

An Evolutionary Approach to Discourse-level Knowledge Discovery from Texts^{*}

John Atkinson and Anita Ferreira

Department of Computer Sciences and Department of Spanish, Universidad de Concepción
Concepción, Chile.
atkinson@inf.udec.cl and aferreir@udec.cl

Abstract. This paper proposes a new approach for mining novel patterns from textual databases which considers both the mining process itself, the evaluation of this knowledge, and the human assessment. This is achieved by integrating Information Extraction technology and Genetic Algorithms to produce high-level explanatory novel hypotheses. Experimental results using the model are discussed and the assessment by human experts are highlighted.

1 Introduction

An important problem in processing real texts for text mining purposes is that 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 [6], 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 [5]. Nevertheless, these discoveries are useless unless they contribute valuable knowledge for users who make strategic decisions. This leads then to a complicated activity referred to as *Knowledge Discovery from Texts* (KDT).

KDT can potentially benefit from successful techniques from Data Mining or KDD [4] which have been applied to relational databases. However, DM/KDD techniques cannot be immediately applied to text data for the purposes of TM as they assume a structure in the source data which is not present in free text. Hence new representations for text data have to be used. Also, while the assessment of discovered knowledge in the context of KDD is a key aspect for producing an effective outcome, the assessment of the patterns discovered from text has been a neglected topic in the majority of the KDT approaches. Consequently, it is

^{*} This research was partially supported by the National Council for Scientific and Technological Research (FONDECYT, Chile) under grant number 1070714: “*An Interactive Natural-Language Dialogue Model for Intelligent Filtering based on Patterns Discovered from Text Documents*”

not proved whether the discoveries are novel, interesting, and useful for decision makers.

The most sophisticated approaches to text mining or KDT are characterized by an intensive use of external electronic resources including ontologies, thesauri, etc., which highly restricts the application of the unseen patterns to be discovered, and their domain independence. In addition, the systems so produced have few metrics (or none at all) which allow them to establish whether the patterns are interesting and novel.

In terms of data mining techniques, Genetic Algorithms (GA) for Mining purposes has several promising advantages over the usual learning methods employed in KDT: the ability to perform global search, the exploration of solutions in parallel, the robustness to cope with noisy and missing data (something critical in dealing with text information as partial text analysis techniques may lead to imprecise outcome data), and the ability to assess the goodness of the solutions as they are produced.

In this paper, we propose a new model for KDT which brings together the benefits of shallow text processing and GAs to produce effective novel knowledge. In particular, the approach put together *Information Extraction* (IE) technology and multi-objective evolutionary computation techniques. It aims at extracting key underlying linguistic knowledge from text documents (i.e., rhetorical and semantic information) and then hypothesizing and assessing interesting and unseen explanatory knowledge. Unlike other approaches to KDT, we do not use additional electronic resources or domain knowledge beyond the text database.

2 Related Work

In the context of KDT systems, some current applications show a tendency to start using more structured or deeper representations than just keywords (or terms) to perform further analysis so to discover unseen patterns. Early research on this kind of approach is derived from seminal work by Swanson [8] on exploratory analysis from the titles of articles stored in the MEDLINE medical database. Swanson designed a system to infer key information by using simple patterns which recognize causal inferences such as "X **cause** Y" and more complex implications, which lead to the discovery of hidden and previously neglected connections between concepts. This work provided evidence that it is possible to derive new patterns from a combination of text fragments plus the explorer's medical expertise.

Further approaches have exploited these ideas by combining more elaborated IE patterns and general lexical resources (e.g., WordNet) [5] or specific concept resources (i.e., thesauri). They deal with automatic discovery of new lexico-semantic relations by searching for corresponding defined patterns in unrestricted text collections so as to extend the structure of the given ontology/thesaurus (i.e., new relations, new concepts).

A different view in which linguistic resources such as WordNet are used to assist the discovery and to evaluate the unseen patterns is followed by Mooney

and colleagues [1] who propose a system to mine for simple rules from general documents by using IE extraction patterns. Furthermore, human subjects assess the real interestingness of the most relevant patterns mined by the system. The WordNet approach to evaluation has proved to be well correlated with human judgments. However, the dependence on a linguistic resource prevents the method from dealing with specific terminology leading to missing and/or misleading information.

