Combining Information Extraction with Genetic Algorithms for Text Mining.

Article *in* Intelligent Systems, IEEE · June 2004

DOI: 10.1109/MIS.2004.4 · Source: IEEE Xplore

John Atkinson
Universidad Adolfo Ibáñez, Santiago
68 PUBLICATIONS **370** CITATIONS

Stuart Aitken
The University of Edinburgh
90 PUBLICATIONS **1,107** CITATIONS

Combining Information Extraction with Genetic Algorithms for Text Mining

John Atkinson-Abutridy, Chris Mellish, and Stuart Aitken, *University of Edinburgh*

An evolutionary approach that combines information extraction technology and genetic algorithms can produce a new, integrated model for text mining.

T*ext mining* discovers unseen patterns in textual databases. But these discoveries are useless unless they contribute valuable knowledge for users who make strategic decisions. Confronting this issue can lead to *knowledge discovery from texts*, a complicated activity that involves both discovering unseen knowledge (through TM) and evaluating this potentially valuable knowledge. KDT can benefit from techniques that have been useful in data mining or *knowledge discovery from databases*.[1] However, you can't immediately apply data mining techniques to text data for TM because they assume a structure in the source data that isn't in free text. You must therefore use new representations for text data.

In many TM applications, you can use more structured representations than just keywords to perform analysis to uncover unseen patterns. Early research on such an approach was based on seminal work on exploratory analysis of article titles stored in the Medline medical database.[2] Other approaches have exploited these ideas by combining more elaborated *information extraction* patterns and general lexical resources such as WordNet[3] or specific concept resources such as thesauri.[4] Another approach, relying on IE patterns, uses linguistic resources such as WordNet to assist the discovery and evaluation of patterns to extract basic information from general documents.[5]

In this context, we propose a new KDT approach that combines fragments of key information extracted from text documents and then uses a multicriteria optimization strategy to produce explanatory knowledge without using external resources. Although researchers have tackled KDD tasks as learning problems, the nature of KDD suggests that data mining is a continuous task of searching for and optimizing potential hypotheses that can maximize quality criteria.

A significant number of successful, practical search-and-optimization techniques exist,[6] but some techniques are more appealing for KDD tasks than others. In particular, genetic algorithms look promising. Compared with classical search-and-optimization algorithms, GAs are much less susceptible to getting stuck in local suboptimal regions of the search space because they perform global searches by exploring multiple solutions in parallel. Being robust, GAs can cope with noisy and missing data. However, to use GAs effectively in KDT, we must tackle several problems first, including devising high-level representations and tailoring new genetic operations.

Evolutionary knowledge discovery

We've brought together the benefits of GAs for data mining and genre-based IE technology to propose a new approach for high-level knowledge discovery. Unlike previous KDT approaches, our model doesn't rely on external resources or conceptual descriptions. Instead, it performs the discovery using only information from the original corpus of text documents and from training data computed from them. The GA that produces the hypotheses is strongly guided by semantic constraints, which means that several specifically defined metrics evaluate the quality and plausibility.

Figure 1 shows our approach's working model divided into two general levels of processing. The input is a corpus of technical and scientific natural language documents; the output is a small set of the hypotheses that the GA discovered.

The first level involves a preprocessing step that

1094-7167/04/$20.00 © 2004 IEEE
Published by the IEEE Computer Society

produces both the training data for further automatic evaluation of the hypotheses and the initial population of the GA from information extracted from the documents. The IE task applies genre-based extraction patterns and then generates a rule-like representation for each document in a domain corpus. That is, after processing *n* documents, the extraction stage will produce *n* rules, each one representing the document's content in terms of a cause-and-effect relation. These rules, along with previously generated training data, will become the semantic model that guides the GA-based discovery. To create the initial population, we create a random set of hypotheses by combining random units from the extracted rules.

The second level is the GA-based knowledge discovery, which aims to produce the explanatory hypotheses. The GA runs for several generations until achieving a maximum number of learning generations. At the end, we obtain a small set of the best *K* hypotheses. *K* is a fixed user-defined value, $1 < K < p$.

