Human-centric intelligent systems for exploration and knowledge discovery

I. C. Parmee

This speculative article discusses research and development relating to computational intelligence (CI) technologies comprising powerful machine-based search and exploration techniques that can generate, extract, process and present high-quality information from complex, poorly understood biotechnology domains. The integration and capture of user experiential knowledge within such CI systems in order to support and stimulate knowledge discovery and increase scientific and technological understanding is of particular interest. The manner in which appropriate user interaction can overcome problems relating to poor problem representation within systems utilising evolutionary computation (EC), machine-learning and software agent technologies is investigated. The objective is the development of user-centric intelligent systems that support an improving knowledge-base founded upon gradual problem re-definition and reformulation. Such an approach can overcome initial lack of understanding and associated uncertainty.

Introduction

Uncertainty and poor problem definition are inherent features during early stages of many problem-solving processes. Immediate requirements for relevant information to improve understanding can be confounded by complex problem descriptions comprising many interacting variable parameters. Problem constraints and multiple objectives that defy complete quantitative representation and therefore require a degree of subjective user evaluation can further inhibit meaningful progression. Indeed, problem representation in the first instance may be merely based upon qualitative mental models arising from experiential knowledge, group discussion and slight empiric investigation. However, such representations, coupled with user intuition, play a significant role in the identification of future direction and further investigation. Initial concepts and hypotheses based upon current understanding require exploration in a breadth-first manner to generate relevant information that supports and enables meaningful progress.

Compound design perhaps presents a typical example where the chemist is faced with a problem of such magnitude in terms of the number of possible solutions that finding an appropriate starting point upon which to base empiric study is a major task involving extensive experiential knowledge, skill and intuition. Although some computational representations may be available to provide an indication of performance of, say, reagent combinations against specific criteria, a degree of uncertainty with regard to the fidelity of their output is generally inherent. Hence the need for human evaluation to eliminate poor reagent combinations that have survived machine-based evaluation whilst identifying high potential combinations for further empiric investigation. Due to the number of possible combinations across multiple reagent libraries some form of computational search and exploration capability is essential to identify potential high performance solutions for further evaluation by the chemist.[1] Thus a machine/human procedure could ensure that experimental effort is concentrated upon 'best' candidates thereby significantly reducing design lead time. The above example is used in the paper to aid understanding of the proposed speculative approaches. Given the potential in the compound design domain it is apparent that the development of similar human/computer based search, exploration and classification capabilities would also be of significant benefit in other biotechnology domains. The analysis of data sets from gene expression experiments to provide insights into gene activity under differing environmental conditions and the identification of gene regulatory networks is another area currently receiving attention.[2]

Problem redefinition and reformulation

Generally, the development of machine-based representations can support exploration through the evaluation of identified combinations against criteria perceived to be relevant. Initial representations may comprise simple rule sets coupled with basic statistical models generated from any available relevant data. Although the degree of user-confidence in the output of each criteria representation may vary considerably such representations can provide essential problem insight despite their apparent shortfalls. Seemingly high performance solutions identified in terms of quantitative criteria followed by qualitative human evaluation utilising experiential knowledge and intuition provides an indication of the viability of concepts and hypotheses and of the fidelity of the initial computational representations. An iterative user/machine-based exploratory process can commence where gradual

improvements in understanding contribute to better representations, a developing knowledge-base and the eventual establishment of computational models that support more rigorous analysis. A highly interactive process thus emerges supporting the development of representation through knowledge discovery. Such a human/machine-based development may run concurrently with, and be enhanced by, conventional empiric investigation and other forms of data/information gathering.

The above could be considered a general description of how we progress when faced with problems that initially seem beyond our perceived analytic capabilities. Using this description the following sections explore the human-centric utilisation of evolutionary computation, machine learning and agent-based approaches integrated with enabling computational technologies to significantly enhance this iterative, knowledge discovery and representation development process. Particular areas requiring attention are:

• the development of meaningful computational representations from experiential knowledge, sparse data and collective reasoning;

• non-linear search and exploration processes that can negotiate the complex solution spaces described by such representations (where the solution space is described by all possible combinations of variables *e.g.* reagent libraries);

• the capture of user experiential knowledge and intuition during re-definition of machine-based representations and reformulation and subsequent exploration of innovative solution spaces;

• development of software agent-based activities for information extraction, processing and succinct presentation to the user resulting in a reduction in cognitive load.