3 Semantically-guided Patterns Discovery from Texts

We developed a semantically-guided model for evolutionary Text Mining which is domain-independent but genre-based. Unlike previous approaches to KDT, our approach does not rely on external resources or descriptions hence its domain-independence. In addition, a number of strategies have been developed for automatically evaluating the quality of the hypotheses. This is an important contribution on a topic which has been neglected in most of KDT research over the last years.

Evolutionary computation techniques (i.e., GA) have been adopted in our model to KDT and others have been designed from scratch.

The proposed model has been divided into two phases. The first phase is the preprocessing step aimed to produce both training information for further evaluation and the initial population of the GA. The second phase constitutes the knowledge discovery itself, in particular this aims at producing and evaluating explanatory unseen hypotheses.

In order to generate an initial set of hypotheses, an initial population is created by building random hypotheses from the initial rules. The GA then runs for a number of generations until a fixed number of generations is achieved. At the end, a small set of the best hypotheses are obtained.

The description of the paper is organized as follows: section 3.1 presents the main features of the text preprocessing phase and how the representation for the hypotheses is generated. In addition, training tasks which generate the initial knowledge to feed the discovery are described. Section 3.2 highlights constrained genetic operations to enable the hypotheses discovery, and proposes different evaluation metrics to assess the plausibility of the discovered hypotheses.

3.1 Text Preprocessing and Training

An underlying principle in our approach is to be able to make good use of the structure of the documents for the discovery process. For this, we have restricted our scope somewhat to consider a scientific genre involving scientific/technical abstracts. These have a well-defined macro-structure (genre-dependent rhetorical structure) to “summarize” what the author states in the full document (i.e., background information, methods, conclusions, etc). From this kind of document’s structure, important constituents can be identified such as *Rhetorical Roles*, *Predicate Relations*, and *Causal Relation(s)* [6].

In order to extract this initial key information from the texts, an IE module was built. Essentially, it takes a set of text documents, has them tagged through a previously trained Part-of-Speech (POS) tagger, and produces an intermediate representation for every document (i.e., template, in an IE sense) which is then converted into a general rule. A set of hand-crafted domain-independent extraction patterns were written and coded.

For this purpose, key training data are captured from the corpus of documents itself and from the semantic information contained in the rules. This can guide the discovery process in making further similarity judgments and assessing the plausibility of the produced hypotheses.

In order to obtain training information from the Corpus, we have designed a semi-structured Latent Semantic Analysis (LSA) representation [7, 3] for text data in which we represent predicate information (i.e., verbs) and arguments separately once they have been properly extracted in the IE phase. In addition, training information from the texts is not sufficient as it only conveys data at a word semantics level.

Accordingly, we perform two kinds of tasks: creating the initial population and computing training information from the rules. In computing training information, two kinds of key training data are obtained: correlations between rhetorical roles and predicate relations which establishes associations between rhetorical information and the predicate action performed (i.e., in certain domains, the *goal* of some hypothesis is likely to be associated with the *construction* of some component, etc) and the co-occurrences of rhetorical information, in which valid hypotheses are assessed in terms of their semantic coherence [7].

3.2 Mining and Evaluating Plausible Patterns

The approach to KDT is strongly guided by semantic and rhetorical information, and consequently there are some soft constraints to be met before producing the offspring so as to keep them coherent.

The GA will start from a initial population, which in this case, is a set of semi-random hypotheses built up from the preprocessing phase. Next, constrained GA operations are applied and the hypotheses are evaluated. In order for every individual to have a fitness assigned, we use a evolutionary multi-objective optimization strategy [9] in a way which allows incremental construction of a Pareto-optimal set and uses a steady-state strategy for the population update.

Patterns Discovery Using the semantic measure above and additional constraints discussed later on, we propose new operations to allow guided discovery such that unrelated new knowledge is avoided, as follows:

- *Selection*: selects a small number of the best parent hypotheses of every generation (*Generation Gap*) according to their Pareto-based fitness.
- *Crossover*: a simple recombination of both hypotheses' conditions and conclusions takes place, where two individuals swap their conditions to produce new offspring (the conclusions remain).