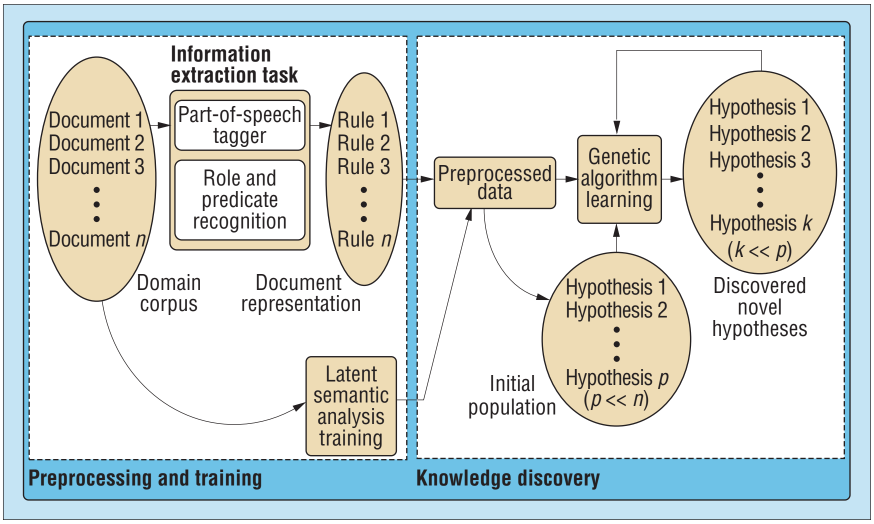


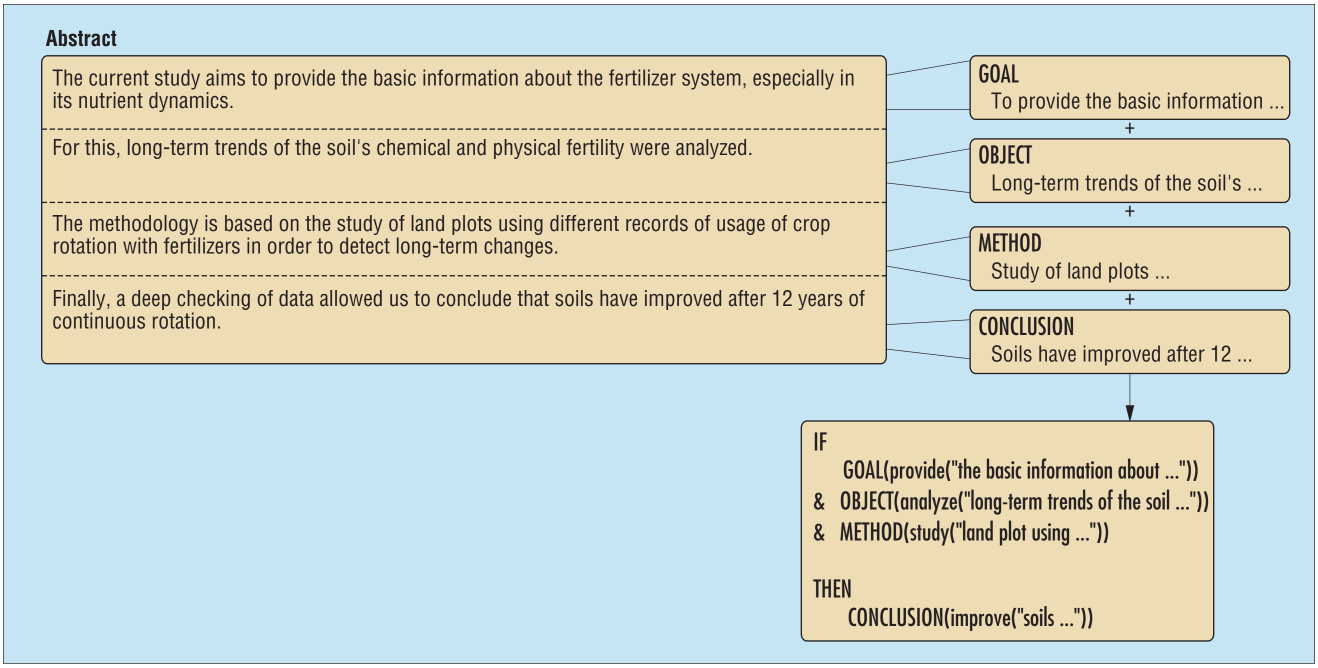
Figure 1. Our genetic-algorithm-based approach for knowledge discovery from texts.

Genre-based IE

To extract key information from the texts, you must define the representation and specify how it's constructed from the extracted information. To this end, we used a genre-based approach. Specifically, we used the genre of technical abstracts. Typically, abstracts have a well-defined structure that authors use to summarize their ideas and state key facts concisely. This makes abstracts suitable for further shallow analysis and avoids many conceptual-level ambiguities related to the restricted use of concepts in specific contexts.

Linguistic evidence shows that an abstract in a given domain follows a prototypical and even modular organization—the genre-dependent rhetorical structure—that its author uses to express the background information, methods, achievements, and conclusions.[7] From a scientific viewpoint, there are also claims that important findings could be searched by linking this kind of information across the documents.

Unlike other researchers that exploit this structure in their work, we use the rhetorical structure and the semantic information in it differently. As Figure 2 shows, starting from an abstract, the IE task extracts information at a rhetorical and semantic level, then uses this information in a rule-like form to represent the documents' key facts. Because we assume that what's stated in the document


Abstract

The current study aims to provide the basic information about the fertilizer system, especially in its nutrient dynamics.

For this, long-term trends of the soil's chemical and physical fertility were analyzed.

The methodology is based on the study of land plots using different records of usage of crop rotation with fertilizers in order to detect long-term changes.

Finally, a deep checking of data allowed us to conclude that soils have improved after 12 years of continuous rotation.

GOAL
To provide the basic information ...
+
OBJECT
Long-term trends of the soil's ...
+
METHOD
Study of land plots ...
+
CONCLUSION
Soils have improved after 12 ...

```
IF
    GOAL(provide("the basic information about ..."))
& OBJECT(analyze("long-term trends of the soil ..."))
& METHOD(study("land plot using ..."))

THEN
    CONCLUSION(improve("soils ..."))
```

Figure 2. Rule representation from the semantic and rhetorical information extracted from an abstract.

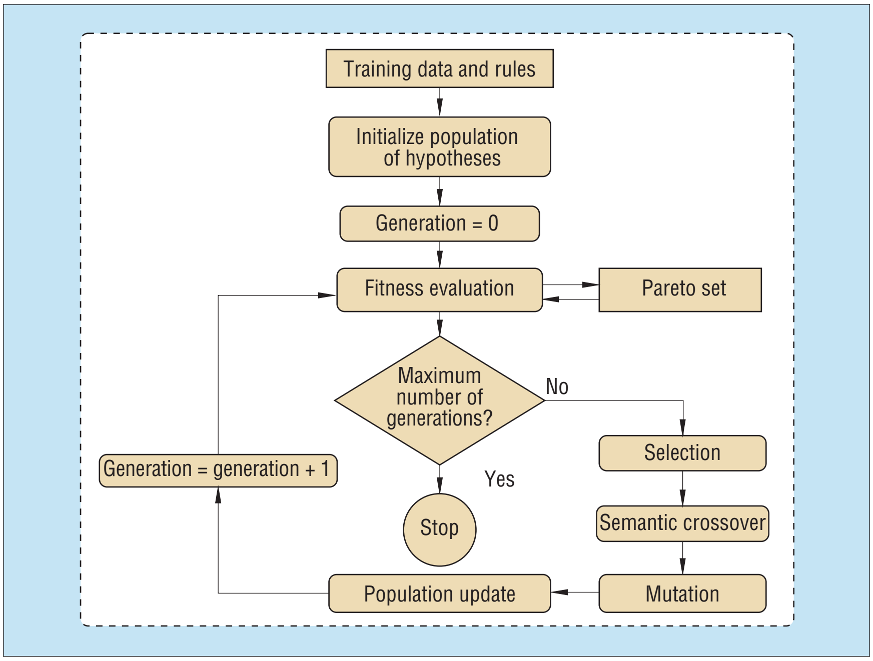


Figure 3. A semantically constrained and multicriteria genetic algorithm.