The overall objective is the establishment of user-interactive computationally intelligent search and exploration environments that support rapid concept and hypothesis formulation, exploration and evaluation. Novel human-centred problem-solving processes integrated with such 'virtual laboratories' may lead to innovation and scientific breakthrough within an academic research environment whilst supporting competitive product development through continuous knowledge discovery in an industrial context.

The author has been actively researching the development and integration of such user-centric CI systems primarily in the field of conceptual engineering design[3,4] Recent involvement with pharmaceutical and biotechnology design through close collaboration with Evotec OAI, Milton Park, Abingdon, UK (http://www.oai.co.uk/) indicates a very real potential for the integration of similar systems with these areas.

Search and exploration in complex space

Concepts relating to multi-dimensional (*i.e.* multi-variable) search and exploration require description to clarify terminology. A *search space* comprises the set of all possible solutions described by combinations of the problem's parameters (*e.g.* reagents within selected reagent libraries). *Dimensionality* of the space relates to the number of variables (*e.g.* reagent libraries) involved. The size of the search space relates to the dimensionality and the number of values relating to each variable. For instance two reagent libraries each containing two hundred reagents would describe a search space of 40,000 possible combinations. Adding a third reagent library comprising a further 200 reagents results in a space containing 8,000,000 possible combinations. This non-linear relationship creates very large search spaces even when considering relatively small numbers of variables. Such spaces can be visualised as *multi-dimensional surfaces* or *landscapes*. For instance, imagine a problem comprising two continuous (real number) variables represented along orthogonal horizontal axis with a vertical axis relating to the relative performance of combinations of these variables in terms of some criteria. A three-dimensional landscape representing all possible solutions at a given resolution can then be plotted and visualised as shown in Fig. 1.

Obviously, a problem defined by larger numbers of variables rapidly becomes impossible to visualize or imagine in terms of the resulting high dimensional landscape. The *complexity* of this high-dimensional landscape is significantly increased when integer variables or when complex mixes of integer and continuous variables are involved in the problem description. In addition, the presence of local optima representing best values relating to the criteria under consideration will create a *multi-modal* surface (region A in Fig. 1) comprising peaks/troughs upon which any form of search
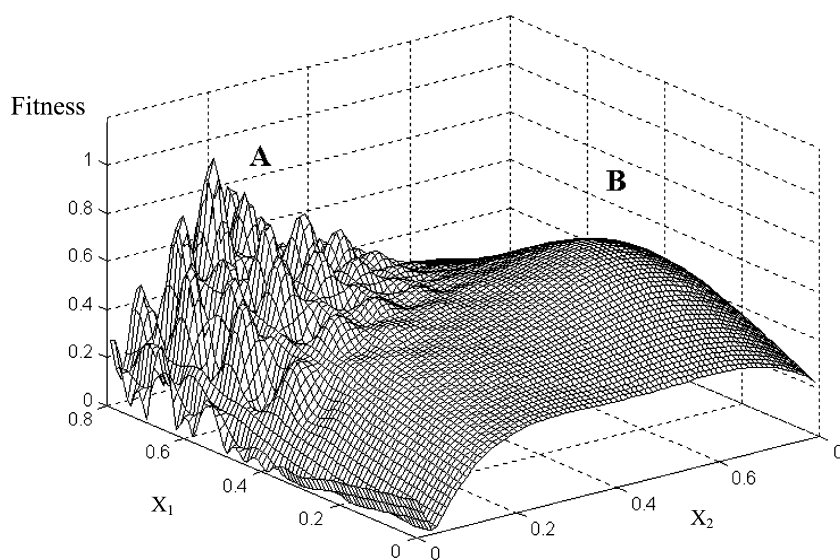

Fig. 1 Search space/fitness landscape described by values of two continuous variable parameters (X_1, X_2) with solution performance indicated on the vertical axis. Regon A: multi-modal characteristics. Region B: unimodal characteristics.

and optimization procedure may prematurely converge. Region B is unimodal *i.e.* only one peak is evident within the region. This offers a far lesser challenge as a variety of gradient-based optimizers would rapidly converge upon the optimal point. The surface described by the reagent combination example would likely be extremely rugged with many local optima due to the random ordering of reagents within the reagent libraries.