Under normal circumstances, crossover works on random parents and positions where their parts should be exchanged. However, in our case this operation must be restricted to preserve semantic coherence. We use soft semantic constraints to define two kind of recombination:

1. *Swanson's Crossover*: based on Swanson's hypothesis [8] we propose a recombination operation as follows:

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

The principle above can be seen in Swanson's crossover between two learned hypotheses as shown in figure 1(a).

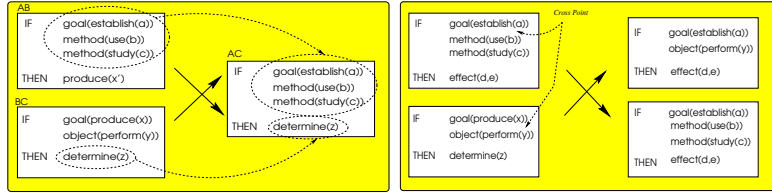


Fig. 1. (a) Semantically-guided Swanson Crossover. (b) Default Semantic Crossover

2. *Default Semantic Crossover*: if the previous transitivity does not apply then the recombination is performed as long as both hypotheses as a whole have high semantic similarity which is defined in advance by providing minimum thresholds (figure 1(b)).

- *Mutation*: aims to make small random changes on hypotheses to explore new possibilities in the search space. As in recombination, we have dealt with this operation in a constrained way, so we propose three kinds of mutations to deal with the hypotheses' different objects: *Role Mutation*, *Predicate Mutation*, and *Argument Mutation*.
- *Population Update*: we use a non-generational GA in which some individuals 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. We use a steady-state strategy in which each individual from a small number of the worst hypotheses is replaced by an individual from the offspring only if the latter are better than the former.

Assessment and Analysis Since each hypothesis in our model has to be assessed by different criteria, usual methods for evaluating fitness are not appropriate. Hence *Evolutionary Multi-Objective Optimization* (EMOO) techniques which use the multiple criteria defined for the hypotheses are needed. Accordingly, we propose EMOO-based evaluation metrics to assess the hypotheses' fitness in a domain-independent way and, unlike other approaches, without using any external source of domain knowledge.

In order to establish evaluation criteria, we have taken into account different issues concerning plausibility, and quality itself. Accordingly, we have defined eight evaluation criteria to assess the hypotheses given by: **relevance, structure, cohesion, interestingness, coherence, coverage, simplicity, plausibility of origin**:

- **Relevance** (*How important is the hypothesis to the target question?*): measures the semantic closeness between the hypothesis' predicates and the target concepts. Relevance is then computed from compound vectors obtained in the LSA analysis which follows work by Kintsch on *Predication* [7]. We then propose an adaptation of the LSA-based closeness so to compute the overall relevance of the hypothesis in terms of the "strength" which determines how closely related two concepts are to both some predicate and its arguments.
- **Structure** (*How good is the structure of the rhetorical roles?*): measures how much of the rules' structure is exhibited in the current hypothesis. Since we have previous preprocessing information regarding bi-grams of roles, the structure is computed by following a Markov chain of the "bi-grams" of the rhetorical information of each hypothesis. From this model, it can be observed that some structures tags are more frequent than others.
- **Cohesion** (*How likely is a predicate action to be associated with some specific rhetorical role?*): measures the degree of "connection" between rhetorical information and predicate actions. The issue here is how likely some predicate relation P in the current hypothesis is to be associated with role r .
- **Interestingness** (*How interesting is the hypothesis in terms of its antecedent and consequent?*): Unlike other approaches to measure "interestingness" which use an external resource (e.g., WordNet) and rely on its organisation we propose a different view where the criterion can be evaluated from the semi-structured information provided by the LSA analysis. Accordingly, the measure for hypothesis H is defined as a degree of unexpectedness, that is, the semantic dissimilarity between the rule antecedent and consequent. Here, the lower the similarity, the more interesting the hypothesis is likely to be. Otherwise, it means the hypothesis involves a correlation between its antecedent and consequent which may be commonsense knowledge.
- **Coherence**: This metrics addresses the question whether the elements of the current hypothesis relate to each other in a semantically coherent way, a property which has long been dealt with in the linguistic domain, in the context of *text coherence* [3].