follows a form of antecedent-consequent reasoning drawn by the author, we convert the intermediate template into a rule-like form that aims to capture the author's scientific evidence and conclusion.

To produce this information, the IE task takes the set of tagged documents and produces a template representation for every document. We then easily convert this representation into an if-then rule. For this purpose, we wrote a set of domain-independent extraction patterns so that we could match them against the input documents. Each extraction pattern constructs an output representation that involves two levels of linguistic knowledge: the rhetorical information expressed in the abstract and the semantic information contained in it, which we later convert into a predicate-like form. We specify these extraction patterns as follows:

con el proposito de VP …: goal(&ACTION[VP])
(In order to ACTION (OBJECT) …),

The left-hand expression states the pattern to be identified (con el proposito de), and the right-hand side (following the colon) states the corresponding semantic action to be produced. We decompose verb phrase (VP) components into two elements: the predicate action and the sequence of terms that represent its argument. For example, if the input tagged text looks like

con/p el/art proposito/n de/p producir/inf tomates/n en/p epoca/s de/p invierno/s …,

the intermediate representation will be

$$\underbrace{goal}_{role}\overbrace{\left(\underbrace{producir}_{predicate}\underbrace{[tomates, en, epoca, de, invierno, \ldots]}_{argument}\right)}^{VP},$$

The product will transform this representation into goal(producir('tomates en epoca de …')). Although we might not know the set of predicates, we must specify the rhetorical roles in advance because they're common across the technical genre.

Generating training data

Rules extracted this way constitute key elements for producing and evaluating new hypotheses as the GA goes on. However, the rules themselves don't suffice. Specifically, to make similarity judgments in producing hypotheses, we must use the whole corpus of documents to obtain initial knowledge at the lexical-semantics level provided by *latent semantic analysis*.[8] We augment this knowledge with syntactic information to represent the predicates in a vector where the predicates and arguments are converted into vectors that represent the meaning according to the context similarities learned by LSA.

The initial extracted rules also convey underlying data and associations that can guide evolutionary discovery. First, we build the initial population of hypotheses randomly by combining rhetorical and semantic information. We keep these basic units in a separate database to provide the genetic operations with new information. Second, information about the associations between rhetorical roles and predicate actions provides insights for discovering coherent hypotheses. Indeed, some predicate actions might be more likely to happen with specific rhetorical information than others. The preprocessing takes this into account through a Bayesian approach that accounts for this kind of association by computing the conditional probabilities of predicates p, given some attached rhetorical role r—namely, $Prob(p \mid r)$.

In producing plausible hypotheses, the rhetorical roles' organization, as in a text discourse, is worth exploring because the meaning of the scientific evidence stated in the abstract can subtly change if the facts' order changes. Indeed, changing the order of the rhetorical information can also alter the coherence between the paragraphs of a text. This suggests that in generating valid hypotheses, some rule structures will be more desirable than others. Therefore, the creation of hypotheses must be fed with information concerning a good structure. To generate this information, we can think of a rule's p roles as a sequence of tags, $<r_1, r_2, \ldots, r_p>$, such that r_i precedes r_{i+1}. So, we generate the conditional probabilities $Prob(r_p \mid r_q)$, for every role p, q—that is, the probability that r_q precedes r_p, which we'll further use to evaluate the new hypotheses.

All the training information, the obtained associations, and the semantic knowledge provided by the semistructured LSA will guide how the GA produces hypotheses. In other words, this constitutes the model the GA uses to guide the search for the plausible hypotheses in the whole search space.

Hypotheses discovery and evaluation

Once we obtain the training data and the semantic information provided by LSA, the GA starts off from the initial population by searching for optimal hypotheses (as shown in Figure 3) that satisfy multiple quality criteria. Next, we apply semantically constrained genetic operations and evaluate the hypotheses according to their ranking obtained from evaluating these criteria.

We designed the genetic operations to guide the GA-based search and avoid incoherent knowledge. We developed three semantically constrained operations: *selection*, *crossover*, and *mutation*.

Selection picks a small number of the best hypotheses of every generation to be reproduced according to their fitness.

Crossover recombines two selected hypotheses and takes place with some probability, where both of them swap their elements at some random position of the hypotheses to produce new offspring. However, because we want to restrict the operation to preserve semantic plausibility, we defined two kinds of recombination (see Figure 4).

On the basis of Don Swanson's inference between the titles of documents,[2] we propose a recombination operation (which we call *Swanson's crossover*) to allow any kind of relationship to hold between hypotheses AB and BC (see Figure 4a). This operation is more flexible than Swanson's patterns. If this transitivity-like operator doesn't apply, we perform the recombination in the usual way by swapping conditions of the hypotheses (see Figure 4b), as long as both hypotheses meet some minimum semantic similarity. In Swanson's crossover, the semantic similarity between parts of the parent hypotheses is kept for a further evaluation. This is crucial because the higher the similarity, the better the accuracy of the offspring of this kind of Swanson-like inference.