Various *constraint*s (*e.g.* maximum allowable weight) will create infeasible regions of a surface and these regions may be convoluted and disjoint. Several quantitative criteria may be involved (*e.g.* similarity, QSAR, docking *etc.*) introducing varying degrees of conflict and creating landscapes with differing characteristics relating to a particular criteria. This introduces a requirement to search for common regions of these landscapes offering best compromise solutions.

Searching the two-dimensional landscape in Fig. 1 presents little problem. Given existing computational capability a coarse exhaustive search of solutions would be possible. An increase in the number of variables and the inclusion of the additional complexities described above presents a far greater challenge. An exhaustive search becomes non-viable and the investigator has to rely upon sophisticated search, exploration and optimisation algorithms.

If the problem is well-defined in terms of variables, constraints and objectives that are quantifiable and of known relative importance then a range of search and optimisation techniques can be utilised that can handle the above complexities to a varying degree. Modern heuristic techniques involving populations of trial solutions and stochastic operators that promote search and exploration and eventual convergence are particularly well-suited to the negotiation of such complex problem spaces. The term evolutionary computation tends to cover techniques such as these and perhaps the genetic algorithm (GA)[5] has become the best known.

In most cases, however, high problem definition is characteristic of the latter stages of a problem-solving process. These final stages may represent the tip of the iceberg given the time and effort involved in initial problem understanding, definition, formulation and representation. During early stages a high degree of assumption, particularly relating to objective representation, generally provides a starting point for our investigations. An initial variable set may be selected with later addition or removal of variables as the sensitivity of the problem to various aspects becomes apparent. Constraints may be treated in the same way with the added option of softening them to allow exploration of non-feasible regions. Included objectives may change as significant payback becomes apparent through a re-ordering of objective preferences. Some non-conflicting objectives may merge whilst difficulties relating to others may require serious re-thinking with regard to problem formulation. The initial problem space is therefore a moving feast rich in information which, when extracted and coupled with the investigators' experiential knowledge and intuition supports significant problem insight and subsequent problem re-formulation. It is quite possible that final solutions will be identified from a space that bears little resemblance to the search space that provided a starting point for our investigations.

We are, perhaps, considering two problem search spaces:

(1) The machine-based quantitative space that is bounded and inflexible when considered stand-alone (*i.e.* the space defined by reagent libraries within a compound design situation). Search and exploration algorithms utilizing machine-based criteria representations to evaluate solutions can rapidly provide novel information from this space that aids problem understanding at a human level. Such understanding and subsequent search space redefinition can remove the initial bounds.

(2) The investigators' mental representations of the problem. These representations are only bounded by current knowledge and understanding. The development of this problem space relies upon external stimuli and human intuition and judgement at both a quantitative and qualitative level.

The indication from previous conceptual design work in the engineering domain is that the appropriate melding of these two spaces will support a holistic, knowledge-based approach that can result in significant step changes to machine-based objective representation and in scientific/technological understanding.

Problem reformulation

The concept of problem formulation and reformulation is well established within the engineering design research community especially when considering innovative and creative design.[6,7] This is associated with the development of a designer's understanding of a problem during the early investigative stages that may result in radical changes in problem representation. Another concept relates to the integration of knowledge from other sources through, say, analogical or metaphorical transfer from another problem area. This can be of significant benefit especially with regard to the development of innovative approaches. Also, much attention has been paid in the design research area to holistic aspects of the design process and the manner in which initial qualitative modelling of a problem domain eventually translates into more definitive representations.

