*, *"For this, a method based on"*, and *"As a result"* are rough translations for "goal", "method" and "conclusion" respectively, may look like:

```
The goal of this work is to produce "a".
For this, a method based on the analysis of "b" was
carried out. ..
As a result, "c" was obtained...
```

Given that the conversion may not produce well connected text (this is not the purpose of the model), in order to produce an understandable text, it was necessary to make slight changes by hand, for example, putting a better connector between paragraphs, etc.

2. The expert's assessment of the current hypothesis in terms of the four subjective criteria previously mentioned. The degree for each criterion can be selected from 5 ("very high"/best) down to 1 ("low"/worst) [3]. Extra space is also provided for the expert to enter optional comments about his/her assessment in a particular criterion of the hypothesis.

[3] Although the order of the questions might affect the answers, this was not investigated in the current experiment.

3. Expert's general comments (if any) about the hypothesis as a whole.

4. The expert's assessment to whether the novel knowledge contributes to an understanding of the nature of the relationship between the two target terms provided (criterion ADD), ranging from 5 ("*Yes, absolutely*") down to 1 ("*Not at all*").

Overall, the expert's assessment aimed at assessing the real quality of the produced hypotheses by domain experts, having two goals in mind: trying to address research question 2 and 3, and complimenting the investigation on the effectiveness of the search ability of the model.

Some aspects of this evaluation experiment are worth highlighting which constrain the kind of analyses performed in the rest of the current section:

- The model is unique in the way the different strategies and elements, and the automatic evaluation are used. This makes it hard to find systems or approaches to compare with, or even to have "gold standards" available. So the analyses and investigation of the results of the assessment have been performed by keeping in mind a simple purpose: to show the quality of the model and to provide explanations of the possible weaknesses.

- Because of limitations on the number of domain experts it was not possible to deal with different evaluation groups or different levels of tests (i.e., pre and post evaluations). In addition, most of the individuals who promised to take part in the experiment either did not finish it or did not do it at all (despite their willingness) which makes deeper cross analyses unfeasible. As a consequence, the data analysis is carried out on the individuals who managed to finish the evaluation. For those who did not do it completely, we just took the assessments that allowed us to keep the sample balanced. That is, the hypotheses which had a different number of experts' assessments were complimented with those from the individuals who did the evaluation partially.

Keeping this issue in mind, the analysis of the results is performed by first considering the quantitative assessment provided by the experts, by investigating qualitative or implicit aspects that allow us to provide some explanations of the weakness of the outcome, and finally, by doing some partial cross analysis.

4.2.2.1 Quantitative Assessment

Once the system hypotheses were produced, the experts were asked to score them according to the five subjective criteria. The results of this evaluation are shown in figures 4.8 and 4.9, in which the average scores of each criterion for each of the 25 hypotheses are drawn (i.e., average score is taken from different experts assessing each hypothesis). Note that each five-hypothesis group corresponds to one run in increasing order, so hypothesis 10 actually represents the 5^{th} hypothesis of run 2, and so on.

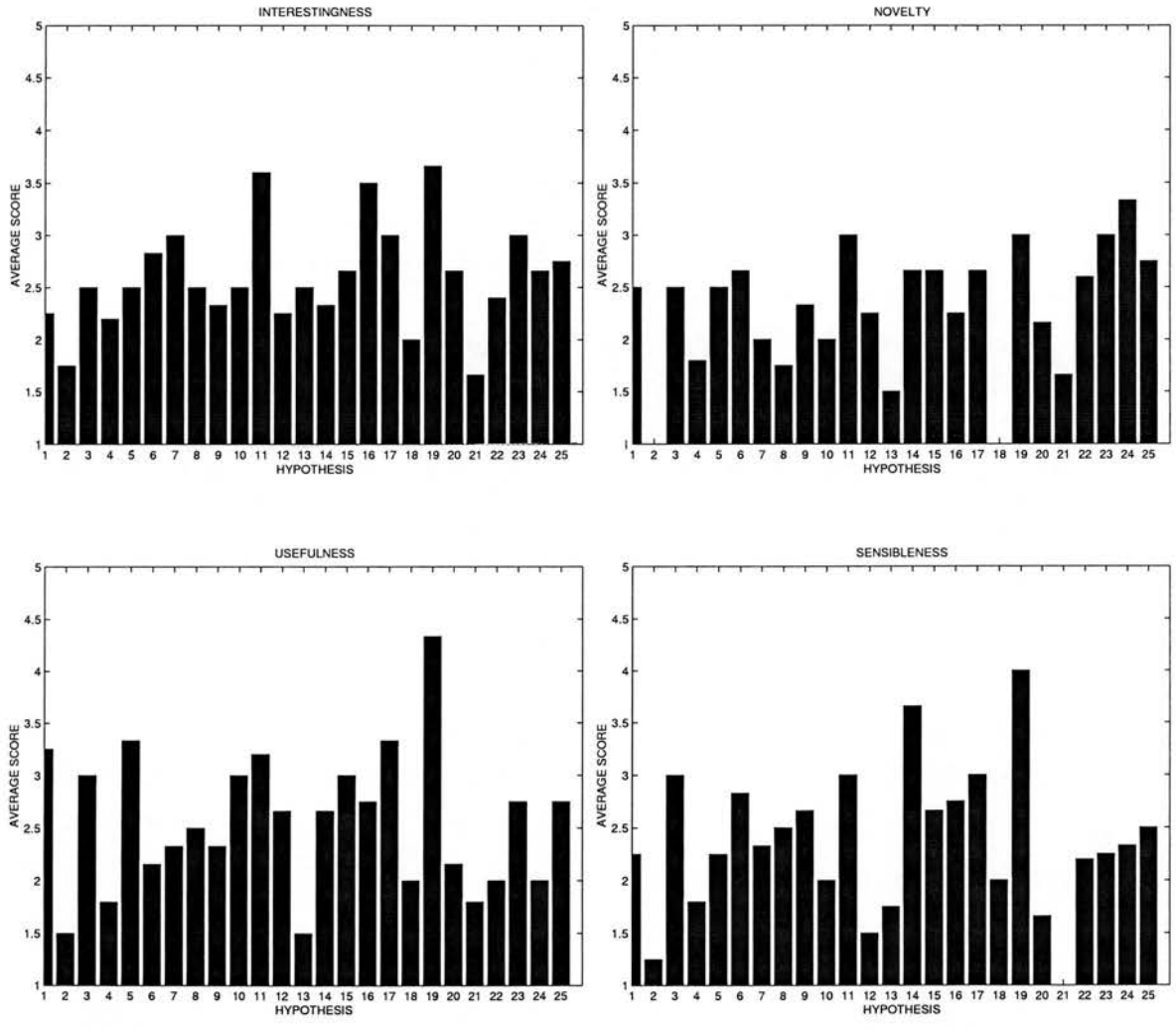

Figure 4.8: Experts Assessment for INT, NOV, USE and SEN

The assessment of individual criteria shows some hypotheses did well with scores

above the average 3 (50%) on a 1-5 scale. This is the case for hypotheses 11, 16 and 19 in terms of INT (hypotheses 7, 17, 23 are just at the average), hypotheses 14 and 19 in terms of SEN (hypotheses 3, 11 and 17 are just at the average), hypotheses 1, 5, 11, 17 and 19 in terms of USE (hypotheses 3, 10 and 15 are just at the average), and hypotheses 24 in terms of NOV (hypotheses 11, 19 and 23 are just at the average).

Criterion	**Num. of Hypotheses**	
	Negative $<$ Average (3)	Positive $\geq$ Average
ADD	20/25 (80%)	5/25 (20 %)
INT	19/25 (76%)	6/25 (24 %)
NOV	21/25 (84%)	4/25 (16 %)
SEN	17/25 (68%)	8/25 (32 %)
USE	20/25 (80%)	5/25 (20 %)

Table 4.9: Distribution of Hypothesis Scores per Criteria

Note also that the assessment seems to be consistent for individual hypotheses across the criteria: hypothesis 19 is well above the average for almost all the criteria (except for NOV), hypothesis 18 always received a score below 2 (25%) except for ADD in which this is slightly higher. Similar situations can be observed for hypotheses 2, 21, etc. The overall distribution of the number of hypotheses which are in this case can be seen in table 4.9.

Although individual hypotheses show very high and very low assessment for such a complex task, it is also important to highlight the overall results. These are summarized in terms of statistical analysis of the whole evaluation in table 4.10. The table indicates that the overall scores are below the average (3) but with large variations for NOV, SEN, and USE.

Although ADD is not an indication of the quality of the hypothesis itself, the previous data still allows us to address question 3 with the scores of figure 4.9 and the statistics of table 4.10. Although below the average, the score for ADD is the highest among the criteria with an average 2.58 (40%) and with the lowest variation of all the

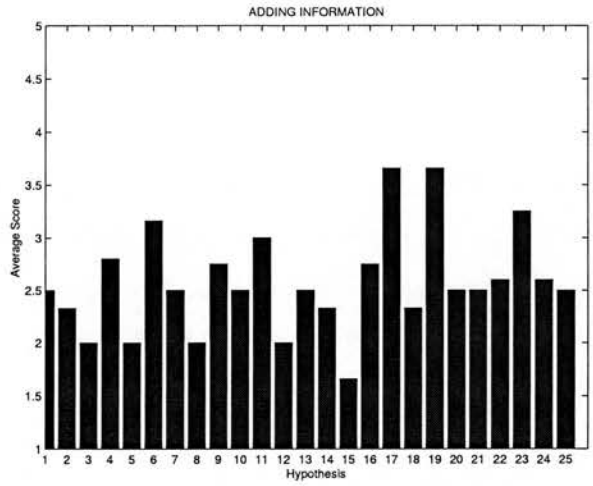

Figure 4.9: Experts Assessment for Hypotheses' Additional Information

Criterion	Mean	Confidence Interval (95%)
ADD	2.60 (40%)	2.60 ± 0.168
INT	2.60 (40%)	2.60 ± 0.173
NOV	2.30 (32%)	2.30 ± 0.205
SEN	2.51 (38%)	2.51 ± 0.237
USE	2.56 (39%)	2.56 ± 0.228

Table 4.10: Statistical Data

criteria (0.168).

