Criticality-Based Task Composition in Distributed Bioinformatics Systems

Konstantinos A. Karasavvas[1], *Richard Baldock*[2] *and*
Albert Burger[1, 2]

[1]*School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK and* [2]*Human Genetics Unit, Medical Research Council, Edinburgh, EH4 2XU, UK*

ABSTRACT

During task composition, such as can be found in distributed query processing, workflow systems and AI planning, decisions have to be made by the system and possibly by users with respect to how a given problem should be solved. Although there is often more than one correct way of solving a given problem, these multiple solutions do not necessarily lead to the same result. Some researchers are addressing this problem by providing data provenance information. In this paper we propose an approach that assesses the importance of such decisions with respect to the overall result. We present a way of measuring decision criticality and describe its potential use. Real bioinformatics examples are used to illustrate the approach.

INTRODUCTION

To solve a complex problem, systems typically break it down into more managable smaller problems, which are then executed following some partial ordering. In general, there is a tradeoff between protecting the user from unnecessary details of these composition activities, i.e. providing a high level of transparency, and giving him/her some control over the answer finding process.

In many cases there is more than one way to solve a particular problem. For example, what scoring matrix should be used for a sequence comparison, and what database should be used to find tissue-specific gene expression data. We use the term *decision point* to refer to situations where such choices exist. Providing some measure of the *criticality* of a decision point can aid the system, as well as the biologist in pruning the possible solution space for any given problem.

For such a measure we compare the results acquired after making one choice to the results acquired from the other available choices of the decision point. Alternative choices are executed according to user preferences and available computational resources. The higher the differentiation between the result sets over a number of similar queries, the more critical that decision point becomes.

The basic concept of decision points and their criticality is applicable to a variety of technologies, such as distributed query composition, workflow composition and hierarchical task network (HTN) planning.

We have implemented and tested a prototype system to demonstrate the ideas mentioned above. It is a multi-agent bioinformatics system integrating gene expression resources for mouse. It comprises GXD (Ringwald *et al.*, 2001), a mouse gene expression database, EMAGE (Davidson *et al.*, 1997), a mouse gene expression database with mappings to a mouse embryo 3D model, and BLAST (Altschul *et al.*, 1990) at NCBI[†], an Internet sequence search tool.

Other query or workflow composition systems, like Geodise (Chen *et al.*, 2002), use domain knowledge to guide the user during the composition process. Our approach, in addition to domain knowledge uses an independent measure, criticality, which is based on the actual data of the resources involved in the composition.

The remainder of the paper is organised as follows. A brief overview of the multi-agent system is given in the next section. Then, decision criticality is explained in more detail, while the following section describes the experiments that were carried out. The last section summarises the paper.

SYSTEM

Our system is a purely communicative multi-agent system: there is no external environmental influence and the agents communicate only by means of messages. The system is based on the FIPA specifications, and a FIPA-compliant development tool, JADE, is used for implementation. Messages exchanged between agents are formed in a high-level language, FIPA Agent Communication Language (ACL), and the ACL content language is SL0—a subset of the FIPA suggested Semantic Language (SL). In turn, the SL0 content conforms to specified ontologies.

In our integration system the user is allowed to take

[†] http://www.ncbi.nlm.nih.gov/BLAST

control of a decision point as it occurs and if certain conditions customised by the user beforehand are met. To help the user make his/her choice, an explanation facility is provided that explains known benefits and/or drawbacks of the available choices.

To achieve the kind of functionality already mentioned we developed a framework to allow the user flexible control over the decision points. To this end, two new ontologies were defined: a system ontology and an application domain one, and a new agent protocol specified (to properly orchestrate the agent interactions). The general architecture in terms of the agents of the system consists of five types of agents: ‘User Agents’, ‘Mediation Agents’, ‘Resource Agents’, ‘Comparison Agents’, and ‘Explanation and Interaction Agents’.

The last two are particularly important for the functionality mentioned in this paper. ’Explanation and Interaction Agent’ is responsible for facilitating user interaction, guiding the user for making his/her choice, as well as providing an explanation of the choices made during the composition/planning process. ’Comparison Agent’ is responsible for calculating criticality, keeping track of the data/statistics collected thus far, and responding to requests for data analysis and statistics. For more information about the system’s design and architecture see (Karasavvas *et al.*, 2002).