Many design research concepts map well onto generic problem-solving and decision-making processes where complexity, high-dimensionality and the inability of the user to concurrently cope with many dimensions of information (cognitive overload) obstructs progress and inhibits exploration. Computational intelligence techniques relevant to and developed within the design domain are now reaching a level of sophistication that allows them to be utilised to support a more holistic approach to problem solving.[8] Machine-based exploratory systems can better handle the complexities of high-dimensional space ensuring that succinct specific information is available to the investigator thus enabling a greater user-concentration upon the significance of emerging results.

Existing appropriate CI and enabling technologies

Existing computational intelligence technologies contribute in a piecewise manner to the development of user-interactive CI systems that meld

machine-based problem processing and user problem solving. The flexibility of stochastic search and exploration processes provided by evolutionary computation (EC)[9,10] means that complex search spaces described by mathematical, statistical, boolean, neural network or fuzzy inference models can be efficiently and rapidly negotiated. This flexibility allows the integration of simple rule-based models to provide initial problem representation and the subsequent development and integration of fuzzy, neural and/or statistical models[11,12] for the evaluation of solutions as data and knowledge accumulates. Cluster-oriented genetic algorithms[13,14] can support information extraction and multi-objective evolutionary computation approaches can handle multiple quantitative criteria. On-line user-centric criteria ranking capabilities can be achieved *via* fuzzy preference techniques.[14,15] The integration of qualitative criteria is implicitly supported during problem re-definition and re-formulation through the influence of user experiential knowledge. Software Agent technologies[17] can support information extraction in terms of the identification of interesting solutions and support background processing of data and subsequent meaningful presentation of results to the user. Multi-agent systems can provide a machine-based negotiating capability that may assist in solution identification and selection, qualitative objective satisfaction and the processing of search direction alternatives.[18] A reduction in perceived problem complexity through machine-based processing of extracted data and the subsequent presentation of succinct information may thus relieve cognitive load upon the investigator. The processing of relatively mundane tasks can be readily achieved *via* single function agents thus allowing the user to concentrate more upon emerging aspects of interest. In the longer term it is also possible that embedded machine learning processes training upon extracted information may enable some degree of autonomous agent-based activity.

The introduction of supporting and enabling technologies such as state-of-the-art data visualisation techniques and high-performance computing (HPC)

would result in interactive CI search and exploration systems where the user becomes immersed within an information-rich computing environment accepting and analysing output and introducing change. High-performance computing capabilities would be essential to achieve a seamless interface between interactive processes. On-line data-mining techniques[11] coupled with agent-assisted data processing and visualisation would contribute greatly to the immersion concept. Overall integration with e-Science technology could lead to the establishment of Grid-based search and exploration capabilities widely available to the UK research community whilst also enabling remote access to very significant HPC resources and possibly to diverse information sources that enhance current knowledge of the problem at hand.

The establishment of a seamless user/machine-based information generation environment as described is ambitious. However highly efficient search across changing fitness landscapes with varying objective preferences and changing constraint conditions is achievable. It is also possible to spawn

concurrent/complementary local search utilising appropriate algorithms. Constraint-handling techniques can be introduced that allow exploration and information extraction relating to constraint sensitivity. Search space sampling techniques can be integrated with exploration processes to rapidly generate concepts of problem complexity as landscapes change. Statistical and CI-based modelling techniques are now available whereas the concurrent utilisation of differing model types to provide better overall representation and increased confidence is accepted practice in some areas.

A possible configuration of the various system components and of user interactivity is simply illustrated in Fig. 2.