Note also that the qualitative explanations previously provided also hold for ADD. Specifically, they suggest that the quality of ADD may depend on how much information contained in one hypothesis is relevant to the target terms, but as dominance conditions must be met, the *relevance* values do not always achieve high values. Regardless of the runs, relevance objective values did not exceed 30% (figure 4.6). This suggests that *relevance* should be given some special status in the optimisation.

Another likely explanation for having below-average scores for ADD is that the target terms do not always appear explicitly in the hypotheses, and as the experts are asked to assess how the hypotheses contribute to understanding of the (unknown) rela-

tionship between the concepts, the criterion is assessed low. In this regard, the evaluation suggests that the experts made little effort to realize that while in most cases the target concept do not appear, some close concepts try to make the hypothesis relevant.

Ultimately, it is up to LSA to judge whether similar information appears in a hypothesis. However, as high similarity does not ensure that the target concepts will appear, more constrained measures should be considered.

In general, although the overall results are below the average, they look promising in that this is a very demanding evaluation in terms of human performance (knowledge discovery). Indeed, it is shown that the model is still capable of producing well assessed hypotheses.

However, low/high scores are not necessarily a sign of failure/sucess when humans are involved in the assessment. For such a complex task, humans may perform poorly when faced with analysing large amounts of text data. For the same reason, humans may not be able to find the hypotheses that the system found. In order to investigate the extent to which the system evaluation is comparable to human judgments, we carried out a correlation analysis between the model evaluation and the expert judgments. The objective here is that if the model's automatic evaluation correlates with human judgments, then the evaluation criteria can be considered successful.

Since both the expert and the system's model evaluated the results considering several criteria, we first performed a normalization aimed at producing a single "quality" value for each hypothesis as follows:

- *For the expert assessment:* the scores of the different criteria for every hypothesis[4], are averaged. Note that this will produce values between 1 and 5, with 5 being the best.

- *For the model evaluation:* for every hypothesis, both the objective values and its fitness are considered as follows: as the values should show the fact that the higher, the better, we subtract the fitness from 1 (the lower the fitness, the better) for that hypothesis and then we add this to the average value of the objective values for this hypothesis. Note that this will produce values between 0 and 2, with 2 being the best.

[4]ADD is not considered here as this does not measure a typical KDD aspect

We calculated the above pair of values for every hypothesis as shown in figure 4.10[5] and obtained a (Spearman) correlation $r = 0.43$ ($t-test = 23.75, df = 24, p < 0.001$). From this result, we see that the model shows a good level of prediction compared to human subjects. Note that in a similar experiment supported by external resources (WordNet), and using simple discovered rules (Basu et al., 2001b), a lower human-system correlation of $r = 0.386$ was obtained. Considering also that the human subjects were not domain experts as in our case, our results are encouraging as these involve a more demanding process which requires understanding of both the hypothesis itself and the working domain.

	Relev	Struc	Coher	Cohesi	Inter	Plaus	Cove	Simp
ADD	-0.06	-0.18	0.14	0.19	0.49	0.28	0	-0.16
INT	0.18	-0.18	0.03	0.033	0.56	0.01	0	-0.09
NOV	0.15	-0.28	0.00	-0.038	0.32	-0.07	0	-0.07
USE	0.27	-0.15	0.21	-0.202	0.59	0.009	0	-0.03
SEN	0.20	-0.26	-0.04	-0.19	0.53	0.14	0	-0.06

Table 4.11: Details of Expert-System Correlations

Since the quality of the hypotheses is measured above by using all the criteria, it does not evaluate particular objectives (in isolation) because of the effect of the trade-offs in the optimization phase. The correlation of these objectives with particular expert's criteria are shown in table 4.11[6]. Some of them are worth highlighting as follows:

- There is a very good correlation $r = 0.56$ ($t-test = 36, df = 24, p < 0.001$) for *interestingness* between the system and the experts. This suggests that the system has a good notion of what the interesting hypotheses are, correlating positively with human judgment. Note that the model is able to achieve this without using any external ontological resource.

[5] In order to allow a better visualisation the system evaluation is scaled up to the range 1 to 5.

[6] Due to dominance conditions, the **coverage** value remains the same for the best hypotheses and so there is no correlation with the different human criteria.

- There is no correlation between the system's *simplicity* and any of the expert's criteria: this suggests that the simplicity of a hypothesis may not be a determining factor for the real novelty or interestingness as assessed by the experts.

- There is a correlation $r = 0.21$ ($t-test = 23.98, df = 24, p < 0.001$) between the system's *coherence* and the experts' assessment of usefulness: as the coherence measures a hypothesis' semantic features, the correlation suggests that for the expert, the hypothesis' comprehensibility and readability may be a key issue in determining its degree of usefulness.

- The system's *Plausibility* is well correlated with the expert's criterion ADD ($r = 0.28, t-test = 18, df = 24, p < 0.001$) and Sensibleness ($r = 0.14, t-test = 18, df = 24, p < 0.001$): the correlation with ADD suggests that the experts may not be considering the explicit explanation of a hypothesis for the relation between the target terms, but the degree of the hypothesis' plausibility. Intuitively, the correlation with sensibleness shows that the hypothesis may need to make sense before analyzing its plausibility.

Although we have not assessed the model in other domains, in which these correlations might vary, it is worth wondering why the assessment for most hypotheses are below the average.

4.2.2.2 Qualitative Analysis

As mentioned in the details of the assessment methodology, the experts had also the opportunity to make (optional) comments about the individual criteria or the hypotheses. We collected and summarized this information to see if it supports different factors (e.g., issues concerning IE, LSA, etc discussed in section 4.1) identified during the experiment which, we believe, may explain at some extent, the low scores in the evaluation. In this closer analysis, the following issues were observed:

1. **Comprehensibility:** some of the text representing the hypotheses' knowledge was not comprehensible to the experts. Some experts suggested improving the realization of the texts as they were unable to assess these properly. Some of

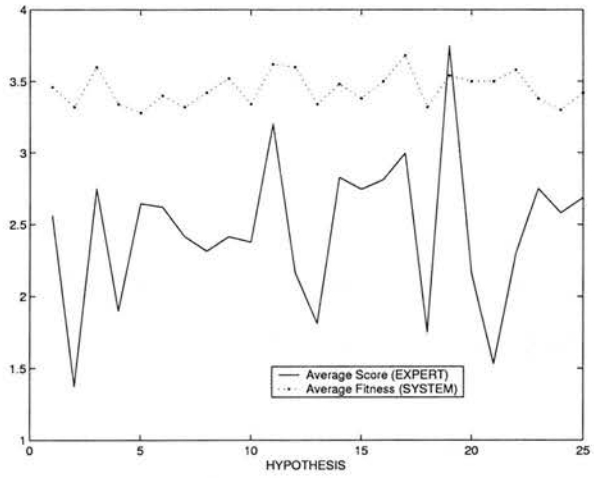

Figure 4.10: Expert Assessment versus Model Evaluation

them argued that the text could not be visualized as a hypothesis, hence they could not figure out its purpose. Representative comments of this kind included:

- The hypothesis is incomprehensible and cannot be understood
- I know about the work, but the text fails to provide the idea in a proper form.
- I think that writing of the text needs to be improved in order to understand the hypothesis.
- I don't see a hypothesis formulation but a story of facts happened.

Likely explanation: Although the model has not aimed at automatically generating the text from the hypothesis to be understood by human evaluators, the evaluation suggests that more effective ways to represent the hypotheses into readable texts should be considered. Sometimes, it turned out to be difficult for the experts to understand the real aim of the experiment in assessing the hypotheses. As a consequence, experts tended to get confused so instead of assessing the communicative goal of the text in terms of the hypothesis, they tried to assess the quality of the text itself, hence the misleading evaluations for some cases.

2. **Inconsistency between paragraphs:** it was noted that for some texts, there was no consistent connection (apparently) between topics of the paragraphs. Despite this, it is claimed that these hypotheses still contain information worth exploring but in putting their parts together, it makes it difficult to figure out the meaning of the hypothesis (this also applies to the connection between the antecedents and consequents). Although useful antecedents in the hypotheses were found, it is noted that in some hypotheses, these have nothing to do with the corresponding conclusions. Representative comments of this kind included:

```
-The second paragraph is out of context.
-The conclusion is unconnected from the statement of the work.
-Definitely the hypothesis sounds useful but  the results seem
 to be addressed in a different direction.
-Little relation between results (digestibity and ecotypes
 performance) which is a intensive agricultural approach, with
 the conservationist topic.
-It is unclear what the relation between the use of ethylene
 and the   digestibity mentioned in the third paragraph is.
```

Likely explanation: in general, relatedness between conditions (e.g., paragraphs) is regarded as a semantic similarity task, specifically evaluated by criteria such as coherence (see chapter 3). However, as we are measuring coherence in terms of LSA-based relatedness, the relation is expressed as a contextual measure, implying that the similarity between concepts depends on the contexts in which these occur. Therefore, the "predictions" made by LSA for the hypotheses are not necessarily correct. For example, suppose that two consecutive conditions contain sentences whose main topics involve `soil` and `cow`, respectively. According to the training information provided from LSA, these two terms are highly similar, whereas for the domain experts, the paragraphs containing the terms above show a completely inconsistent relation between two unconnected topics.

Another key factor that might be influencing this kind of undesirable outcome is the fact that even if the coherence is correctly determined for the hypothesis

(i.e., LSA has sufficient information to make similarity judgments), this does not ensure that those containing higher coherence values become the final good solutions due to the trade off decision between criteria. For example, in run 3 (figure 4.3), there is no hypothesis with a coherence value higher than 0.74 (out of 1), which indicates that these hypotheses are not so good from that point of view, but since that the method has to consider other criteria as well, these solutions (i.e., "combinations" of objectives) are the best ones found so far.

