3 Satisfação de Restrições

Este capítulo define inicialmente o conceito de *problema de satisfação de restrições* (RPR). Em seguida, indica como modelar o problema de escalonamento como um *problema de satisfação de restrições temporais* (PSRT). Por fim, indica como representar uma instância do PSRT sob forma de um grafo.

Este capítulo prepara portanto a discussão do capítulo seguinte que focaliza a questão de planejamento e escalonamento com restrições temporais.

3.1 O Problema de satisfação de restrições (PSR)

O *problema de satisfação de restrições* (PSR) consiste num conjunto de variáveis $\boldsymbol{Z}= \{x_1,\ldots x_n\}$, com domínio discreto e finito $\boldsymbol{D}=\{D_1,\ldots D_n\}$, e um conjunto de $\boldsymbol{m}$ restrições $\boldsymbol{C}=\{c_1,\ldots c_m\}$, que são predicados $ck(x_i,\ldots x_j)$ definidos sobre o produto Cartesiano $D_i\times\ldots\times D_j$. Se $c_k(x_i=v_i,\ldots x_i =v_j)$ é certo, podemos dizer que a avaliação das variáveis é *consistente* em relação a c_k , e c_k se *satisfaz.* Uma solução é uma atribuição de valor a cada variável, em seu respectivo domínio, tal que todas as restrições sejam satisfeitas. Uma instância de um PSR (***Z,D,C***), pode ser representada como um grafo de restrições (ou como uma rede de restrições) G={***V,E***}. Para cada variável *v* em ***Z***, existe um nó *n* em ***V***. Para cada conjunto de variáveis conectadas por uma restrição *c* em ***C***, teremos um hiperarco *e* correspondente de ***E***, (ou arco binário se as restrições forem binárias). O problema de escalonamento pode portanto ser visto como um caso especial do PSR [Fox, 1984].

Como exemplo, podem ser mencionadas as seguintes aplicações:

- Hubble Space Telescope Scheduler (HSTS) [Muscettola, 1994], onde o planejamento e o escalonamento são integrados e o domínio do problema é dividido num conjunto de variáveis de estado. As variáveis de estado representam os componentes do domínio de trabalho, os quais assumem um conjunto de valores no tempo, cada um deles como um intervalo de tempo.

A integração entre o planejamento e o escalonamento consiste em encontrar o conjunto de evoluções temporais, ou seja, a seqüência de intervalos assumidos pelas variáveis de estado.

- IxTET [Ghallab, 1994], que utiliza a álgebra de intervalos, onde dado um problema e um conjunto de tarefas o sistema irá gerar uma solução através do refinamento sucessivo do problema inicial. O plano parcial gerado é composto de um conjunto de proposições temporais (eventos, ou utilização de recursos), um mapa temporal (para o gerenciamento das restrições temporais) e uma tabela de variáveis (para o gerenciamento das restrições).

Um PSR pode ser resolvido através da instanciação sistemática das variáveis e obtendo o conjunto solução das restrições. O custo computacional envolvido é exponencial [Dechter, 1991]. Por isso, é comum serem utilizadas técnicas para reduzir o custo computacional, basicamente isto pode ser feito minimizando a quantidade de *backtrackings* (heurísticas de decisão), realizando um pré-processamento da rede, verificando a consistência para reduzir o domínio [Sadeh, 1996, Belhadji, 1998, etc.]. No processo de busca da solução, no caso de uma solução, uma instanciação parcial não cumpre alguma das restrições e assim não é possível encontrarmos os valores a serem instanciados, sendo necessário a realização de um *backtracking*, para um ponto anterior eliminando assim uma parte do espaço de busca. A ordenação adequada das variáveis levará a "eliminação", e a um novo ramo na árvore de busca. Este procedimento no entanto pode levar ao aumento da complexidade devido ao grande número de vezes que é necessário realizar voltar a traz.

Na literatura, encontramos diferentes formas de modelar o problema de escalonamento como um PSR. [Sadeh, 1989], [Sadeh, 1991] e [Sadeh, 1995] apresentam uma solução do problema de encontrar um conjunto consistente de tempos de início para cada operação, onde uma aproximação é implementada, implicando em operações de escalonamento com janelas temporais pré-definidas, com tempos de começo mais cedo e mais tarde. [Caseau, 1994] introduz o conceito de intervalo de tarefas, que deve ser consistente com todas as operações que podem ser seqüenciadas entre o tempo mais cedo de começo e o mais tarde de término de uma operação.