DECISION CRITICALITY

Criticality of a decision point is calculated by comparing the results received after execution of two or more possible choices. Intuitively, to measure the similarity of two sets we take their intersection (common elements) and divide the resulting set’s cardinality by the average cardinality of the two sets; for sets A_1 and A_2 their similarity s_{12} is: $s_{12} = \frac{|A_1 \cap A_2|}{\frac{|A_1|+|A_2|}{2}} = 2 \cdot \frac{|A_1 \cap A_2|}{|A_1|+|A_2|}$.

In the case of more than two sets we need to make all possible comparisons in pairs. A result of k sets consists of $n = k(k-1)/2$ pairs, and thus will make that many comparisons. Then, $\overline{s_k} = \frac{1}{n}\left(\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} s_{ij}\right)$ calculates the mean similarity of k sets.

Now we also have the mean differentiation of the k sets, $\overline{d_k} = 1 - \overline{s_k}$. Finally, we can express both similarity and differentiation means as percentages with $\overline{S_k} = \overline{s_k} \cdot 100$ and $\overline{D_k} = \overline{d_k} \cdot 100$ respectively.

Of course $\overline{D_k}$ reflects the differentiation of a decision point over only one query. While this is useful for analysing the possible fluctuation on a specific query, we would also like to have a general differentiation measure for a decision point over the history of all queries. We call this criticality of a decision d, and it is simply the mean: $\overline{C_d} = \frac{1}{h}\sum_{i=1}^{h}\overline{D_k^i}$, where h is the history—number of queries—of that decision point.

So far we measured the criticality of one decision point. A query usually contains more than one such decision point. More formally, a query Q can be said to consist of a set of decision points—each one with its own local criticality—that influence its global (overall) criticality: $Q : (DP_1, DP_2, \ldots, DP_n)$. To specify dependencies between decision points we use ‘$\rightarrow$’, as in: $Q : (DP_1 \rightarrow DP_2)$, where DP_2 needs data acquired after DP_1 is resolved.

EXPERIMENTS

The experimental results presented here are based on the following query: “Find the mouse tissues that express the genes which match the given protein sequence.”. This query can be decomposed into two sub-queries:

(a) Which mouse genes correspond to the protein sequence given as input, and

(b) Which mouse tissues express these genes at a particular developmental stage.

We tested around 50 queries where the input of each was a short protein sequence randomly taken from known existing genes.

In our integration system we use an online BLAST sequence tool for sub-query (a). There are many parameters to be considered when querying BLAST, e.g. sequence database, scoring matrix, gap costs, etc. Each parameter could constitute another decision point during the task composition process. For this experiment we examine only one, the scoring matrix, and more specifically we only consider two matrix choices, BLOSUM62 and PAM70.

For sub-query (b) we use two different gene-expression resources, GXD and EMAGE. Thus, the decision to be made is which of the two databases should be used to acquire the final results.

In each sub-query, either automatically (by the system), or interactively (by the user) one choice would be selected and followed to acquire the final results—mouse tissues for this query. For example, matrix BLOSUM62 could be used to get the mouse genes and the GXD resource to get the mouse tissues. A different combination may provide different results, both intermediate and final. Indeed, comparing the results would enable us to calculate their differences and consequently how critical a decision is.

Our example query can be written as: $Q_e : (DP_1 \rightarrow DP_2)$, where DP_1 and DP_2 are the decision points for the scoring matrix and the gene expression database respectively, and the arrow specifies the dependence.