As we have semantic information provided by the LSA analysis which is complemented with rhetorical and predicate-level knowledge, we developed a simple method to measure coherence, following work by [3] on measuring text coherence. Semantic coherence is calculated by considering the average semantic similarity between consecutive elements of the hypothesis.

- **Coverage**: The coverage metric tries to address the question of how much the hypothesis is supported by the model (i.e., rules representing documents and semantic information). For this, we say that a hypothesis covers an

extracted rule only if the predicates of the hypothesis are roughly (or exactly, in the best case) contained in this rule. Once the set of rules covered is computed, the criterion can finally be computed as the proportion of rules covered by the hypothesis.

- **Simplicity** (*How simple is the hypothesis?*): shorter and/or easy-to-interpret hypotheses are preferred. Since the criterion has to be maximized, the evaluation will depend on the length (number of elements) of the hypothesis.
- **Plausibility of Origin** (*How plausible is the hypothesis produced by Swanson's evidence?*): If the current hypothesis was an offspring from parents which were recombined by a Swanson's transitivity-like operator, then the higher the semantic similarity between one parent's consequent and the other parent's antecedent, the more precise is the evidence, and consequently worth exploring as a novel hypothesis.

Note that since we are dealing with a multi-objective problem, there is no simple way to get independent fitness values as the fitness involves a set of objective functions to be assessed for every individual. Therefore the computation is performed by comparing objectives of one individual with others in terms of *Pareto dominance* [2] in which non-dominated solutions (Pareto individuals) are searched for in every generation.

Next, three important issues had to be faced in order to assess every hypothesis' fitness: Pareto dominance, fitness assignment and the diversity problem [2]. In particular, Zitzler [9] proposes an interesting method, *Strength Pareto Evolutionary Algorithm* (SPEA) which uses a mixture of established methods and new techniques in order to find multiple Pareto-optimal solutions in parallel, and at the same time to keep the population as diverse as possible. We have also adapted the original SPEA algorithm to allow for the incremental updating through a steady-state replacement method.

4 Analysis and Results

The quality (novelty, interestingness, etc) of the discovered knowledge by the model was assessed by building a Prolog-based KDT system. 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. The initial rules are represented by facts containing lists of relations both for antecedent and consequent.

For the purpose of the experiments, the corpus of documents has been obtained from the *AGRIS* database for agricultural and food science. We selected this kind of corpus as it has been properly cleaned-up, and builds upon a scientific area which we do not have any knowledge about so to avoid any possible bias and to make the results more realistic. A set of 1000 documents was extracted from which one third were used for setting parameters and making general adjustments, and the rest were used for the GA itself in the evaluation stage.

We then tried to provide answers a basic question concerning our original aims: How good are the hypotheses produced according to human experts in terms of text mining’s ultimate goals: interestingness, novelty and usefulness, etc.

In order to address this issue, we used a methodology consisting of two phases: the system evaluation and the experts’ assessment.

1. *System Evaluation*: this aims at investigating the behavior and setting the parameter values used by the evolutionary model for KDT.

We set the GA by generating an initial population of 100 semi-random hypotheses. In addition, we defined the main global parameters such as *Mutation Probability* (0.2), *Cross-over Probability* (0.8), *Maximum Size of Pareto set* (5%), etc. We ran five versions of the GA with the same configuration of parameters but different pairs of terms to address the quest for explanatory novel hypotheses.

2. *Expert Assessment*: this aims at assessing the quality of the discovered knowledge on different criteria by human domain experts. For this, we designed an experiment in which 20 human experts were involved and each assessed 5 hypotheses selected from the Pareto set. We then asked the experts to assess the hypotheses from 1 (worst) to 5 (best) in terms of the following criteria: Interestingness (INT), Novelty (NOV), Usefulness (USE), Sensibleness (SEN), etc.

In order to select worthwhile terms for the experiment, we asked one domain expert to filter pairs of target terms previously related according to traditional clustering analysis. The pairs which finally deserved attention were used as input in the actual experiments (i.e., **glycocide and inhibitors**).

Once the system hypotheses were produced, the experts were asked to score them according to the five subjective criteria. Next, we calculated the scores for every criterion as seen in the overall results in table 1 (for length’s sake, only some criterion are shown).