The virtual laboratory

Taking all of the above into account the concept of a virtual laboratory begins to emerge. Imagine developing relatively basic machine-based criteria representations from current data and constitutive theories associated with current understanding and then being able to rapidly explore the multi-variate space described
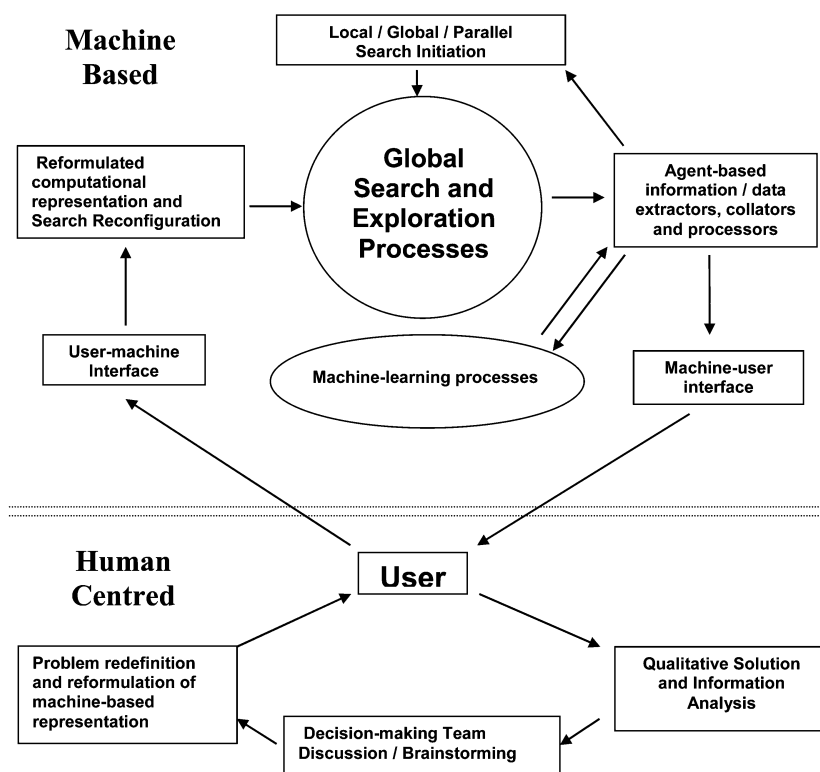

Fig. 2 Simple illustration of a possible system configuration.

by these representations using combinations of evolutionary algorithms and other local and global search techniques. As search progresses the system is extracting and accumulating information relating to complex characteristics of the problem domain whilst also discovering viable solutions. Solutions are initially identified that best satisfy objectives/constraints seemingly relevant in terms of current understanding whilst background processes extract information from areas of the problem space previously visited and present this, in a succinct manner, on-screen to the user. The degree of difficulty of satisfying initial objectives within existing variable bounds or within existing objective preference ranking becomes quantifiable and presentable through background data processing as search progresses. On-line user actions such as constraint softening, objective preference variation or modification of variable ranges may change the nature of the space and search direction whilst machine-based software agents acting as information collators, processors and presenters provide indications of the effects of such changes. These agents constantly advise the user on interesting solution correlation or re-direct you to previously visited areas now of more interest. Concurrent, finer-grained, localised search processes may be spawned to explore specific regions. These actions become semi-autonomous as, through a machine-learning capability, the agents become more 'aware' of your requirements. The environment becomes more immersive as the user reacts to the information being presented and user on-line actions become an integral part of the exploration process negotiating this high dimensional space reacting to feedback from the system to make iterative changes to the problem landscape.

At any point this relatively continuous exploration process can be paused and relevant information downloaded and presented to the decision-making team for discussion. An easily understood graphic provides a recorded history of user-instigated change thereby supporting traceability and allowing analysis of the logical progression of the team's thinking based upon extracted information. The presentation of such material promotes discussion and allows the

perspectives of others to be integrated in further exploratory interactive activity *via* appropriate problem re-definition and re-formulation.

As this iterative interactive process continues so confidence in the developing criteria representations increases, the knowledge-base becomes well-founded and uncertainty significantly decreases. A natural result is a reduction in user-interaction as we move from a high-risk problem definition phase through an intermediate phase of increasing confidence to the final stages of detailed analysis of a well-defined problem space. This could be considered analogous to the conceptual, embodiment and detailed stages of engineering design.[19]

Conclusions

There is obviously much further research required to achieve the goal of the seamless user-centric system described above. However, many of the component parts are at a stage of development where their collective utilisation is possible and current research is pushing hard towards achieving this. Problems specific to the biotechnology field will necessitate appropriate development of any working system based upon these concepts. It is suggested that the flexibility of CI technologies is such that specific problems are unlikely to be insurmountable. For instance, although a machine-based representation of an evaluation function may cause problems the user-centric approach supports complete or partial human evaluation of solutions and this can initially play an integral role.