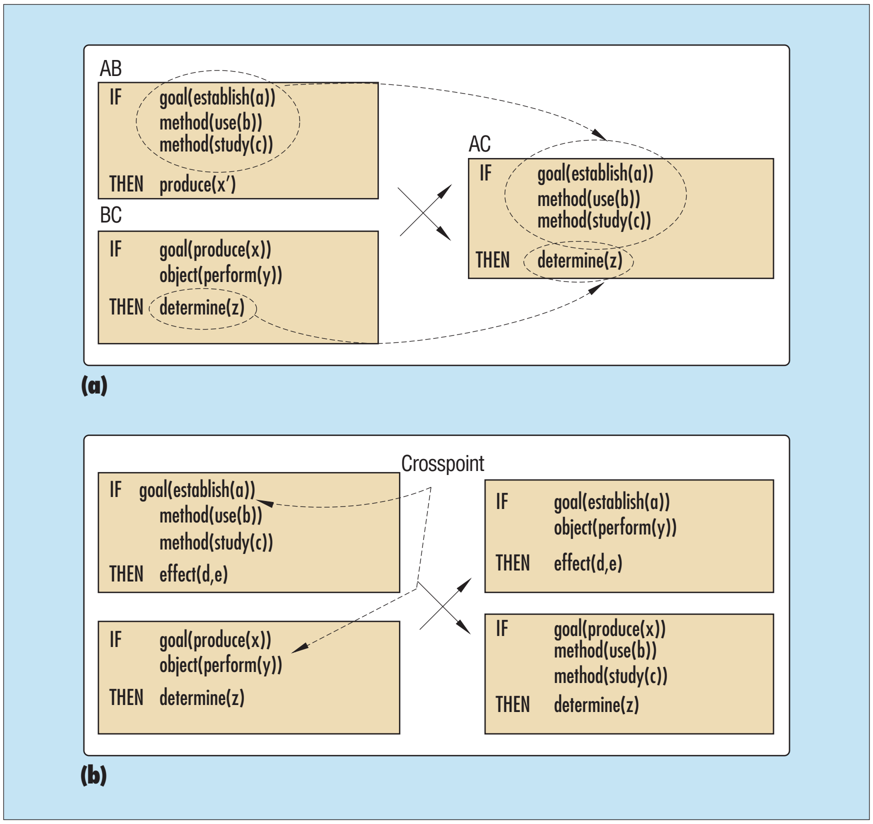


Figure 4. Two kinds of recombination: (a) Swanson's crossover and (b) the default semantic crossover.

Mutation aims to make small random changes in the hypotheses to explore new possibilities in the search space. We've developed three kinds of constrained mutations. *Role mutation* selects one rhetorical role (including its contents) and randomly replaces it with one from the initial database of roles and predicates. *Predicate mutation* selects one predicate action and argument and randomly replaces them with others, which modifies the association between semantics and rhetoric. *Argument mutation*, because we have no information about the arguments' semantic types, follows a constrained procedure that randomly chooses a new argument from those predicates that have the same name and number of arguments as the current one.

These operators' role is only to produce new hypotheses that might become part of the new generation. However, we must evaluate these individuals' goodness to establish whether they provide plausible hypotheses. Consequently, the fittest individuals will survive to the next learning generation, and others will be eliminated. For this to happen, we must develop appropriate evaluation criteria to measure the hypotheses' quality (from a KDD viewpoint) and design an optimization strategy that trades off between these multiple criteria to create a pool of better hypotheses.

In developing evaluation metrics, we must consider two different issues. The first is metrics related to the hypotheses' meaning and semantics that aim at ensuring that the hypotheses are coherent and meaningful. The second is metrics related to quality from a KDD viewpoint. We designed a set of eight domain-independent evaluation criteria. The GA's goal is to explore new hypotheses that maximize these criteria. This process produces the set of vectors representing the criteria values for every hypothesis. That is, each hypothesis contains an eight-dimensional objective vector. We then obtain the best hypotheses of every generation through a multicriteria optimization strategy. We define the different proposed criteria in the following sections.

Relevance

The relevance model aims to not only discover novel and interesting knowledge but also provide explanation for the discovered knowledge. When comparing this approach to bag-of-word approaches, you'd expect the model to provide a meaningful relationship between different concepts. For example, these concepts might be the terms of a produced cluster in a text-mining approach that searches an unknown relationship.

Because we aim to produce explanatory hypotheses about the novel discovered knowledge, these target concepts can also be part of a major request for the model: to find the set of the best novel hypotheses that help us understand the relationship between a pair of user-defined target concepts. With this in mind, the relevance criterion measures how relevant a hypothesis is to the user-defined target concepts. Specifically, this criterion evaluates the semantic closeness between the predicates of a hypothesis and the target concepts to provide more information about the unknown relationship between these concepts.

Because the LSA-based method we use to evaluate this similarity doesn't consider the context to represent the predicate information, we measure relevance in terms of these target concepts' influence on the hypothesis.

We perform this mainly by using a variation of the strength concept (proposed by Walter Kintsch) between a predicate and the surrounding terms in the semantic neighborhood.[9]

We compute the relevance of a hypothesis H with predicates $P_i(A_i)$ and target concepts ($<term1>$ and $<term2>$) as the average semantic similarity of every predicate (and argument) of H to the target concepts. That is,

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

where $|H|$ denotes the length of H (that is, the number of predicates).

Structure and cohesion

The structure criterion addresses the question of how good the rhetorical roles' structure is, which we can approximate by determining how much of the initial extracted rules structure is exhibited in the current hypothesis. To this end, this metric uses the training information provided at the beginning to compute the structure's quality according to a bigram model in which the roles r_i are a sequence of tags. That structure's quality is

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

In this equation, r_i represents the ith 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, and $Prob(r_i)$ denotes the probability that no role precedes r_i (the beginning of the structure).