3. **Specificity of topics:** some experts felt unable to assess the hypotheses properly because some specific topics or methods stated in these hypotheses were unknown for them. Representative comments of this kind included:

```
- The topics concerned  tree-like species are unknown to me.
- I have no knowledge to assess the novelty of this hypothesis.
- What is  the forage production for cow ribs?
- I have no idea what "capulin" and "furnival" are.
- The process of "sawing" gets the statement confused.
```

Likely explanation: A straightforward reason for this has to do with the level of expertise of the evaluators. As shown in table 4.8, experts are quite heterogeneous in terms of degrees, ages, and experience. By observing individual details of the evaluators, it is noted that the same hypothesis is assessed by individuals with different experience and education, leading to misleading results as what is quite obvious or well-known for the experienced, is often not for someone less experienced.

In this regard, we believe that this problem also involves idiosyncratic issues, such as cultural aspects and specialization level. While most of the experts have higher degrees in the field, they come from different Spanish-speaking countries where they have been educated with different levels of preparation. In other words, most of their knowledge, expertise and skills in the field depend on their own countries' conditions (i.e., problems or species found in one country- because of its climatic characteristics, do not exist in the others).

4. **Domain specific issues:** it was observed that some specific aspects of the domain influenced the quality of the hypothesis, which, in some cases could not be captured by the model. Representative comments of this kind included:

```
- Physiology and Fertilization should be connected.
- There is confusion about the above-mentioned plant ("Cocotero")
- The second sentence is methodological in nature.
- Determining the quantitative level in plants would compliment
  the degree of the knowledge interestingness.
- Digestibity both "in vitro" and "in situ" form,  is a response
  variable not the cause of the different ages.
- The process is normally performed in animal production.
- Salinity restrict some nutrient's absorption.
- No distance of plantation is specified.
```

   Likely explanation: clearly, the model does not deal with domain specific knowledge beyond what is stated in the documents, so it is difficult for this to provide more precise information. However, some aspects can be due to the lack of a larger training corpus. If an insufficiently large corpus is used, the data analysis performed on the texts might be insufficient to provide adequate lexical knowledge for the similarity judgments. For example, noting the first comment (e.g., "Physiology and .."), the model may fail to have the concepts connected because the training corpus is not sufficient to capture the context features.

5. **Incomplete hypotheses:** some texts are not properly assessed as they were not considered good hypothesis because they missed some key elements. For example, a hypothesis may be composed of only the goal and the results, but additional elements are needed to figure out what the hypothesis tries to state (e.g., methods). Representative comments of this kind included:

```
- More information is needed to assess the novelty.
- The cause-effect connection can not be seen.
- The relation cause-effect is unclear.
```

- The usefulness of the middle sentence cannot be seen.

Likely explanation: while the IE task affects the information (complete or incomplete) that will lie in the hypotheses, the fact that there may be missing or incomplete facts in them may be due to two factors: the training data extracted from the corpus and the GA itself. If some associations between roles or predicates are not significantly represented in the corpus, these will not be captured in the training outcome. Hence the final hypotheses will fail to express the kind of underlying associations referred by the experts (e.g., cause-effect). The optimization strategy of the GA is then responsible for getting hypotheses that contain incomplete structures.

Because of the trade off between criteria (specifically involving the criterion "structure"), higher values for the criterion may not necessarily achieve the final stage. In fact, for some of the runs, the average hypotheses' structure value does not exceed 35% (figure 4.6), which suggests that despite becoming part of the final set of solutions, the individual objective values are not as high as desired but they meet dominance requirements.

4.2.2.3 Examples

In order to show what the final hypotheses look like and how the good characteristics and less desirable features as above are exhibited, we picked some of the average best and worst hypotheses as assessed by the experts (the full set of sample hypotheses can be seen in appendix B). Specifically, the 2 best hypotheses and the 2 worst hypotheses were taken and are highlighted in tables 4.12 (AEA denotes *Average Expert Assessment*) and 4.13.

As we do not have domain knowledge to analyse the content of these selected hypotheses, some general descriptions of the predicates' arguments are provided to give a flavour of the knowledge involved as follows:

- *Content of Hypothesis 19:*

IF work aims at performing the genetic grouping of populations..

Hyp.	Target Terms	AEA	Objective Vector (Rel,Str,Coher,Cohes,Int,Pla,Cov,Simp)
19	degradation erosive	3.74	[0.92,0.09,0.50,0.005,0.7,0.00,0.30,0.25]
11	antinutritious cyanogenics	3.20	[0.29,0.18,0.41,0.030,0.28,0.99,0.30,0.50]
21	cyanogenics inhibitor	1.53	[0.29,0.48,0.49,0.014,0.2,0.00,0.30,0.50]
2	enzyme zinc	1.37	[0.30,0.12,0.47,0.023,008,0.98,0.30,0.50]

Table 4.12: Assessment and Evaluation for some of the best and worst hypotheses

Hyp.	Description IF-part	THEN-part
19	[goal(perform(19311)),goal(analyze(20811)), goal(establish(22911))]	[establish(111)]
11	[goal(present(11511)),method(use(25511))]	[effect(1931,1932)]
21	[object(perform(20611)),object(perform(2631))]	[effect(1931,1932)]
2	[goal(determine(25011)),object(perform(8821))]	[produce(29051)]

Table 4.13: Structure of hypotheses from table 4.12

```
 AND to analyse the vertical integration for elaborating
     Pinus timber...
 AND to establish the setting values in native timbers
THEN the best agricultural use for land lots  of organic
     agriculture must be established....
```

The hypothesis appears to be more relevant and coherent than the rest in table 4.12. However, this is not complete in terms of cause-effect. For instance, the methods are missing.

- *Content of Hypothesis 11:*

```
IF the goal is to present the forest restoration ..
  AND the method is based on the use of micro-environments
      for capturing farm  mice
THEN digestibity "in vitro" should have an effect on the
     bigalta cuttings
```

This hypothesis looks more complete (goal, methods, etc) but is less relevant than the previous hypothesis despite its close coherence. Note also that the plausibility is much higher than for hypothesis 19, but the other criteria seemed to be a key factor for the experts.

- *Content of Hypothesis 21:*

```
IF the object of the work is to perform the analysis of
   the fractioned honey..
 AND to carry out observations for the study of    pinus
     hartwegii
THEN digestibity "in vitro" should have an effect on the
     bigalta  cuttings ..
```

Note that the structure (48%) is better than for hypothesis 11. However, as the hypothesis is not complete, this has been scored less than for the previous one. This might be explained because the difference in structure between object-object and goal-method (see chapter 3, figure 3.3) is not significant and as both hypothesis (11 and 21) become final solutions, the expert scored best those which better explain the facts. Note that as the model relies on the training data, this does not ensure that every hypothesis is complete. In fact, recall from section 4.1.1 that only 26% (out of 326 rules) contain some sort of "method".

- *Content of Hypothesis 2:*

```
IF the goal was to determine the features of stake in
```

```
   raining soils..
AND the absorption of a  dose of furadan into pinus pringlei
    must be performed
THEN the analysis of the heights  must be produced for the
     study and treatment of the cepa ..
```

This hypothesis is more relevant than hypothesis 11 and 21, however this does not have a clear cause-effect connection. Note also that this is more coherent than hypothesis 11 but the latter is preferred because the additional information which is provided (e.g., method) among other factors (11 has better cohesion than 2, etc).

4.3 Summary

In this chapter, we have described the investigation of the ability of our model (via an implemented prototype), in terms of the different proposed mechanisms, to extract information from text documents, to make semantic similarity judgements, and to search for good hypothesis according to a set of novel evaluation criteria. The effectiveness of the knowledge produced by the model was assessed by human experts through a web-based experiment.

The expert assessment showed very high and low scores for some of the KDD criteria for individual hypotheses. However, the overall evaluation is well correlated with that performed by human experts. In order to investigate why some scores are below the average, we also carried out a qualitative analysis that provides some evidence for the likely reasons.

A closer cross analysis between the system's evaluation and the expert assessment shows very promising results in terms of their correlations. This suggests that despite the overall low-average expert assessment, the system and the experts are well correlated, indicating that for such a complex task (knowledge discovery), the model's behavior is not too different from the experts'.

While the strategies used by the model to evaluate the hypotheses do not use explicit domain knowledge, further experiments in other domains will be necessary to

provide better supporting evidence for our strategies' ability to work across domains.

Nevertheless, the obtained results show that some domain-independent aspects could be improved. In particular, the IE task could benefit from trainable mechanisms to extract the information via the IE patterns. This could also handle specific discourse-level techniques to deal with underlying information contained in the texts and currently not detected via deeper semantic analysis and/or anaphora resolution strategies.

The results also show that some limitations come from the way the GA, specifically the optimization phase, deals with the solutions. Because of the trade-off between hypotheses, some well evaluated aspects of the hypotheses are lost as the GA goes on.

Chapter 5

Conclusion and Further Work

5.1 Conclusion

In this thesis, a model for knowledge discovery from texts has been proposed. Here, the process of search for potentially novel knowledge is performed by an evolutionary algorithm which is guided by semantic and rhetorical information so as to create plausible hypotheses.

The initial (input) knowledge for the model involves a set of rules that represent the documents, and training information obtained from these rules and from the whole corpus. The approach assumes that only parts of a document need to be represented so as to capture key facts by making use of the underlying genre. For this, IE patterns have been designed to extract domain-independent but genre-based information (for scientific abstracts) from texts. Training information is extracted from the analysis of associations from this rule set, and from the lexical semantic knowledge provided by *Latent Semantic Analysis* (LSA) by analysing the whole corpus. The structured information contained in the rules is used to compliment the structure-free information provided by LSA so as to make further semantic similarity judgements.