Um problema semelhante consiste em estabelecer relações de ordem entre pares de operações que necessitam utilizar o mesmo recurso. [Cheng, 1997] propõe uma busca com backtracking, em que a solução é obtida mediante a a ordenação de pares de tarefas seqüenciadas. Este procedimento poda o espaço de busca e combina o uso de heurísticas de ordenação de valores com variáveis.

Ambas as abordagens assumem que são conhecidas de antemão todas as restrições relacionadas ao problema. Isto implica que, em sistemas dinâmicos, ou quando temos a necessidade de utilizar em conjunto planejamento e escalonamento, esta abordagem não é adequada, mesmo que estejamos considerando abordagens básicas, tais como duração, precedência, tempo de início e término dos trabalhos e disjunções.

3.2 O Problema de satisfação de restrições temporais (PSRT)

O problema de escalonamento pode ser enquadrado como um caso particular de um *problema de satisfação de restrições temporais* (PSRT), formulado por [Dechter, 1991]. A particularização consistirá, então, em etiquetar cada um dos intervalos que formam parte de uma restrição temporal entre duas variáveis, distinguindo-o dos outros intervalos que fazem parte da mesma restrição e do resto das restrições [Barber, 2000].

Assim as operações são consideradas como unidades de processamento elementares, onde cada uma delas será realizada por um recurso disponível initerruptamente.

Neste contexto, será de grande importância representar eventos, tais como início e fim da execução dos operadores, através de variáveis temporais (pontos temporais) t_{pi} e t_{pj}. Cada restrição, portanto, terá a forma $a \leq t_{pi} - t_{pj} \leq b$, onde a e b são constantes. Então, se t_{pi} e t_{pj} são início e o fim da execução do operador O_p, a duração da restrição pode ser representada como $d_j \leq t_{pj} - t_{pi} \leq d_i$.

Considerando o papel fundamental do tempo no problema, é necessário o estabelecimento de um marco temporal como referência em que serão identificados os trabalhos e operações, assim como as restrições entre eles. Será considerada um modelo de tempo ***T*** discreto e finito, tendo como primitiva temporal o ponto $t_i \in \boldsymbol{T}$. Sobre estes pontos de tempo serão estabelecidas as restrições métrico-disjuntivas [Dechter, 1991], cuja forma geral é:

$t_i \{[d_1\ D_1], [d_2\ D_2], \ldots, [d_n\ D_n]\}\ t_j$, com $d_k \leq D_k$, para todo $k \in [1,n]$

e cuja semântica é:

$(t_j - t_i \in [d_1\ D_1]) \vee (t_j - t_i \leq [d_2\ D_2]) \vee \ldots\ldots\ldots \vee (t_j - t_i \leq [d_n\ D_n])$

Definimos também dois pontos especiais $\boldsymbol{T_0}$ e $\boldsymbol{T_f}$ como o *tempo inicial* do escalonamento e o *tempo final* do escalonamento, respectivamente. Uma *operação* $o_{ij} \in \boldsymbol{O}$ está associada a um par de pontos $on(o_{ij})$ *e* $off(o_{ij})$ tais que $on(o_{ij}) \leq off(o_{ij})$ ou, na nossa notação, $on(o_{ij})\{[0,\infty]\}off(o_{ij})$. A duração de uma operação $dur(o_{ij})=d_{ij}$ pode ser expressada como $on(o_{ij})\ [d_{ij}\ d_{ij}]\ off(o_{ij})$.

Um *trabalho* $J_i \in \boldsymbol{J}$ é uma seqüência de operações $o_{i0}, o_{i1}, \ldots, o_{im}$. As restrições de precedência entre as operações terão portanto a seguinte forma $off(o_{ij})\{[0,\infty]\}on(o_{i(j+1)})$, para todo $j \in [1,m]$, que serão denotadas *precedências*(J_i).

É possível também representar o intervalo de tempo mínimo $[d_{min}\ D_{min}]$ *(ready time)* a partir do qual um trabalho j_i pode começar como $\boldsymbol{T_0}\{[d_{min}\ D_{min}]\}on(o_{i0})$, assim como o intervalo de tempo máximo $[d_{max}\ D_{max}]$ *até quando* um trabalho j_i pode terminar como $\boldsymbol{T_0}\{[d_{max}\ D_{max}]\}off(o_{im})$.