The cohesion criterion addresses the question of how likely a predicate action will be associated with some specific rhetorical role. The underlying issue here is that some predicate relations P_i will be more likely than others to be associated with some rhetorical role r_i. For this, hypotheses containing this kind of association should be "rewarded" in the search-optimization phase.

Using the conditional probabilities provided by the training data, we express H's cohesion as

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

where $Prob(P_i \mid r_i)$ is the conditional probability of the predicate P_i given a rhetorical role r_i.

Coherence and coverage

The coherence metric addresses the question about whether the elements of the current hypothesis relate to each other in a semantically coherent way. Unlike rules produced by data mining techniques in which the order of the conditions is not an issue, the hypotheses produced in our model rely on pairs of adjacent elements that should be semantically sound, a property that has long been dealt with in the linguistic domain in the context of text coherence. Because we have semantic information provided by the LSA analysis that is complemented with rhetorical and predicate-level knowledge, we developed a simple method to measure coherence, following work on measuring text coherence.[10]

> The coherence metric addresses the question about whether the elements of the current hypothesis relate to each other in a semantically coherent way.

We calculate coherence by considering the average semantic similarity between consecutive elements of the hypothesis. However, we compute this closeness only on the semantic information that the predicates and their arguments convey because we considered the role structure in a previous criterion. Accordingly, we express the criterion as

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

where $(|H| - 1)$ denotes the number of adjacent pairs.

The coverage metric addresses the question of how much the model supports the hypothesis. KDD approaches have usually measured coverage of a hypothesis by considering some data structuring that isn't in textual information. In addition, most KDD approaches have assumed the use of linguistic or conceptual resources to measure coverage. We designed a hybrid method that combines semantic constraints and organization-related aspects to obtain the coverage. In general terms, a hypothesis H covers a rule R_i only if R_i contains the predicates of H.

The first step involves establishing the semantic similarity between H and R_i. However, because this relation is symmetrical, the second step analyzes whether the hypothesis contains elements of the rules. Formally, we define it as

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

In this equation, $SemanticsSimilarity(HP_k, P_j)$ represents the LSA-based similarity between predicates 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 H, and P_j denotes a predicate (with arguments) contained in RU_i. Once we compute the set of rules, we can compute the criterion as

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

where $|RulesCovered|$ denotes the size of the set of rules covered by H, and $|RuleSet|$ denotes the size of the initial set of extracted rules.

Simplicity and interestingness

The simplicity criterion addresses the question of how simple the hypothesis is. For this, we focus on hypothesis length. Because the criterion must be maximized, and shorter or easy-to-interpret hypotheses are preferable, the evaluation is simply

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

where $<MaxElems>$ denotes a fixed maximum number of elements allowed for any hypothesis.

The interestingness criterion captures the degree of surprisingness and/or unexpectedness in what the hypothesis conveys. Unlike another approach that uses a linguistic resource for this purpose,[11] we measure this criterion in terms of the unexpectedness of the relation between the antecedent and consequent of the hypothesis.

We can evaluate this criterion from the semistructured information provided by the LSA analysis. Because we're looking for interesting or unexpected connections, we assess this criterion through the semantic dissimilarity between antecedent and consequent—that is,

$$\begin{aligned}&\mathit{Interestingness}(H)\\ &= <\mathit{Dissimilarity\ between\ Antecedent\ and\ Consequent}>\\ &= 1 - \mathit{SemanticSimilarity}(An(H), Co(H)),\end{aligned}$$

where $An(H)$ and $Co(H)$ represent the antecedent and consequent of H.

Plausibility of origin

As we previously mentioned, Swanson's crossover encourages the production of potentially novel hypotheses in terms of a transitivity-like inference whose precision (degree of similarity of the parts of the parent hypotheses being recombined) might hopefully be higher for the cases considered worth exploring.

Thus, the *plausibility of origin* criterion measures the potential plausibility of the current hypothesis by remembering the quality of the inference when this hypothesis was created. This is necessary because other operations might have created the current hypotheses. In other words, keeping this similarity of Swanson's crossover provides a simple mechanism to remember that plausibility of origin considers hypotheses that only this operator has recombined.

Accordingly, the criterion for a hypothesis H is simply

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

Steady-state strategy

Once we evaluate the hypotheses' criteria, we must determine the best and worst individuals to choose which ones will survive and which ones won't. Consequently, we'll modify the current population so that (hopefully) we can consider better individuals in the next generation. How do we establish the set of best hypotheses in every generation? In a traditional GA, the notion of best or worst is clear because it involves picking the individuals with higher or lower fitness value. However, because we're dealing with multiple criteria, no unique fitness value exists. So, we must redefine the notion of optimum, which is usually a relation of dominance.[6]

A hypothesis H_1 *dominates* another hypothesis H_2 if both these conditions are true:

1. H_1 isn't worse than H_2 in all the criteria f_j, or $f_j(H_1) \not< f_j(H_2)$ for all $j = 1, 2, \ldots, N$ criteria.
2. H_1 is strictly better than H_2 in at least one criterion, or $f_j\,(H_1) > f_j\,(H_2)$ for at least one $j \in \{1, 2, \ldots, N\}$.

The best individuals are those with low fitness values because they contribute positively to improve the dominated individuals.

In other words, we trade off the vector of criteria of each hypothesis against the vector of the other competing hypotheses to determine the best ones. You usually refer to the set of nondominated individuals as the *Pareto set*. Computing this dominance relation to determine which hypotheses are part of the Pareto set doesn't say anything about their fitness.