The assessment of individual criteria shows some hypotheses did well with scores above the average (50%) on a 1-5 scale. This is the case for hypotheses 11, 16 and 19 in terms of INT, hypotheses 14 and 19 in terms of SEN, hypotheses 1, 5, 11, 17 and 19 in terms of USE, and hypotheses 24 in terms of NOV, etc.

These results and the evaluation produced by the model were used to measure the correlation between the scores of the human subjects and the system’s model evaluation. 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.

We then calculated the pair of values for every hypothesis and obtained a (Spearman) correlation $r = 0.43$ (t -test = 23.75, $df = 24$, $p < 0.001$). From this result, we see that the correlation shows a good level of prediction compared to humans. This indicates that for such a complex task, the model’s behavior is not too different from the experts’.

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 one of the

Criterion	No. of Hypotheses	
	Negative < Average	Positive ≥ 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 1. Distribution of Experts’ assessment of Hypothesis per Criteria

best hypotheses as assessed by the experts (out of 25 best hypotheses) considering the average value of the 5 scores assigned by the user. For example, hypothesis 65 of run 4 looks like: **IF goal(perform(19311)) and goal(analyze(20811)) THEN establish(111)**

Where the numerical values represent internal identifiers for the arguments and their semantic vectors, and its resulting criteria vector is [0.92, 0.09, 0.5, 0.005, 0.7, 0, 0.3, 0.25] (the vector’s elements represent the values for the criteria relevance, structure, coherence, cohesion, interestingness, plausibility, coverage, and simplicity) and obtained an average expert’s assessment of 3.74. In natural-language text, this can roughly be interpreted as:

- The work **aims** at **performing** the genetic grouping of seed populations and investigating a tendency to the separation of northern populations into different classes.
- The **goal** is to **analyze** the vertical integration for producing and selling Pinus Timber in the Andes-Patagonia region.
- As a **consequence**, the best agricultural use for land lots of organic agriculture must be **established** to promote a conservationist culture in priority or critical agricultural areas.

The hypothesis appears to be more relevant and coherent than the others (relevance = 92%). However, this is not complete in terms of cause-effect. For instance, the methods are missing.

In addition, there is also qualitative evidence that there were other subjective factors which influenced some hypotheses’ low scores, which was extracted from the experts’ overall comments such as the origin and expertise of the experts, the hypotheses understanding, etc.

5 Conclusions

In this work we contribute a novel way of combining additional linguistic information and evolutionary learning techniques in order to produce novel hypotheses which involve explanatory and effective novel knowledge.

We also introduced a unique approach for evaluation which deals with semantic and Data Mining issues in a high-level way. In this context, the proposed representation for hypotheses suggests that performing shallow analysis of the documents and then capturing key rhetorical information may be a good level of processing which constitutes a trade off between completely deep and keyword-based analysis of text documents. In addition, the results suggest that the performance of the model in terms of the correlation with human judgments are slightly better than approaches using external resources. In particular criteria, the model shows a very good correlation between the system evaluation and the expert assessment of the hypotheses.

The model deals with the hypothesis production and evaluation in a very promising way which is shown in the overall results obtained from the experts evaluation and the individual scores for each hypothesis. However, it is important to note that unlike the experts who have a lot of experience, preconceived concept models and complex knowledge in their areas, the system has done relatively well only exploring the corpus of technical documents and the implicit connections contained in it.

References

1. S. Basu, R. Mooney, K. Pasupuleti, and J. Ghosh. 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, June 2001.
2. Kalyanmoy Deb. *Multi-objective Optimization Using Evolutionary Algorithms*. Wiley, 2001.
3. P. Foltz, W. Kintsch, and T. Landauer. The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse processes*, 25(2):259–284, 1998.
4. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan-Kaufmann, 2001.
5. M. Hearst. Text Mining Tools: Instruments for Scientific Discovery. *IMA Text Mining Workshop, USA*, April 2000.
6. D. Jurafsky and J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, 2000.
7. W. Kintsch. Predication. *Cognitive Science*, 25(2):173–202, 2001.
8. D. Swanson. 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*, 27(3), February 2001.
9. E. Zitzler and L. Thiele. An Evolutionary Algorithm for Multiobjective Optimisation: The Strength Pareto Approach. Technical Report 43, Swiss Federal Institute of Technology (ETH), Switzerland, 1998.