Normalmente cada operação deve portanto ser executada em uma máquina, de forma exclusiva, caracterizando as restrições disjuntivas: $on(o_{uj})\{[dur(o_{uj})\ \infty], [-\infty\ dur(o_{vj})]\}on(o_{vj})$. Deve ser observado, no entanto, que esta restrição insere uma disjunção, que consequentemente aumenta a complexidade do processo. A escolha destas disjunções será a base do método de busca proposto.

Assim, o padrão de fluxo do processo será determinado pelas restrições de precedência para cada um dos trabalhos, *precedências*(J_i), para $i \in [1,n]$. Por exemplo, se estivermos lidando com um escalonamento do tipo *flow-shop*, a ordenação dos elementos de cada trabalho será a mesma:

precedências(J_x)=*precedências*(J_y), *para todo* $x,y \in [1,m]$. Poderá ser diferente se estivermos lidando com um *job-shop*.

Para a representação de todos os elementos envolvidos no problema, utilizaremos uma *rede de restrições temporais* (RRT) [Dechter, 91], o que nos permitirá obter uma seqüência de ordenação de cada uma das operações sobre os recursos para realizar todos os trabalhos num mínimo de tempo necessário.

3.3 Representação de instâncias do problema de satisfação de restrições temporais

Seja ***o*** uma operação, ***on(o)*** o ponto temporal que representa o seu inicio ***e*** ***off(o)***, o ponto temporal que representa o seu final:

- $on(o) = t_i$, supondo que o inicia em t_i

- $off(o) = t_j$, supondo que *o termina em* t_j

Graficamente estes pontos temporais t_i *e* t_j podem ser representados como nós de uma rede de restrições ligados por uma aresta que representa o intervalo de tempo transcorrido entre o início e o fim da operação, que neste caso corresponde a duração *d* da operação.

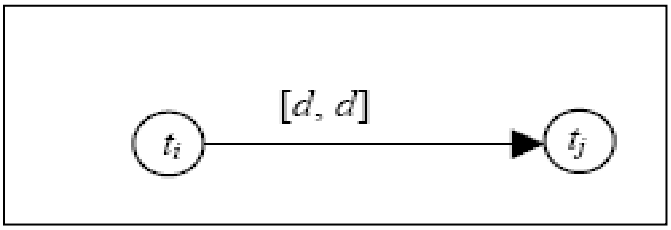

Figura 7 – Representação temporal

A rede temporal que será utilizada como marco de representação estará então formada por restrições **binárias** entre pontos de tempo. Em função da cardinalidade dos intervalos temporais das restrições, serão classificadas em dois grupos:

1 – *Restrições não disjuntivas* ou formadas por um único intervalo de tempo:

São todas as restrições representadas por um único intervalo temporal. Para especificar a duração $dur(o_i)$, de uma operação o_i, a restrição temporal seguinte será, $(on(o_i)\ \{[d\ d]\}\ off(o_i))$.

Já que a representação das restrições pode ser feita através de intervalos de tempo, podemos então representar as restrições que sejam fixas e as variáveis.

A relação de precedência entre as atividades o_i *e* o_j, será representada pela restrição temporal $(off(o_i)\ \{[0\ \infty]\}\ on(o_j))$. Esta restrição indica que do término da operação o_i, até o começo da operação seguinte o_j deve transcorrer um período de tempo ≥ 0.

Uma tarefa J_i *será* composta por um conjunto de m operações $\{o_{ij}\}$, i=1..n, j = 1..m, sendo n o número total de atividades. Estas operações estarão então ordenadas como restrições de precedência entre elas sejam estas de tempo ou recursos, assim uma dada tarefa J_i, formada por m operações ordenadas o_{i1}, o_{i2}, o_{im}, a primeira operação será chamada de ***first(J$_i$)*** e a ultima de ***last(J$_i$)*** então :

- $first(J_i) = o_{i1}$

- $last(J_i) = o_{im}$

Para especificar o **tempo de início** r, ou seja o tempo a partir do qual o trabalho J_i pode começar a utilizar os recursos, usaremos as restrições $(\boldsymbol{T0}\ \{[r\ \infty]\}\ on(first(J_i)))$, indicando que J_i pode começar a partir de r. Da mesma forma, supondo um **tempo limite** *dd,* ou tempo máximo permitido para o término J_i será especificado mediante a restrição $(\boldsymbol{T0}\ \{[0\ dd]\}\ off(last(J_i)))$, significando que a última operação de J_i pode terminar com atraso em *dd*. A figura a seguir representa esquematicamente esta situação, onde um trabalho J_i, formado por m operações com tempo de início r e um *deadline dd.*