To obtain the "quality" of each created hypothesis, we must do a fitness measure. To get an individual fitness value from several evaluation criteria, the multiobjective optimization strategies must complement the Pareto set. We use one of these strategies, called *strength Pareto evolutionary algorithm*,[12] which deals with the diversity of the solutions and the fitness assignment as a whole. An important aspect of SPEA is that some individuals remain unmodified from one generation to another. Because we're handling rule-like hypotheses, we must avoid losing some hypotheses' good material, so we must keep the offspring only if they're better than the population's worst parent. To do so, we adopt a steady-state strategy in which we replace a small portion of the worst parents with the offspring (from the best parents) only if the latter are better than the former.

The whole strategy's outcome is the computed Pareto set for every generation, which we incrementally update. We compute the fitness for the nondominated and dominated individuals differently. For nondominated individuals, the fitness is a proportion of the dominated individuals in the population. For dominated individuals, the fitness is the accumulated sum of the fitness of all the individuals that dominate the current solution. The algorithm will produce fitness values (such as strength) between 0 and 1 for the nondominated individuals, and fitness values greater than 1 for the dominated individuals. We prefer low fitness values because they represent good solutions.

If the number of Pareto members exceeds some user-defined threshold, we reduce the Pareto set by clustering its elements in terms of the similarity between their criteria vectors. Because of this clustering, the actual set maintained only approximates the true Pareto set. Because the optimization strategy aims to improve the solutions in the Pareto set as the GA goes on, we've explored two main choices:

- Improving the solutions in the Pareto set through the genetic operations
- Improving the dominated solutions by bringing them into the Pareto set or by modifying their genetic material to improve individual fitness values

The best individuals are those with low fitness values because they contribute positively to improve the dominated individuals.

Once we've computed the fitness and produced the offspring, we update the population of the worst parents with the offspring through the steady-state strategy.[13]

Evaluation and results

To evaluate our approach's search ability, we implemented a prototype of our model for GA-based KDT in Prolog and integrated it with the rest of the system as seen in Figure 1. Then we selected and cleaned up a corpus of documents in an example domain (agriculture). We used one-third of the documents for setting parameters and making general adjustments; the rest we used for the GA in the evaluation stage. From the documents, the

Table 1. Pairs of user-defined target terms used for each run.

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

Table 2. Overall evaluation.

Criterion	Confidence (95%)
Additional Information	2.60 ± 0.168
Interestingness	2.60 ± 0.173
Novelty	2.30 ± 0.205
Sensibleness	2.51 ± 0.237
Usefulness	2.56 ± 0.228

IE task extracted the corresponding 1,000 rules and training information, which we used to create an initial population of 100 semirandom hypotheses.

We then ran five versions of the GA, each one using the same global parameters but a different pair of target terms (see Table 1). These pairs of terms were regarded as relevant by a domain expert and therefore deserved further attention. Each run produced the overall best five hypotheses—that is, the best 25 hypotheses that contain optimum criteria according to the system.

We then designed a Web-based experiment in which we converted the different hypotheses into a readable natural language form to be assessed by 20 domain experts. To avoid a huge workload, each expert assessed no more than five hypotheses, and three experts assessed each hypothesis. We asked the experts to assess each hypothesis in terms of four KDD criteria of quality: Interestingness, Novelty, Usefulness, and Sensibleness. We added the Additional Information criterion to determine whether (according to the target concepts of the corresponding run from which the hypothesis was produced) the hypothesis contributes additional information to help the experts understand the unseen relationship between the previously defined target concepts.

Unlike other TM and KDT approaches, both the system evaluation and the expert's assessment consider multiple features of quality for the discovered knowledge. We can regard the hypotheses' overall quality as assessed by the experts as an average of these criteria. We performed the whole assessment in a scale between 1 (worst) to 5 (best). Figure 5 shows the average resulting scores in assessing 25 hypotheses for each criterion.

The assessment of individual criteria (see Table 2) illustrates that some hypotheses did well with scores above the average (3). This is the case for Hypotheses 11, 16, and 19 in terms of Interestingness (Hypotheses 7, 17, and 23 are just at the average), Hypotheses 14 and 19 in terms of Sensibleness (hypotheses 3, 11, and 17 are just at the average), Hypotheses 1, 5, 11, 17, and 19 in terms of Usefulness (Hypotheses 3, 10, and 15 are just at the average), and Hypothesis 24 in terms of Novelty (Hypotheses 11, 19, and 23 are just at the average). 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 Novelty), Hypothesis 18 always received a score below 2 (25 percent), except for Additional Information, in which its score is slightly higher.

Although there are very good individual hypotheses, the average scores show that except for Novelty and Usefulness, the assessments are below average. The average scores for Additional Information (along with Interestingness) are slightly above the rest at 2.60 (40 percent) with an expected mean with the lowest variation of all the criteria (0.168). One reason for the assessment of Additional Information is that its quality depends on how much information in one hypothesis is relevant to the target terms. However, because dominance conditions must be met, the relevance values aren't