While the training information and the lexical-level data obtained provide useful knowledge of a domain, it is also revealed that there are limitations which make the process of creating full cause-effect hypotheses difficult. For example, the ability of a hypothesis to explain the discovered facts is restricted by the kind of roles found

in the IE phase. Although the IE can be improved so as to have higher recall, by manually examining many abstracts of the corpus, it can be observed that the cause-effect relation often does not exist at all. Instead, the scientific abstracts analysed seem to be only focused on what the authors did, and what they got, in a way that can be far away from being informative.

As part of the content of a hypothesis is measured in terms of what the model learns from the corpus and the rules, the lack of further rhetorical information which explains the cause-effect relation for some hypotheses suggests that the model may not perform much better than what the texts state. However, the evolutionary strategy used enables the model to combine and evaluate different elements within the search space that humans might not be able to do manually in an efficient way. This ability to explore different and multiple solutions even with incomplete information is indeed one of the main assets of the strategy in searching for plausible hypotheses.

In terms of the discovery in the model, this is defined as the search for hypotheses which maximize quality (and plausibility) criteria, hence it is considered a continuous process of multi-criteria optimization. In order to measure this "quality", key evaluation criteria are proposed and designed which use the semantic and rhetorical information obtained from the training phase, the content of the current hypotheses, and the knowledge provided by the corpus itself, to objectively evaluate these hypotheses.

An evolutionary method for multi-objective optimization used in previous research is adapted to handle the multiple evaluation metrics. This adaptation constitutes a plausible and extensible way to deal with the intensive linguistic information, the document representation, and the search strategy itself.

Experiments carried out with a computer prototype of the model show that the technique copes well with the trade off between the multiple criteria. The outcome of these experiments supports the belief that the most plausible hypotheses are not necessarily those having high values for all the evaluation criteria but those that succeed to have a good "combination" of objectives when compared to other hypotheses (Deb, 2001). In fact, given the conditions of dominance that must be met by the discovered hypotheses and the requirement for a certain number of solutions to be obtained, it is not possible for a hypothesis to get high values for all the criteria evaluated because

this would imply that there is (potentially) only one non-dominated hypothesis. In addition, the experiments carried out show that some good objective values are lost because of dominance decisions, so the trading off is not perfect.

It is worth noting that this optimization constitutes a fundamental feature in the model as this worked on a "blind" basis considering that

- Unlike in other multi-objective problems in Data Mining where typical solutions and optimum criteria values are known in advance, there are no target solutions to compare with.

- Given the unique evaluation criteria proposed, it might be very hard to experimentally establish weighting factors that allow for the aggregation of the different criteria in a single evaluation metrics. Indeed, there is no evidence that the relation between them can be linearly defined.

Little effort has been made in previous research on TM/KDT to evaluate the discovered knowledge. Our optimization strategy using these new evaluation criteria constitutes then a good starting point as an approach to KDT which deals with multiple quality objectives.

For the purpose of this thesis, the model along with its underlying strategy of search and optimization has been assessed by human experts in terms of the quality of the discovered knowledge (i.e., explanatory hypotheses). The sucess of the model was shown as its ability to filter good hypotheses rather than finding a whole set of good solutions, and to correlate with human judgments (i.e., for such a complex discovery task and given the amount of information to deal with, a human may not be able to perform much better than the model).

The experimental assessment by domain experts of individual hypotheses generated by our model shows some very good hypotheses and others which were poorly evaluated in terms of their effectiveness. An overall assessment by domain experts shows that the quality for the hypotheses according to subjective criteria such as novelty and interestingness, is well below the average for some of the criteria despite the few good solutions discovered. However, the evaluation is well correlated with human

experts. Among other factors, it has been noted that IE-related issues may have affected the other criteria. However, considering the complexity of the discovery task, the overall results are promising because, unlike other approaches to TM/KDT, our model does not require external resources or domain-specific knowledge, beyond that stated in the corpus.

Overall, this assessment highlights the effectiveness of the model to produce "nuggets" from a KDD perspective, and to produce more useful explanations about isolated discovered terms than simple BOW-based text mining systems.

Compared to Mooney's approach which also takes human evaluation into account, the experiments and results in testing our model reveal some differences worth noting:

- Although human-system correlations for Mooney's experiments ($r = 0.386$ for one of the human evaluators group) and for our model ($r = 0.43$) are encouraging, the approaches are using different levels of structuring and sophistication in the extracted information. As a consequence, while the agreement between different humans might be an issue in this kind of evaluation, the experiments suggest that in Mooney's case, human evaluators do not have too much difficulty in assessing the discovered rules in the form of attributes and values. Whereas for our model, assessing the content of the discovered hypotheses is a more demanding process which requires further understanding of both the hypothesis itself and the working domain. In this context, having a few good hypotheses created by our model still represent an encouraging result considering the complexity of the knowledge discovery task and the sophistication of the discovered relations.

- Since we were using a specific technical domain, the experiments required domain human experts to participate, which made the assessment very demanding for the model. For this, the outcome of the model was compared against individuals having an extensive knowledge and expertise in the field, a feature that the model does not have beyond the implicit knowledge stated in the text documents. Note that in Mooney's approach, the information is extracted from general-purpose documents available from `Amazon` which contain the descriptions of books, and where no specialized human evaluators were required.

- Unlike Mooney's approach, our model does not use any external resources or domain-specific knowledge, hence our model is potentially more domain independent than Mooney's. Although this may restrict the ability of knowledge discovery of the model by having no "previously seen" knowledge, the correlation between the model and the experts showed better results than for Mooney. This suggests that our model may have a better notion of what the quality of the discovered knowledge means when compared to humans. Note also that, unlike Mooney's, our approach deals with quality in a multi criteria way, and the correlation between human judgements and some individual criteria evaluated by the model is good. This indicates that our model has a clearer notion of different issues involved in good-quality knowledge, and so is not restricted to a single notion of "novelty" as in Mooney's.

- Another interesting feature in Mooney's work is that in mining "novel" rules, the learning algorithm uses the attribute values of the rules and not the "roles" (e.g., `name, author, ..`). As a consequence, the mined patterns do not neccesarily contain suprising or unexpected relationships as the values will be attached to fixed roles, which may restrict the discovery process in the whole search space. In our model, all the basic knowledge is used to explore the search space including the rhetorical and semantic information which can lead to unusual common findings. Although there is rhetorical information that is more likely to be associated with certain semantic information than other, our approach manages to make good "combinations" of material extracted from different hypotheses so as to find other relationships different from those initially extracted from the rules.

5.2 Further Issues

In this thesis, we used a specific domain corpus as an example of input for our model of knowledge discovery from texts, although the approach developed here can be applied to other domains. Despite the original domain independence of both the representation scheme and the evaluation criteria, qualitative investigation of the assessment revealed that there were a few issues concerning specific topics which may need to be checked

with at least one more scientific domain.

This thesis also established the fact that the use of information extraction techniques based on a technical/scientific genre can be useful to produce explanatory knowledge. Although the aim of the thesis was not to design a highly accurate IE task for the purpose of the discovery, the results obtained and the investigation carried out on these suggest that improvements on the IE component may assist the model to produce a more effective outcome. The evaluation strategy also raises some issues in terms of the trade off between criteria and the preference of these criteria in finding fit hypotheses. Hence it is important to envisage how we can cope with these limitations.

This section discusses these issues with regards to information extraction, domain independence, evaluation, knowledge representation, and extension of the model to other tasks.

5.2.1 Information Extraction

By observing the results in the experiments of this thesis, it can be seen that key weaknesses come from the IE task performed as a part of the preprocessing stage for the model. Having IE patterns defined by hand constitutes a limitation which may be improved by providing trainable ways to extract the semantic and rhetorical information of interest. Indeed, for summarization tasks, Teufel (1998b) proposes an efficient method to extract sentences containing rhetorical information from full texts. For this, the extraction process is seen as a classification task of segments of rhetorical information identified in the texts against the most likely patterns. Compared with predefined IE patterns, the method above can produce higher recall for the extracted information. The method may also benefit from using shallow parsing techniques so as to analyse in more detail the extracted information in terms of their rhetorical roles and semantic subcategorization terms (i.e., predicate and arguments). However, it must be kept in mind that whatever the deeper analysis is, this should consider the trade off between the granularity of the representation obtained in the extracted information and the requirements that could be demanded by the search/optimization search. In our work, semantic-level information such as a predicate's arguments are still considered a "bag of words" mainly because we wanted to preserve most of their semantic/rhetorical

meaning as further analyses (i.e., genetic operators) are carried out. If a deeper representation is looked for, one should be aware that the evolutionary search strategy may not be able to check and/or validate whether the different combinations through the search space are plausible. In addition, when the human assessment comes to play, it must be noted that the deeper and more sophisticated the representation, the harder the translation into understandable language will be.

While incorporating more sophisticated approaches to extraction and representation may somewhat improve the quality of the results, the thesis also showed by manually examining the corpus of abstracts used, that there is a significant amount of information that is not present. Enhanced extraction methods will still fail to detect this information. It is no surprise that the abstracts represent relatively little knowledge when compared to the full text documents, but at the same time, represent a simpler and more compact source of knowledge which makes it relatively easy to perform text analysis tasks.

5.2.2 Domain Issues

In this thesis, both the document representation and the evaluation criteria for the model's search strategy were proposed without using any domain-specific knowledge or external linguistic resources. Despite there being specific domain knowledge that influenced the quality of the outcome in the experimental analyses, there is not sufficient evidence of whether this is a determining factor or not. Additional assessments using other technical/scientific domains will need to be accomplished to establish the extent to which these issues are latent. These extra experiments may also reveal whether the results obtained in this thesis can be more or less extended to other domains within a technical genre.

Further experiments and the detailed investigation of the outcome across different domains may tell us whether the specific problems are due to imprecision in tasks such as information extraction, or whether they are due to conceptual knowledge aspects which by only using a target corpus, the model will fail to deal with.

Since we are using specialized knowledge based only on what the text documents state, these experiments should also make the distinction clear between the model's

way so as to measure the closeness between the target concepts and the hypotheses being evaluated. Ideally, the evaluation should both provide valuable hypotheses and effectively explain the relationship. The experiments suggest however that this is not always the case: some hypotheses are assessed as very interesting but they do not provide a useful explanation for the unknown relation between the concepts.