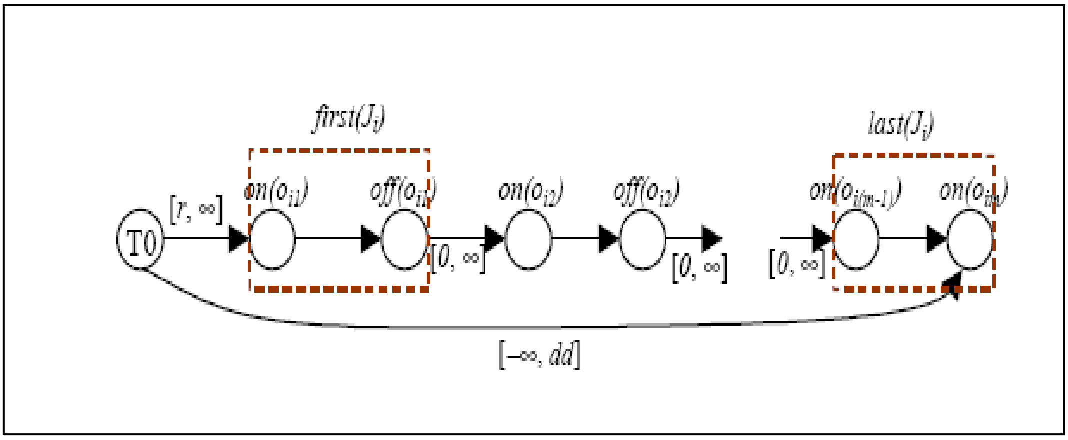

Figura 8 - Restrições não disjuntivas

Se a duração das atividades é fixa, então podemos simplificar a rede de restrições utilizando unicamente os pontos temporais que especificam o começo das tarefas, $on(o_{ij})$. Com este procedimento conseguimos reduzir o número de pontos temporais necessários. A figura anterior demonstra este fato.

2 – Restrições Disjuntivas

Estão relacionadas ao fato de que cada recurso só pode processar uma operação de cada vez, então os operadores o_i e o_j, que utilizariam o mesmo recurso devem ficar nesta ordem ou em ordem inversa. A restrição disjuntiva que representa esta situação $on(o_i)$ $\{[d_i\ ],[-\infty,-d_j]\}$ $on(o_j))$, sendo d_i e d_j as durações das operações o_i *e* o_j respectivamente. O intervalo $[d_i\ \infty]$ expressa a ordem de seqüenciamento, que significa que a operação o_i ocorre antes do que a operação o_j, ainda que $[-\infty,-dj]$ indica a ordem. A representação gráfica aparece na figura a seguir.

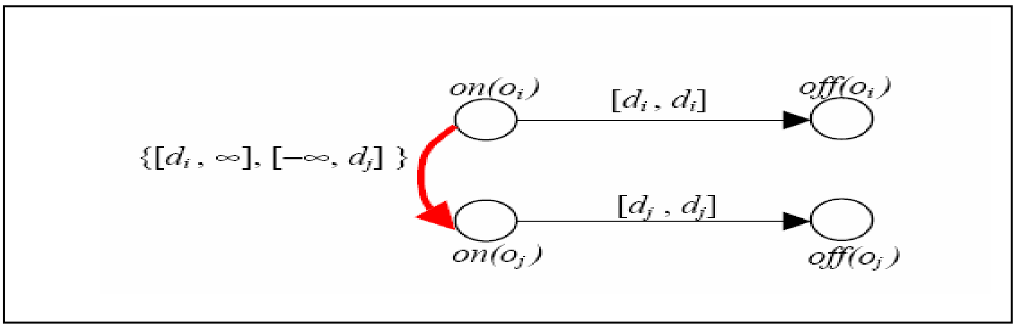

Figura 9 - Restrições Disjuntivas

Cada uma das restrições binárias da rede será denominada de restrição *básica*. Entre as restrições não disjuntivas que lidam com a especificação da duração de uma atividade, e também com as relações de precedência entre as atividades e o tempo de início e final para cada uma das tarefas. As restrições disjuntivas são utilizadas para expressar que um mesmo recurso não pode realizar simultaneamente duas operações distintas. Na especificação das restrições dos problemas de escalonamento como um conjunto de restrições temporais, serão consideradas dois tipos de especificações segundo o número de restrições:

- especificações básicas onde cada uma esta associada com uma única restrição temporal; e
- especificações de alto nível, são um conjunto de restrições temporais.