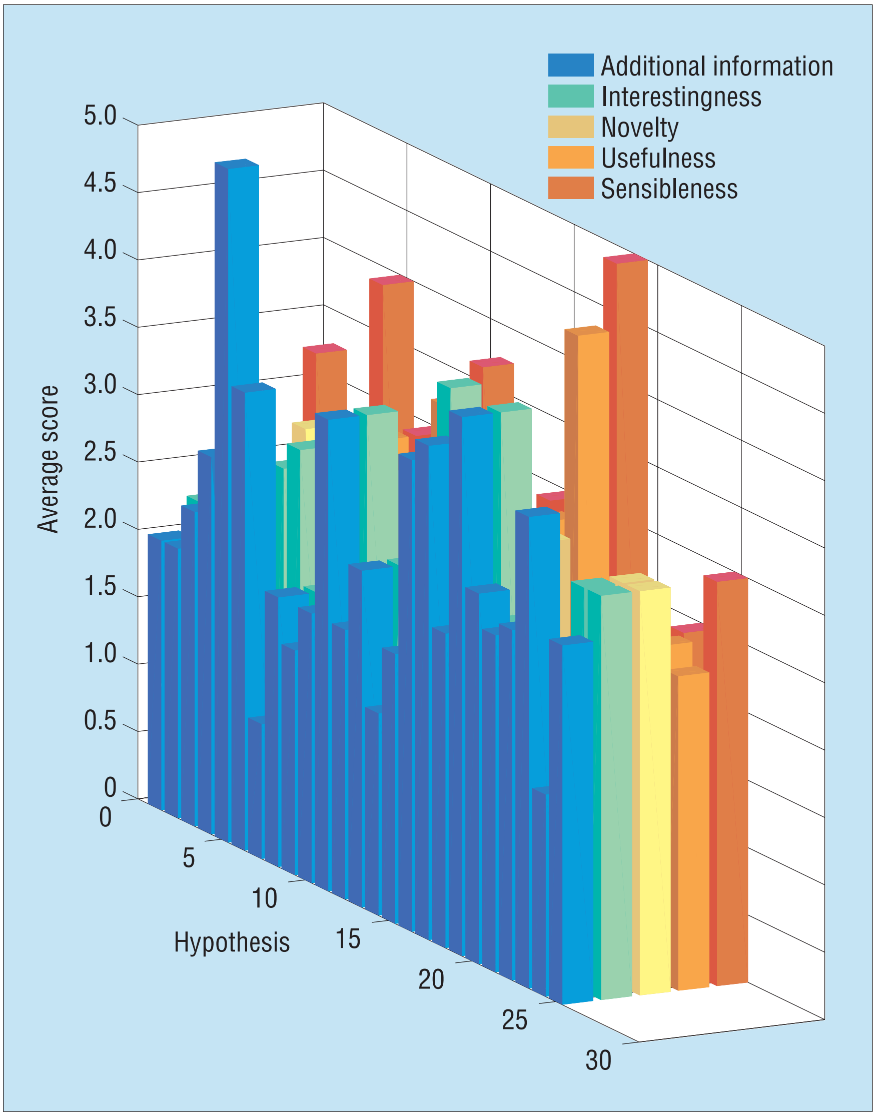


Figure 5. Experts' assessment of hypotheses.

always high. The semantic similarity also affects the contribution of a hypothesis to explain the relation between the concepts.

Indeed, although LSA can regard a hypothesis as very close to the terms, the similarity doesn't ensure that these terms will appear as they are. Some close semantic neighbor might appear instead. Details of the evaluation suggest that the experts made little effort to realize that, while in most cases the target concepts didn't appear, some close concepts try to make a hypothesis about the target concepts understandable. Although some hypotheses show scores below the average, several hypotheses look encouraging in that this is a demanding evaluation in terms of a very complex human task. The model shows promise in terms of finding (and filtering) good hypotheses rather than discovering overall high results, which might not be possible.

On the other hand, we were able to discern no direct evidence on how a human would perform on the same task. For such a complex task, humans might perform poorly when analyzing large amounts of text data. For the same reason, humans might not be able to find the hypotheses that the system found. To address this issue, we measured the correlation between the scores of the human subjects and the model evaluation. Because both the expert's 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. For the expert assessment, we averaged the scores of the different criteria for every hypothesis (values between 1 and 5).

For the system evaluation, we considered both the objective values and the fitness of every hypothesis to come up with a single value between 0 and 2. For visualization, we scaled these values up to the same range as the expert assessment (1 to 5). We then calculated the pair of values for every hypothesis and obtained a (Spearman) correlation $r = 0.43$ ($t - test = 18$, $df = 25$, $p < 0.001$). From this result, we see that the correlation indicates a promising level of correspondence between the system and the human judgments. On the other hand, it suggests that an expert wouldn't perform much better than the system for such a complex task with such a large amount of information.

We collected and summarized the domain expert information to produce factors identified during the experiment that, we believe, might explain somewhat the low scores in the assessment. In this closer analysis, some of the issues included comprehensibility, inconsistency between paragraphs, specificity of topics, domain-specific issues, and incomplete hypotheses. To show what the final hypotheses look like, we picked some of the average best and worst hypotheses as assessed by the experts (out of the 25 best hypotheses). Specifically, we highlighted two of the best hypotheses and one worst hypothesis. Because we do not have domain knowledge to analyze the content of these hypotheses, we provide brief descriptions of the predicates' argument to give a flavor of the knowledge involved:

Although some hypotheses show scores below the average, several hypotheses look encouraging in that this is a demanding evaluation in terms of a very complex human task.

Hypothesis 65 of Run 4. (Table 1 shows the target concepts used for the runs.) The hypothesis is represented by this rule:

IF goal(perform(19311)) and goal(analyze(20811))
THEN establish(111)

(Numerical values represent internal identifiers for the arguments and their semantic vectors.) This hypothesis has a criteria vector [0.92, 0.09, 0.50, 0.005, 0.7, 0.00, 0.30, 0.25] (the vector's elements represent the values for criteria relevance, structure, coherence, cohesion, interestingness, plausibility, coverage, and simplicity). It obtained an average expert's assessment of 3.74.

In natural language text, this rule can roughly be interpreted as

IF work aims at performing the genetic grouping of populations …
    AND to analyze 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 others (its relevance is 92 percent). However, this isn't complete in terms of cause and effect. For instance, the methods are missing.

Hypothesis 88 of Run 3. This is represented by this rule:

IF goal(present(11511)) and method(use(25511))]
THEN conclusion(effect(1931,1932))

The hypothesis has the criteria vector [0.29, 0.18, 0.41, 0.030, 0.28, 0.99, 0.30, 0.50] and obtained an average expert's assessment of 3.20. In natural language text, this can roughly be interpreted as

IF the goal is to present the forest restoration …
    AND the method is based on the use of microenvironments for capturing farm mice …
THEN digestibility "in vitro" should have an effect on the bigalta cuttings …

This hypothesis looks more complete (goal, methods, and so forth) but is less relevant than the previous hypothesis despite its close coherence (50 percent versus 41 percent). Also, the plausibility is much higher than for Hypothesis 65, but the other criteria seemed to be a key factor for the experts.