A very promising task would be to encourage (with a reinforced semantic similarity mechanism if necessary) a stronger explanation of the relation between target concepts so as to make the model extendible to Question-Answering (QA) systems (Harabagiu et al., 2000; Vlado, 2000). Although current QA systems deal with more diverse and complex questions than that which our model provides, further improvement of the model could potentially contribute as a first step to evolutionary QA. Note that at present the explanation only concerns a general relationship between a pair of target concepts. If different relationships are to be looked for, new evaluations concerning the semantic and rhetorical knowledge, among others, should be considered.

5.2.6 Summary

In this thesis, a new model for evolutionary knowledge discovery from texts has been proposed. The model deals with issues concerning shallow text representation and processing for mining purposes in an integrated way. Its aim is to look for novel and interesting explanatory knowledge across text documents. The approach uses Natural-Language technology and Genetic Algorithms to produce explanatory novel hypotheses. The proposed model is interdisciplinary, involving concepts not only from evolutionary algorithms but also from many kinds of text mining methods. Accordingly, new kinds of genetic operations suitable for text mining have been proposed. The principles behind the representation and a new proposal for using multi-objective evaluation at the semantic level have been described. Some promising results and their assessment by human experts were also discussed which indicate the plausibility of the model for effective KDT.

Appendix A

Information Extraction Patterns

```
# ***************************************************************
# Specification of Basic IE patterns
#
# Notation:
#
#  @TAG: matches any word of POS label TAG
#  @WORD/TAG: matches WORD with POS label TAG
#  $CAT: match grammar category CAT (i.e., noun, verb, etc)
#  ? : matches any word
#  &TERM : translates TERM into a predicate
#         (i.e.,  &TERM[Args] --> TERM(Args) )
#  ^ : matches the beginning of the line
#
# ***************************************************************
#  Roles: GOALS, AIMS, OBJECTS
# ***************************************************************
con @art proposito|objetivo|fin|objeto  de @ACTION/inf OBJ se TASK:
        goal(&ACTION[OBJ]),method(realizar(TASK))
^ para @ACTION/inf OBJ se TASK:
          goal(&ACTION[OBJ]),method(realizar(TASK))
@art ? proposito|objetivo|fin|objeto ? es|fue|era @ACTION/inf OBJ:
```

```
          goal(&ACTION[OBJ])
SRC tiene|tuvo como|por ? objeto|objetivo|fin ? @ACTION/inf OBJ:
          goal(&ACTION[OBJ])
pretende|intenta ? @ACTION/inf OBJ: goal(&ACTION[OBJ])
se estudiaron|estudio|estudian  SUBJ:   object(estudiar(SUBJ))
se discute|discuten|expone|exponen  un|una|el|los|las  SUBJ :
          object(presenta(SUBJ))
se desarrollo|desarrolla|desarrollan @art|@s  SUBJ:
          object(desarrollar(SUBJ))
desarrollar SUBJ:  object(desarrollar(SUBJ))
llevar|llevado|llevo a cabo  SUBJ:  object(desarrollar(SUBJ))
determinar|establecer SUBJ : object(establecer(SUBJ))
describio|describe|describen SUBJ : object(describe(SUBJ))
realizo|realizaron|realizan  @art|@s TASK: object(realizar(TASK))
se muestra|muestran los resultados de TASK: object(realizar(TASK))
# ************************************************************
# Roles:  METHODS/PROCEDURES
# ************************************************************
se  usa|uso|usaron|utilizo @art|@s OBJ: method(utilizar(OBJ))
utilizando @art|@s OBJ: method(utilizar(OBJ))
tecnica|metodo|procedimiento usado|utilizado fue|es|sido @art MTD:
          method(utilizar(MTD))
uso|utilizacion de  MTD:   method(utilizar(MTD))
# ************************************************************
#  Roles: CONCLUSIONS, RESULTS
# ************************************************************
concluir|concluyo|concluye  que ARG1 @ACTION/par|@ACTION/v ARG2:
          conclusion(&ACTION[ARG1,ARG2])
resultados|resultado  establece|sugiere|sugieren  que CONC:
          conclusion(establecer(CONC))
muestra|muestran|mostro|indica|indicaron @art RESULTS:
```

```
            conclusion(producir(RESULTS))
muestra|muestran|mostro|indica|indicaron que ARG1 @ACT/v ARG2:
            conclusion(&ACT[ARG1,ARG2])
senalan|mostro|mostraron RESULTS: conclusion(producir(RESULTS))
encontro que RESULT: conclusion(producir(RESULT))
X afecta|afecto|afectaron Y: conclusion(efecto(X,Y))
la influencia de X @p  Y: conclusion(efecto(X,Y))
correlacion|asociacion|relacion|relaciones entre X y Y:
            conclusion(efecto(X,Y))
X se  relaciona|relaciono con Y: conclusion(relacion(X,Y))
```

Appendix B

Sample Hypotheses

Pairs of target terms used for each run in the experiment:

Run	Term 1	Term 2
1	enzyme	zinc
2	glycocide	inhibitor
3	antinutritious	cyanogenics
4	degradation	erosive
5	cyanogenics	inhibitor

Description of the 25 Best hypotheses produced in the experiment (rough translation from the originals in Spanish):

1. Hypothesis:

```
objectiveVector(36,[0.4628,0.05553,0.6532,0.0227,0,0,0.3,0.25]
hypothesis(36,[object(analiza(8831)),object(determinar(25011)),
    method(utilizar(22011))], [conclusion(encontrar(24511))])
```

   Contents (brief):

   - The resulting performance of the four-year bay timber and its effect on the quality, diameter and cutting pattern is analyzed.

- The object is to determine the size and diameter of stake for high performance in coicote soils and to determine its relation with the forage production in raining soils at Tabasco state.
- White quebracho tree is used in combination with aspen wood.
- Finally, an equation is obtained to estimate the fodder production of cow rib with regression coefficients r*2 of 0.44316 and 0.2952 for the cup size.

2. Hypothesis:

```
objectiveVector(71,[0.303,0.122,0.47,0.023,0.008,0.98,0.3,0.5])
hypothesis(71,[goal(determinar(25011)),object(realizar(8821))],
        [conclusion(producir(29051))])
```

Contents:

- The goal of the work is to determine the size and diameter of stake for high performance in coicote soils and to determine its relation with the forage production in raining soils at Tabasco state.
- The object is the absorption of a dose of furadan into pinus pringlei.
- As a result, an analysis of the heights must be produced for the study and treatment of the cepa in which significant differences were observed for the several varieties.

3. Hypothesis:

```
objectiveVector(9,[0.314874,0.483935,0.484,0.0125,0,0,0.3,0.5])
hypothesis(9,[object(realizar(16321)),object(realizar(8821))],
        [conclusion(encontrar(24511))])
```

Contents:

- The goal is to prepare and to carry out the sawing process from a sample of 64 pieces of woods during the 1981-1982 harvest in the experimental farm of the faculty of agronomy.

- The object is the absorption of a dose of furadan into pinus pringlei.
- Finally, an equation is obtained to estimate the fodder production of cow rib with regression coefficients r*2 of 0.44316 and 0.2952 for the cup size.

4. Hypothesis:

```
objectiveVector(27,[0.2267,0.0399586,0.59,0.0437,0,0,0.3,0.25])
hypothesis(27,[goal(determinar(25011)),method(utilizar(22111)),
       method(utilizar(2111))], [conclusion(describe(311))])
```

Contents:

- The object is to determine the size and diameter of stake for high performance in coicote soils and to determine its relation with the forage production in raining soils at Tabasco state.
- The method uses between 100 to 500 stakes with dimensions lower than 15cm for diameter and 5cm for the height, replacing them at least every two years.
- A cific-brand volumetric universal machine along with an electric scale is used for the experiments.
- As a result, the laboratory breeding of eight species of "coleoptera cerambycidae" is described from diverse larvae states to Imago into two artificial media.

5. Hypothesis:

```
objectiveVector(51,[0,0.483935,0.484488,0.0282159,0,0,0.3,0.5])
hypothesis(51,[object(realizar(16321)),object(realizar(8821))],
       [conclusion(producir(2651))])
```

Contents:

- The goal is to prepare and to carry out the sawing process from a sample of 64 pieces of woods during the 1981-1982 harvest in the experimental farm of the faculty of agronomy.
- The object is the absorption of a dose of furadan into pinus pringlei.
- As a consequence, a higher preference of northern exposition of the infested trees is produced. In addition, it was observed that some lower portions of the infected "fuste" showed a larger amount of built galeries and niches compared to the spatial layout of the atacking adults.

6. Hypothesis:

```
objectiveVector(8,[0,0.483935,0.5,0.0263418,0,0,0.3,0.5])
hypothesis(8,[object(describe(1611)),object(describe(1611))],
        [conclusion(producir(18851))])
```

Contents:

- A new association is described which develops in soils of certain salinity degree in docker areas and in cliffs of the Cantabric seaside poo annuae-spergularietum salinae.
- A new association is described which develops in soils of certain salinity degree in docker areas and in cliffs of the Cantabric seaside poo annuae-spergularietum salinae.
- Finally, sequential applications of amonium nitrate are produced but the concentration of free aminoacids did not increase the concentration of free aminoacids.