The left diagram in Figure 1 illustrates the differentiation based on the choice of using either BLOSUM62 or PAM70 (DP_1) by comparing the sets of genes returned

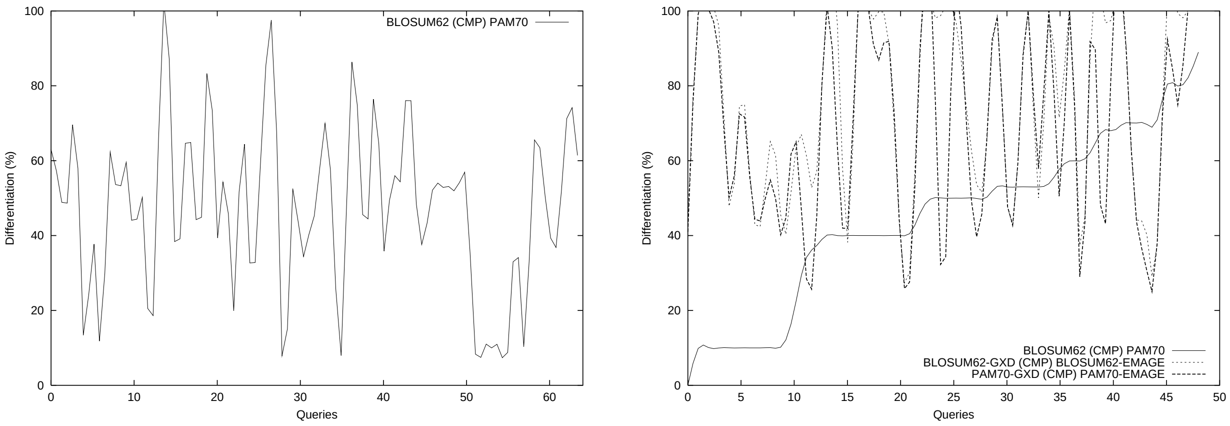

Fig. 1. Left: Differentiation between BLOSUM62 and PAM70. Right: Differentiation between BLOSUM62 and PAM70 (sorted) and differentiation of final results.

for each query. In brief, for some of the given protein sequences (represented by the queries along the x-axis), it mattered greatly which matrix was used, while for others it didn't. One can conclude that given in isolation, this decision point should be given some attention.

On the right diagram we can again show the differentiation of the mouse gene result sets acquired from BLOSUM62 and PAM70, but this time they are sorted in ascending order. For each result set, the second sub-query has been executed to find the appropriate tissues. Here the choice (DP_2) was between the two gene expression databases. For the mouse genes returned from the BLOSUM62 choice we acquired results from both the GXD and EMAGE databases and then compared the results (see normal dashed line in the diagram). Similarly, we compared the tissue results acquired from GXD and EMAGE using the PAM70 genes result set as input (see the bold dashed line in the diagram).

Comparing the two dashed lines, we observe only minor differences, no matter how high the matrix criticality is. The results suggest that even though the local criticality of the matrix choice is quite high, i.e. result sets are quite different, when used to further query for the mouse tissues the impact of this difference is less significant, i.e. its global criticality (with respect to the given query) is relatively low. Hence, even though DP_2 depends on the results of DP_1, the latter's high local criticality does not influence the final results significantly. This suggests that, for the given example, most of the genes that are expressed in mouse were found by both, BLOSUM62 and PAM70.

CONCLUSIONS

Bioinformatics integration systems need facilities that can help users to better understand and control the consequences of certain choices made during task composition. We have introduced the concepts of *decision points* and their *local* and *global criticality*. A concrete Bioinformatics example was used to show how certain decisions may not be critical in the context of more complex tasks. The work was carried out using a prototype multi-agent system for mouse gene expression research, developed at the MRC Human Genetics Unit in Edinburgh. Future work includes a widening of the query examples, a more comprehensive assessment of the criticality functions, and a systematic evaluation of the approach by biologists.

REFERENCES

Ringwald,M. (2001) The mouse gene expression database (GXD). *Nucleic Acids Research*, **29**, 98–101.

Davidson,D. (1997) The mouse atlas and graphical gene-expression database. *Seminars in Cell and Developmental Biology*, **8(5)**, 509–517.

Altschul,S.F. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Karasavvas,K. (2002) A multi-agent bioinformatics integration system with adjustable autonomy. *Lecture Notes in Computer Science*, **2417**, 492–501.

Chen,L. (2003) Towards a Knowledge-based Approach to Semantic Service Composition. *Lecture Notes in Computer Science*, **2870**, 319–334.