Hypothesis 52 of Run 5. This is represented by this rule:

IF object(perform(20611)) and
object(perform(2631))
THEN effect(1931,1932)

The hypothesis has a criteria vector [0.29, 0.48, 0.49, 0.014, 0.2, 0.00, 0.30, 0.50] and obtained an average expert's assessment of 1.53. In natural language text, it can roughly be interpreted as

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 …

The structure (48 percent) is better than for Hypothesis 88 (18 percent). However, because Hypothesis 52 isn't complete, it received a lower score than Hypothesis 88. This might be because the difference in structure between object-object and goal-method is not significant. Because both hypotheses became final solutions, the expert scored best on those that better explained the facts. Because the model relies on the training data, it doesn't ensure that every hypothesis is complete. In fact, training data show that only 26 percent of the 326 rules contain some sort of method.

Evaluating the model shows that it's plausible to produce new knowledge by combining shallow text analysis with evolutionary techniques without using external resources. Our work also contributes to the evaluation and assessment of quality criteria that most other approaches have neglected, by proposing new evaluation criteria to measure the plausibility of the hypotheses as they are produced.

In addition, our proposed semantic representation can help capture key information concerning the creation of hypotheses in a way that goes beyond the structural information in the rules. Our approach handles the problem of the diversity of solutions by having semantic and rhetorical constraints in mind. Unlike traditional methods, this helps deal with the underlying text knowledge without needing to perform further deep analyses.

On the evolutionary-learning side, further research must investigate how the time the evolutionary system spends affects the results' quality. We could have traded off some criteria to test how to improve the results. Other complex issues require more extensive experimental testing.

Overall, the model shows a good level of prediction in terms of its correlation with human judgments, which is comparable to or even better than related approaches.[5] However, the most outstanding feature is how our approach achieves this without using external resources. ■

The Authors

John Atkinson-Abutridy is an associate professor at the Departamento de Ingeniería Informática, Universidad de Concepción, Chile. His research interests include natural language processing, knowledge discovery from texts, artificial intelligence, and evolutionary computation. He received his PhD in artificial intelligence from the University of Edinburgh. He's a member of the ACM, AAAI, and Association for Computational Linguistics. Contact him at the Departamento de Ingeniería Informática, Universidad de Concepción, Concepción, Chile; atkinson@inf.udec.cl.

Chris Mellish holds a chair in the Department of Computing Science at the University of Aberdeen. His research interests include natural language processing and logic programming. He's especially interested in natural language generation. He received his PhD in artificial intelligence from the University of Edinburgh. Contact him at the Dept. of Computing Science, Univ. of Aberdeen, King's College, Aberdeen AB24 3UE, UK; cmellish@csd.abdn.ac.uk.

Stuart Aitken is a member of the University of Edinburgh's Artificial Intelligence Applications Institute. His research interests include ontology, bioinformatics, intelligent tools for knowledge acquisition, and machine learning. He received his PhD in engineering from the University of Glasgow. Contact him at the Artificial Intelligence Applications Inst., Univ. of Edinburgh, Appleton Tower, Room 4.10, Crichton St., Edinburgh EH8 9LE, UK; stuart@aiai.ed.ac.uk.

References

1. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.
2. D. Swanson, "On the Fragmentation of Knowledge, the Connection Explosion, and Assembling Other People's Ideas," *Bull. of the Am. Soc. for Information Science and Technology*, vol. 27, no. 3, Feb./Mar. 2001, pp. 12–14; www.asis.org/Bulletin/Mar-01/swanson.html.
3. M. Hearst, "Automated Discovery of WordNet Relations," *WordNet: An Electronic Lexical Database*, MIT Press, 1998, pp. 131–151.
4. C. Jacquemin, "Syntagmatic and Paradigmatic Representation of Terms Variation," *Proc. 37th Ann. Meeting Assoc. for Computational Linguistics*, Assoc. for Computational Linguistics, 1999, pp. 341–348.
5. S. Basu et al., "Using Lexical Knowledge to Evaluate the Novelty of Rules Mined from Text," *Proc. NAACL 2001 Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Assoc. for Computational Linguistics, 2001, pp. 144–149.
6. K. Deb, *Multiobjective Optimization Using Evolutionary Algorithms*, John Wiley & Sons, 2001.
7. S. Teufel and M. Moens, "Discourse-Level Argumentation in Scientific Articles: Human and Automatic Annotation," *Proc. ACL 1999 Workshop towards Standards and Tools for Discourse Tagging*, Assoc. for Computational Linguistics, 1999.
8. T. Landauer, P. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 10, no. 25, 1998, pp. 259–284.
9. W. Kintsch, "Predication," *Cognitive Science*, vol. 25, no. 2, 2001, pp. 173–202.
10. P. Foltz, W. Kintsch, and T. Landauer, "The Measurement of Textual Coherence with Latent Semantic Analysis," *Discourse Processes*, vol. 25, no. 2, 1998, pp. 259–284.
11. S. Basu et al., "Evaluating the Novelty of Text-Mined Rules Using Lexical Knowledge," *Proc. 7th Int'l Conf. Knowledge Discovery and Data Mining*, ACM Press, Aug. 2001, pp. 233–238.
12. E. Zitzler and L. Thiele, *An Evolutionary Algorithm for Multiobjective Optimisation: The Strength Pareto Approach*, tech. report 43, Swiss Federal Inst. of Technology (ETH), 1998.
13. M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, 1996.