7. Hypothesis:

```
objectiveVector(42,[0.620307,0.0967,0.34566,0.029,0,0,0.3,0.5])
hypothesis(42,[method(utilizar(3821)),object(empleadas(16651))],
        [conclusion(efecto(1931,1932))])
```

Contents:

- Arbitrary blocks with three treatments of "raleo" were used and then overcame by pseudostrobus which currently dominates and is still increasing.
- The aim is to use VG at levels of 48% of diets as in the study for lambs on the grow.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

8. Hypothesis:

```
objectiveVector(31,[0.218754,0.0555,0.655,0.0318,0,0,0.3,0.25])
hypothesis(31,[object(realizar(2631)),object(realizar(3421)),
       method(utilizar(23011))],[conclusion(efecto(1931,1932))])
```

Contents:

- The object of the work is to carry out observations for the study of pinus hartwegii at the maxican snowed hills aimed at complement the previusly existing information about the development states of Adjunctus and its biology.
- The object is to perform the basic study of the resources soil, vegetation, and land water as well as the socioeconomic conditions of the inhabitants in the region.
- The method uses coropletics thematic maps.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

9. Hypothesis:

```
objectiveVector(85,[0.291672,0.096738,0.482,0.036,0,0,0.3,0.5])
hypothesis(85,[method(utilizar(3821)),object(realizar(1921))],
       [conclusion(efecto(1931,1932))])
```

Contents:

- Arbitrary blocks with three treatments of "raleo" were used and then overcame by pseudostrobus which currently dominates and is still increasing.
- The work aims to perform the growing at the postgraduate school in a raining soil of clayey nature.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

10. Hypothesis:

```
objectiveVector(99,[0.620307,0.483,0.4899,0.0073, 0,0,0.3,0.5])
hypothesis(99,[object(establece(24821)),object(empleadas(16651))],
       [conclusion(efecto(1931,1932))])
```

Contents:

- The research's object is to establish the changes in the seeds' permeability of one leucocephala "huaxin" which allows it to promote its germination.
- The VG is used at levels of 48% of diets as in the study for lambs on the grow.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

11. Hypothesis:

```
objectiveVector(88,[0.29,0.185,0.417,0.029,0.27,0.997,0.3,0.5])
hypothesis(88,[goal(presenta(11511)),method(utilizar(25511))],
        [conclusion(efecto(1931,1932))])
```

Contents:

- The goal is to present a two-dimensional scheme for forest restoration in which two regression with Pinus and without Pinus are identified by inspiring in the natural restoring dynamics.
- The method is based on the use of micro-environments for capturing farm mice apodemus sylvaticus and the use of capture traps at a ration of 1464 traps per night.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

12. Hypothesis:

```
objectiveVector(32,[0.291,0.0627,0.4701,0.0361558,0,0,0.3,0.5])
hypothesis(32,[object(realizar(10811)),method(utilizar(25511))],
        [conclusion(efecto(1931,1932))])
```

Contents:

- In this work an analysis of the "hayedos basofilos cantabricos" in 1995 was carried out to determine the existence of two sub-alliances scillo-fagenion oberdorfer and cephalanthero-fagenion TX.
- The method is based on the use of micro-environments for capturing farm mice apodemus sylvaticus and the use of capture traps at a ration of 1464 traps per night.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

13. Hypothesis:

```
objectiveVector(60,[0.317327,0.48393,0.370,0.0213,0,0,0.3,0.5])
hypothesis(60,[object(efecto(16661,16662)),
      object(establecer(25141))],[conclusion(producir(13121))])
```

Contents:

- The aim of the work is to measure the effect of the fodder treatment compared to diets with higher levels as those used in the study of cows on the grow that allows higher consumption of MS (24%).
- The objective is to establish the optimum distance between plants for the production of fodder of interest in 1992 at the council of Huimanguillo Tabasco.
- The work concludes that a high inheritance is produced in terms of the plants' heights. Hence it is possible to get varieties with low contents of rubber and high-performance seed.

14. Hypothesis:

```
objectiveVector(11,[0.620,0.0627,0.3068,0.0069,0,0.995,0.3,0.5])
hypothesis(11,[object(describe(26211)),method(empleadas(16651))],
       [conclusion(efecto(1931,1932))])
```

Contents:

- The object is to describe how to determine the levels of ba, ca, fe, and mg.
- The VG is used at levels of 48% of diets as in the study for lambs on the grow.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

15. Hypothesis:

```
objectiveVector(45,[0.9376,0.06273,0.4940,0.001540,0,0,0.3,0.5])
hypothesis(45,[object(efecto(16661,16662)),
        method(empleadas(16651))],[conclusion(efecto(1931,1932))])
```

Contents:

- The aim of the work is to measure the effect of the fodder treatment compared to diets with higher levels as those used in the study of cows on the grow that allows higher consumption of MS (24%).
- The VG is used at levels of 48% of diets as in the study for lambs on the grow.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

16. Hypothesis:

```
objectiveVector(85,[0.23083,0.1344,0.5856,0.02708,0,0,0.3,0.25])
hypothesis(85,[goal(estudiar(18911)),method(utilizar(14411)),
        object(realizar(31121))], [conclusion(establecer(111))])
```

Contents:

- The goal of the work is to study the growing at the postgraduate school in a raining soil of clayey nature and the harvest of the "capulin" fruit.
- A furnival index is employed for the distribution of sewage coefficients to be used in selecting the best adjusted models.
- The object is to carry out a fertilization with 25 and 50 kg-ha which are suggested to be appropriate for native grasses.

- Finally, the best agricultural use of for land lots of organic agriculture must be established to promote a conservationistic culture in priority or critical areas of agriculture use.

17. Hypothesis:

```
objectiveVector(97,[0.2308,0.04551,0.6303,0.03068,0,1,0.3,0.25])
hypothesis(97,[object(realizar(3411)),method(utilizar(14411)),
        object(realizar(29921))],[conclusion(establecer(111))])
```

Contents:

- The object is to carry out a basic study of different resources including soil, vegetation and water in the lands as well as socieconomic conditions of the inhabitants of the area.
- A furnival index is employed for the distribution of sewage coefficients to be used in selecting the best adjusted models.
- A biomass study was performed to determine its relationship with the environment, the growing and the biomass production and to obtain its correlation with light amd temperature.
- Finally, the best agricultural use of for land lots of organic agriculture must be established to promote a conservationistic culture in priority or critical areas of agriculture use.

18. Hypothesis:

```
objectiveVector(81,[0.307803,0.4839,0.4799,0.01165,0,1,0.3,0.5])
hypothesis(81,[object(realizar(12121)),object(describe(2821))],
        [conclusion(establecer(111))])
```

Contents:

- The object of this work is to carry out the sampling of 26 timber specimen between April and December at the "Encinar" located in the council premises of Almadrones y Mandayona.

- Different aspects related to the timber and its biology are described so to explain how the disease is developed in the plants along with its syhtoms for the "cocotero".
- Finally, the best agricultural use of for land lots of organic agriculture must be established to promote a conservationistic culture in priority or critical areas of agriculture use.

19. Hypothesis:

```
objectiveVector(65,[0.923,0.09673,0.5,0.00577338,0.7,0,0.3,0.5])
hypothesis(goal(realizar(19311)),goal(analisis(20811)),
        goal(establecer(22911))],[conclusion(establecer(111))])
```

Contents:

- The work aims at performing the genetic grouping of populations aimed to show a tendency to the separation of the northern populations into different classes.
- The goal is to analyse the vertical integration for elaborating and selling Pinus timber at the Andes-Patagonia region.
- The goal is to establish the setting values in native timbers and exotic woods currently in use in the construction industry.
- Finally, the best agricultural use of for land lots of organic agriculture must be established to promote a conservationistic culture in priority or critical areas of agriculture use.

20. Hypothesis:

```
objectiveVector(96,[0.92341,0.483935,0.5,0.0115463,0,0,0.3,0.5])
hypothesis(96,[object(establecer(111)),object(establecer(111))],
        [conclusion(establecer(111))])
```

Contents:

- The object is to establish the best agricultural use of for land lots of organic agriculture to promote a conservationistic culture in priority or critical areas of agriculture use.
- The object is to establish the best agricultural use of for land lots of organic agriculture to promote a conservationistic culture in priority or critical areas of agriculture use.
- Finally, the best agricultural use of for land lots of organic agriculture must be established to promote a conservationistic culture in priority or critical areas of agriculture use.

21. Hypothesis:

```
objectiveVector(52,[0.2916,0.4839,0.4913,0.01413,0.2,0,0.3,0.5])
hypothesis(52,[object(realizar(20611)),object(realizar(2631))],
        [conclusion(efecto(1931,1932))])
```

Contents:

- The object of the work is to perform the analysis of the fractioned honey in Brazil aimed at improving the producers' income and profitability.
- The object of the work is to carry out observations for the study of pinus hartwegii at the maxican snowed hills aimed at complement the previusly existing information about the development states of Adjunctus and its biology.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

22. Hypothesis:

```
objectiveVector(43,[0.6203,0.06273,0.3053,0.0078383,0,0,0.3,0.5])
hypothesis(43,[object(realizar(19311)),method(empleadas(16651))],
        [conclusion(efecto(1931,1932))])
```

Contents:

- The work aims at performing the genetic grouping of populations aimed to show a tendency to the separation of the northern populations into different classes.
- The VG is used at levels of 48% of diets as in the study for lambs on the grow.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

23. Hypothesis:

```
objectiveVector(37,[0.291,0.0627,0.4955,0.0298,0,0.9837,0.3,0.5])
hypothesis(37,[object(utilizar(711)),method(utilizar(17511))],
       [conclusion(efecto(1931,1932))])
```

Contents:

- The work aims at using "dibromuro de etileno" and some anastrepha-genre species due to quarantine-related restrictions in the USA and the regulation on the use of crop-spraying such as EDB.
- The method uses Chlorine-based products for the whitening.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

24. Hypothesis:

```
objectiveVector(45,[0.21874,0.05553, 0.5725,0.0314,0,0,0.3,0.25])
hypothesis(45,[object(establecer(111)),object(realizar(3421)),
       method(utilizar(28221))],[conclusion(efecto(1931,1932))])
```

Contents:

- The object is to establish the best agricultural use of for land lots of organic agriculture to promote a conservationistic culture in priority or critical areas of agriculture use.
- The object is to perform the basic study of the resources soil, vegetation, and land water as well as the socioeconomic conditions of the inhabitants in the region.
- A timekeeping-based method is used for each work sequence to calculate the sample size and to obtain additional data such as type of machinery, protection devices, number of workersi, etc.
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

25. Hypothesis:

```
objectiveVector(13,[0.6089,0.4839,0.402547,0.007313,0,0,0.3,0.5])
hypothesis(13,[object(establecer(26911)),object(efecto(16661,16662))],
       [conclusion(efecto(1931,1932))])
```

Contents:

- The goal is to determine the inhibitory ability of the "taninos" extracted from broad bean seeds on the alfa-milasa and tripsina and to establish the effect of different levels of taninos in the portion on the growth of chickens.
- The object is is to measure the effect of the fodder treatment compared to diets with higher levels as those used in the study of cows on the grow that allows higher consumption of MS (24%).
- Digestibity "in vitro" of three cutting ages in six ecotypes has an effect on the bigalta cuttings which got its higher performance in the cutting at 63 days.