Both research council and industrial funding plus close interdisciplinary working will be required to resolve arising problems. From an industrial point of view, however, user-centric CI search and exploration systems could best utilise seemingly endless increases in desktop computational processing capability especially considering that in-house networked machines potentially support access to very high levels of distributed computing power. Such systems continuously running as background processes could support the development of in-house knowledge and expertise whilst reducing lead times to

the discovery of innovative products when allied with complementary investigative processes.

From a more academic research-oriented point of view the further development and utilisation of such systems within a research environment could support significant leaps in understanding relating to the characteristics of poorly defined complex problem space. The ability to rapidly and efficiently play 'what-if' whilst concurrently gathering high-quality information that either confirms or contradicts current thinking suggests an environment well-suited to the support of knowledge discovery and scientific breakthrough. The role of human intuition, experience and judgement within such an environment would be paramount whilst the inherent support of agent-based entities in terms of information processing and presentation would be invaluable.

I. C. Parmee
Advanced Computation in Design and Decision-making CEMS, University of the West of England, Bristol, UK.
E-mail: ian.parmee@uwe.ac.uk

References

1 V. J. Gillet, W. Khatib, P. Willet, P. J. Fleming and D. V. Green, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 375–385.
2 C. Spieth, F. Streichert, N. Sper and A. Zell, *Proceedings IEEE Congress on Evolutionary Computation*, 2004, vol. 1, pp. 152–157.
3 I. C. Parmee, C. Cvetkovic, A. H. Watson and C. R. Bonham, *Evol. Comput.*, 2000, **8**, 2, 197–222.
4 I. C. Parmee, *Artif. Intell. Eng. Des., Anal. Manuf. J.*, 2002, **16**, 3.
5 D. E. Goldberg, *Genetic Algorithms in Search, Optimisation & Machine Learning*, Addison-Wesley Publishing Co., Reading, Massachusetts, 1989.
6 A. K. Goel, *IEEE Exp., Intell. Syst. Appl.*, 1997, **12**, 3, 62–70.
7 N. P. Su, *The Principles of Design*, Oxford University Press, New York, 1990.
8 I. C. Parmee, *Evolutionary and Adaptive Computing in Engineering Design*, Springer Verlag, London, 2001.
9 *Handbook of Evolutionary Computation*, ed. T. Bäck, D. B. Fogel and Z. Michalewicz, IOP Publishing/Oxford University Press, Oxford, 1997.
10 *New Ideas in Optimization*, ed. D. Corne, M. Dorigo and F. Glover, McGraw Hill, London, 1999.
11 T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer-Verlag, Berlin, 2001.

12 R. O. Duda, P. E. Hart and D. E. Stork, *Pattern classification*, John Wiley, London, 2nd edn., 2001.

13 I. C. Parmee and C. R. Bonham, *Artif. Intell. Eng. Des., Anal. Manuf. J.*, 1999, **14**, 3–16.

14 I. C Parmee and J. R. Abraham, *IEEE Congress on Evolutionary Computation*, Portland, USA, 2004, pp. 395–402.

15 J. Fodor and M. Roubens, *System Theory, Knowledge Engineering and Problem Solving*, Kluwer Academic Publishers, 1994, vol. 14.

16 D. Cvetkovic and I. C. Parmee, *IEEE Trans. Evol. Comput.*, 2001, **6**, 42–57.

17 M. J. Woodridge and N. R. Jennings, *Knowl. Eng. Rev.*, 1995, **10**, 2, 115–152.

18 D. Cvetkovic and I. C. Parmee, *Artif. Intell. Eng. Des., Anal. Manuf. J*, 2003, in press.

19 G. Pahl and W. Beitz, *Engineering Design. Design Council*, Springer-Verlag, London, 1984.