Bibliography

Abrams, W. (2002). Text Mining: The Next Gold Rush. *Second Moment: The News and Business Resource for Applied Analytics.* `http://www.secondmoment.org/articles/textmining.php`.

Appelt, D. and Israel, D. (1999). Introduction to Information Extraction Technology. *Tutorial for IJCAI-99, Stockholm, Sweden.*

Banzhaf, W., Nordin, P., Keller, R., and Francone, F. (2000). *Genetic Programming: An Introduction.* Morgan Kaufmann.

Basu, S., Mooney, R., Pasupuleti, K., and Ghosh, J. (2001a). Evaluating the Novelty of Text-mined Rules Using Lexical Knowledge. *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD-2001), San Francisco*, pages 233–238.

Basu, S., Mooney, R., Pasupuleti, K., and Ghosh, J. (2001b). Using Lexical Knowledge to Evaluate the Novelty of Rules Mined from Text. *Proceedings of NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg*, pages 144–149.

Bellegarda, J. (2000). Exploiting Latent Semantic Information in Statistical Language Modeling. *Proceedings of the IEEE*, 88(8):1279–1296.

Bergstron, A., Jaksetic, P., and Nordin, P. (2000). Acquiring Textual Relations Automatically on the Web Using Genetic Programming. *EuroGP 2000, Edinburgh, Scotland*, pages 237–246.

Berry, M. and Browne, M. (2001). *Understanding Search Engines: Mathematical Modeling and Text Retrieval.* SIAM Book series: Software, Environments, and Tools.

Berthold, M. and Hand, D. (2000). *Intelligent Data Analysis.* Springer.

Bratko, I. and Muggleton, S. (1995). Applications of Inductive Logic Programming. *Communications of the ACM*, 38(11):65–70.

Brill, E. (1997). Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. In *Natural Language Processing Using Very Large Corpora.* Academic Press.

Califf, M. (1998). Relational learning techniques for natural language information extraction. Technical Report AI98-276, University of Texas, USA.

Ciravegna, F. and Cancedda, N. (1995). Integrating Shallow and Linguistic Techniques for Information Extraction from Text. *Lecture Notes in Artificial Intellligence 992*, pages 127–138.

Coello, C. (2000). An Updated Survey of GA-based Multiobjective Optimisation Techniques. *ACM Computing Surveys*, 32(2):109–143.

Coello, C. (2001). A Short Tutorial on Evolutionary Multiobjective Optimisation. *First International Conference on Evolutionary Multi-Criterion Optimization, Springer, Lecture Notes in Computer Science No. 1993*, pages 21–40.

Cover, T. and Thomas, J. (1991). *Elements of Information theory.* John Wiley and Sons.

Deb, K. (2001). *Multi-objective Optimization Using Evolutionary Algorithms.* Wiley.

Dijk, T. V. and Kintsch, W. (1983). *Strategies of Discourse Comprehension.* Academic Press.

Ding, J., Berleant, D., and Nettleton, D. (2002). Mining MEDLINE: Abstract, Sentences or Phrases? *Pacific Symposium on Biocomputing, Lihue, Hawaii, USA*, pages 326–337.

Fayyad, U., Piatesky-Shapiro, G., and Smith, P. (1996). From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–36. MIT Press.

Feldman, R. (1998a). Knowledge Management: A Text Mining Approach. *Proc. of the 2nd Int. Conference on Practical Aspects of Knowledge Management (PAKM98), Basel, Switzerland*, pages 1–10.

Feldman, R. (1998b). Text Mining at the Term Level. *Lecture Notes in Artificial Intellligence 1510*, pages 65–73.

Feldman, R. and Dagan, I. (1995). Knowledge Discovery in textual databases (KDT). *Proceedings of the first international conference on knowledge discovery and data mining (KDD-95), Montreal, Canada*, pages 112–117.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Finn, R. (1998). Program Uncovers Hidden Connections in the Literature. *The Scientist*, 10(12):12–13.

Foltz, P., Kintsch, W., and Landauer, T. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse processes*, 25(2):259–284.

Fonseca, C. and Fleming, P. (1995). An Overview of Evolutionary Algorithms in Multiobjective Optimisation. *Evolutionary Computation*, 3(1):1–16.

Freitas, A. (1997). A Genetic Programming Framework for Two Data Mining Tasks: Classification and Generalized Rule Induction. *Genetic Programming 1997, Stanford, USA*, pages 96–101.

Freitas, A. (1998). On Objective Measures of Rule Surprisingness. *Proceedings of the 2nd European Symposium on Knowledge Discovery on Databases, Nantes, France*, pages 1–9.

Freitas, A. (2001a). A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery. In *Advances in Evolutionary Computation*. Springer.

Freitas, A. (2001b). Evolutionary computation. In *Handbook of Data Mining and Knowledge Discovery*. Oxford University Press.

Gaines, B. (1996). Transforming Rules and Trees into Comprehensible Knowledge Structures. In *Advances in Knowledge Discovery and data Mining*, pages 205–226. MIT Press.

Gaizauskas, R. and Wilks, Y. (1997). Information Extraction: Beyond Document Retrieval. Technical Report CS-97-10, Department of Computer Science, University of Sheffield.

Ganter, B. and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*. Springer.

Godin, R., Mineau, G., and Missaoui, R. (1995). Incremental Structuring of Knowledge Bases. *Proceedings of the International Knowledge Retrieval, Use and Storage for Efficiency Symposium (KRUSE-95), Santa Cruz, USA*, pages 179–198.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley.

Graesser, A., Wiemer-Hastings, P., and Kreuz, R. (1999). AutoTutor: A Simulation of a Human Tutor. *Journal of Cognitive Systems Research*, 1:35–51.

Grishman, R. (1997). Materials for Information Extraction: Techniques and Challenges. *International Summer School SCIE-97*.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan-Kaufmann.

Harabagiu, S. and Moldovan, D. (1998). Knowledge processing on an extended wordnet. In *WordNet: An Electronic Lexical Database*, pages 379–403. MIT Press.

Harabagiu, S. and Moldovan, D. (1999). Enriching WordNet texonomy with Contextual Knowledge Acquired from Text. In Swanska, L. and Shapiro, S., editors, *Natural Language Processing and Knowledge Representation*, pages 301–333. MIT Press.

Harabagiu, S., Pasca, M., and Maiorano, S. (2000). Experiments with Open-Domain Textual Question Answering. *Proceedings of COLING-2000, Saarbruken, Germany*, pages 292–298.

Hartley, J. and Benjamin, M. (1998). An Evaluation of Structured Abstracts in Journals Published by the British Psychological Society. *The British Journal of Educational Psychology*, 3(68):443–456.

Hearst, M. (1997). Distinguishing between Web Data Mining and Information Access: Position Statement. *Panel on Web Data Mining KDD-97, Newport Beach, California, USA*.

Hearst, M. (1998). Automated Discovery of WordNet Relations. In *WordNet: An Electronic Lexical Database*, pages 131–151. MIT Press.

Hearst, M. (1999). Untangling Text Data Mining. *Proceedings of the 37th Annual Meeting of the ACL, University of Maryland, USA (invited paper)*.

Hearst, M. (2000). Text Mining Tools: Instruments for Scientific Discovery. *IMA Text Mining Workshop, USA*.

Hilderman, R. (1999). Heuristics for Ranking the Interestingness of Discovered Knowledge. *3rd Pacific-Asia Conference, PAKDD-99, Beijing, China, April*, pages 204–209.

Holland, J. (1992). *Adaptation in Natural and Artificial Systems*. MIT press.

Humphreys, K. and Gaizauskas, R. (2000). LaSIE Technical Specifications. Technical Report CS-00-09, Department of Computer Science, University of Sheffield.

Jacquemin, C. (1999). Syntagmatic and Paradigmatic Representation of Terms Variation. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics , University of Maryland, USA*, pages 341–348.

Jacquemin, C. (2001). *Spotting and Discovering Terms through NLP*. MIT Press.

Jacquemin, C. and Tzoukermann, E. (1999). NLP for Term Variant Extraction: Synergy between Morphology, Lexicon, and Syntax. In *Natural Language Information Retrieval*. Kluwer Academic.

Jaroszeqicz, S. and Simovici, D. (2001). A General Measure of Rule Interestingness. *Principles of Data Mining and Knowledge Discovery, Freiburg, Germany*, pages 253–265.

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall.

Kando, N. (1999). Text Structure Analysis as a Tool to Make Retrieved Documents Usable. *The Fourth International Workshop on Information Retrieval with Asian Languages, IRAL 99, Taiwan*, pages 126–135.

Kintsch, E. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Kintsch, E. and D. Steinhart, G. S. (2000). Developing Summarization Skills through the use of LSA-based Feedback. *Interactive Learning Environments: An International Journal*, 8(2):87–109.

Kintsch, W. (2001). Predication. *Cognitive Science*, 25(2):173–202.

Klavans, J. and Resnik, P. (1996). *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press.

Klebanov, B. (2001). Using Lantent Semantic Analysis for Pronominal Anaphora Resolution. MSc thesis, School of Cognitive Science, Division of Informatics, University of Edinburgh.

Klementinen, M. (1994). Finding Interesting Rules from Large Sets of Discovered Association Rules. *Proc. Third International Conference on Information and Knowledge Management, New York*, pages 401–407.

Knight, K. (1999). Mining Online Text. *Communications of the ACM*, 42(11):58–61.

Kodratoff, Y. (2000). Applying Data Mining Techniques in Text Analysis. Technical Report unpublished, Laboratoire de Recherche en Informatique (LRI), Inference and Learning Group, Universite Paris Sud, France.

Kourie, D. and Oosthuizen, G. (1998). Lattices in Machine Learning: Complexity Issues. *Acta Informatica*, 35(4):269–292.

Koza, J. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press.

Lamirel, J. and Toussaint, Y. (2000). Combining Symbolic and Numeric Techniques for DL Contents Classification and Analysis. *Proceeding of DELOS-2001, first workshop on information seeking, searching and querying in digital libraries, Zurich, Switzerland*, pages 253–258.

Landauer, T., Foltz, P., and Laham, D. (1998a). An Introduction to Latent Semantic Analysis. *Discourse Processes*, 10(25):259–284.

Landauer, T., Laham, D., and Foltz, P. (1998b). Learning Human-like Knowledge by Singular Value Descomposition: A Progress Report. In Jordan, M., Kearns, M., and Solla, S., editors, *Advances in Neural Information Processing Systems*, volume 10, pages 45–51. MIT Press.

Langley, P. (1987). Towards an Integrated Discovery System. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence, Milan, Italy*, pages 198–200.

Lee, T. (2000). An Introduction to Coding Theory and the Two-part Minimum Description Length Principle. Technical Report 2000/8, Department of Statistics, Colorado State University.

Liu, B., Hsu, W., and Chen, S. (2000). Analyzing the Subjective Interestingness of Association Rules. *IEEE Intelligent Systems*, 15(5):47–54.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Manzur, R., Orellana, A., and Cachi, M. (1998). Aplicacion del Método Científico en el Análisis de Resumenes de Trabajos Científicos. *Revista de la Federacion Argentina de Cardiología*, 27(2):221–227.

Michalewicz, Z. and Fogel, D. (1999). *How to Solve It: Modern Heuristics*. Springer.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill, New York.

Mitchell, T. (1999). Machine Learning and Data Mining. *Communications of the ACM*, 42(11):30–36.

Mitra, S. (2002). Data Mining in Soft Computing Framework: A Survey. *IEEE Transactions on Neural Networks*, 13(1):3–14.

Moens, M. and de Busser, R. (2001). Information Extraction: Current Technologies and Promising Research Directions. *Internal Report TR-IE-1, Leuven, ICRI, Belgium.*

Moldovan, D. (2000). Domain-Specific Knowledge Acquisition from Text. *Proceedings of the ANLP-NAACL, Seattle, Washington (USA)*, pages 268–275.

Morin, E. and Jacquemin, C. (1999). Projecting Corpus-based Semantic Links on a Thesaurus. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics , Maryland, USA*, pages 389–396.

Muggleton, S. (1999). Inductive Logic Programming: Issues, Results and the Challenge of Learning Language in Logic. *Artificial Intelligence*, 114(1-2):283–296.

Mugnier, M. (2000). Knowledge Representation and Reasoning Based on Graph Homomorphism. *Proceedings of ICCS-2000, Darmstadt, Germany*, pages 172–192.

Muller, C. (1997). Acquisition et Structuration des Connaissances en Corpus: Elements Methodologiques. Rapport de Recherche 3198, INRIA-Lorraine, Nancy, France.

Nahm, U. and Mooney, R. (2000a). A Mutually Beneficial Integration of Data Mining and Information Extraction. *Proceedings of the 17th National Conference on Artificial Intelligence, Austin, Texas*, pages 627–632.

Nahm, U. and Mooney, R. (2000b). Using Information Extraction to Aid the Discovery of Prediction Rules from Text. *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining, Boston, USA*, pages 51–58.

Nahm, U. and Mooney, R. (2002). Text Mining with Information Extraction. *AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, Stanford, USA*, pages 60–67.

Noda, E., Freitas, A., and Lopes, H. (1999). Discovering Interesting Prediction Rules with a Genetic Algorithm. *Proceedings of the Congress on Evolutionary Computation (CEC-99), Washington*, pages 1322–1329.

Padmanabham, B. and Tuzhilin, A. (1998). A Belief-driven Method for Discovering Unexpected Patterns. *Proc. ACM International Conference on Knowledge Discovery and Data Mining, New York*, pages 94–100.

Pazzani, M. (2000). Knowledge Discovery from Data? *IEEE Inteligent Systems*, 15(2):10–13.

Pazzani, M., Mani, S., and Shankle, W. (1997). Comprehensible Knowledge-Discovery in Databases. *Proceedings 19th Annual Conference of the Cognitive Science Society, Stanford University, USA*, pages 596–601.

Piatesky-Shapiro, G. and Matheus, C. (1994). The Interestingness of Deviations. *Proc. ACM International Conference on Knowledge Discovery and Data Mining, New York*, pages 25–36.

Polanco, X. and Francois, C. (1998). Informetrics and Knowledge Engineering: Data Mining and Information Analysis Aimed at Discovering Knowledge. *Scientometrics*, 41(1):69–82.

Radcliffe, N. and Surry, P. (1994). Co-operation through Hierarchical Competition in Genetic Data Mining. Technical Report EPCC-TR94-09, University of Edinburgh.

Rajman, M. and Besancon, R. (1998). Text Mining: Knowledge Extraction from Unstructured Textual Data. *6th Conference of the International federation of classification societies (IFCS-98), Rome, Italy*, pages 478–480.

Reinartz, T. (1998). Focusing Solutions for Data Mining: Analytical Studies and Experimental Results in Real-World Domains. *Lecture Notes in Artificial Intellligence 1623*.

Riloff, E. and Lorenzen, J. (1999). Extraction-based Text Categorization: Generating Domain-specific Role Relationship Automatically. In *Natural Language Information Retrieval*, pages 20–35. Kluwer Academic.

Sahami, M. (1995). Learning Classification Rules using Lattices. *Proceedings of the Eight European Conference on Machine Learning, Crete, Greece. Lecture Notes in Computer Science 912*, pages 343–351.

Schneider, T. (2000). Information Theory Primer. *Personal Communication.*

Silbershatz, A. and Tuzhilin, A. (1996). What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Transactions on Knowledge and Data Engineering*, 6(8):970–974.

Stadler, W. (1988). *Fundamentals of Multicriteria Optimization*. Plenum Press, New York.

Swanson, D. (1988). Migraine and Magnesium: Eleven Neglected Connections. *Perspectives in Biology and Medicine*, 4(31):526–557.

Swanson, D. (2001). On the Fragmentation of Knowledge, the Connection Explosion, and Assembling Other People's ideas. *Annual Meeting of the American Society for Information Science and Technology, Bulletin of ASIST*, 27(3):12–14.

Teufel, S. (1998). Meta-Discourse Markers and Problem-structuring in Scientific Articles. *Workshop on Discourse Structure and Discourse Markers, ACL 1998, Montreal*, pages 43–49.

Teufel, S. and Moens, M. (1998). Sentence Extraction as a Classification Task. *AAAI Spring Symposium on Intelligent Text Summarization, Stanford, USA*, pages 16–16.

Toussaint, Y. and Simon, A. (2000). Building and Interpreting Term Dependencies using Association Rules Extracted from Galois Lattices. *Proceeding of RIAO-2000, Content-based Multimedia Information Access, Paris, France*, pages 80–85.

Valtchev, P. and Missaoui, R. (2001). Building Concept (Galois) Lattices from Parts: Generalizing the Incremental Methods. *Lecture Notes in Artificial Intellligence 2120*, pages 290–303.

Van der Tol, M. (1998). The Abstract as an Orientation Tool in Modular Electronic Articles. *First International Conference on Document Design, Tilburg*, pages 25–30.

Vlado, K. (2000). TREC-9 Question Answering: Some General and Some Specific Issues. Technical report, Dalhousie Faculty of Computer Science,Waterloo, Canada.

Weaver, W. and Shannon, C. (1963). *A Mathematical Theory of Communication*. University of Illinois Press (republished).

Whitley, D. (1989). The genitor algorithm and selective pressure: Why rank-based allocation of reproductive trials is best. *Proc. 3th International Conf. on Genetic Algorithms*, 4:116–121.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, 4:65–85.

Wiemer-Hastings, P. (1999). How Latent is Latent Analysis. *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence, San Francisco, USA*, pages 932–937.

Wiemer-Hastings, P. (2000). Adding Syntactic Information to LSA. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society, University of Pennsylvania, USA*, pages 989–993.

Wiemer-Hastings, P. and Zipitria, I. (2001). Rules for Syntax, Vectors for Semantics. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society, Edinburgh, Scotland*, pages 1140–1145.

Wilks, Y. and Catizone, R. (1999). Can we make Information Extraction more Adaptive? Technical Report CS-00-02, University of Sheffield, Computer Science Dept.

Wille, R. (2001). Why can Concept Lattice Support Knowledge Discovery in Databases. *International Workshop on Concept Lattice-based theory, methods and tools for Knowledge Discovery in Databases, Stanford University, USA*, pages 22–25.

Williams, G. (1999). Evolutionary Hot Spots Data Mining. *3rd Pacific-Asia Conference, PAKDD-99, Beijing, China, April*, pages 184–193.

Zhang, C. (2002). *Association Rule Mining: Models and Algorithms*. Lecture Notes in Computer Science 2307.

Zitzler, E. and Thiele, L. (1998a). An Evolutionary Algorithm for Multiobjective Optimisation: The Strength Pareto Approach. Technical Report 43, Swiss Federal Institute of Technology (ETH), Switzerland.

Zitzler, E. and Thiele, L. (1998b). Multiobjective Optimisation using Evolutionary Algorithms: A Comparative Case Study. *Parallel Problem Solving from Nature, Amsterdan*, pages 292–301.