SIMPÓSIO DOUTORAL 2006

Relatório de Desenvolvimento/Resultados do Projecto de Doutoramento

A. IDENTIFICAÇÃO

Doutorando: Jorge Alexandre de Albuquerque Loureiro
jloureiro@di.estv.ipv.pt

Título da Tese: Reestruturação Dinâmica de Estruturas Multidimensionais de Dados em Tempo Útil

Orientador: Professor Doutor **ORLANDO BELO**, Universidade do Minho, Departamento de Informática
obelo@di.uminho.pt

Data Início: - Oficial 2 Abril 2003 - Efectiva 1 Janeiro 2004

Data Término: Prevista: 30 Setembro de 2006

B. Resumo

B.1 Área de Investigação e Desenvolvimento (I&D)

Informática – Optimização OLAP; Selecção de Estruturas Multidimensionais de Dados.

B.2 Resumo

A existência de estruturas multidimensionais de dados sob a forma de vistas materializadas ou subcubos é condição *sine qua non* de desempenho em sistemas de processamento analítico (OLAP). Estas estruturas não são mais do que o resultado da pré-agregação e materialização dos dados segundo as diversas dimensões/hierarquias. Em sistemas OLAP de dimensionalidade elevada, o número destas estruturas pode ser muito alto (milhares ou mesmo mais). Desta forma, importa limitar o seu número, dados os constrangimentos espaciais de armazenamento e, especialmente, os custos de manutenção destas estruturas, já que importa providenciar à sua actualização periódica, reflectindo o estado das relações base. Assim, se por um lado, a existência destas estruturas é benéfica para o tempo de resposta das interrogações analíticas e, tendencialmente mais benéfica, quão maior for o número de subcubos materializados, por outro lado, cada novo subcubo implica, em regra, um aumento de custos de manutenção (embora nem sempre, dada a sua conhecida natureza não-monotónica). Estamos, assim, em presença de duas grandezas de sentidos opostos, cuja minimização importa estudar e que abriu uma nova área de investigação, conhecida como problema de selecção de vistas ou cubos, que continua a merecer atenção redobrada por parte da comunicada científica. Trata-se de um problema de optimização de carácter combinatório (cada subcubo pode ou não ser materializado), reconhecidamente NP-hard, para o qual foram propostas inúmeras soluções, utilizando heurísticas várias (greedy, evolucionárias e pseudo-aleatórias), e baseadas em modelos de custos lineares e, mais recentemente, não lineares, quase totalmente endereçadas a estruturas multidimensionais centralizadas. Só, recentemente, o âmbito do problema foi alargado a uma abordagem distribuída, sob a forma de uma arquitectura OLAP multi-nó. Uma outra questão, que vem introduzir uma nova dimensão ao problema, consiste na periodicidade da recalibração das estruturas, já que, cada alteração do perfil das interrogações colocadas, deveria, idealmente, sugerir modificações nas selecções feitas. Em resumo, o problema da selecção de vistas ou subcubos pode ser classificado, na sua generalidade, segundo uma perspectiva tridimensional: 1) espacial, relativa à distribuição das estruturas, 2) lógica de selecção (heurísticas e outras lógicas de optimização); e 3) tempo

(periodicidade de recalibração). Outras dimensões possíveis poderiam ser 4) os constrangimentos aplicados ao processo, e 5) a natureza do modelo de custos subjacente ao cálculo de custos (de interrogação e manutenção).

Este trabalho de doutoramento está enquadrado na problemática descrita, também numa abrangência multidimensional: 1) propõe soluções para o problema da selecção de subcubos para OLAP centralizado e distribuído, 2) usa algoritmos evolucionários e de enxame de partículas discretos (além dos greedy) e 3) pode considerar-se maioritariamente de selecção estática, mas também alargar-se para uma perspectiva dinâmica (se considerarmos a recalibração fina das estruturas mais "leves"). Ainda quanto às referidas dimensões adicionais, nos algoritmos propostos, são aplicados constrangimentos espaciais e temporais de manutenção ao processo de selecção e é usado um modelo linear e também não linear de base à estimação dos custos de interrogação e manutenção. Estes custos são calculados usando algoritmos especificamente desenvolvidos para o efeito que, no caso de estarmos em presença de arquitecturas OLAP multi-nó, simulam o processamento paralelo das tarefas, utilizando assim o paralelismo inerente da arquitectura. Os algoritmos propostos podem, futuramente, ser integrados num middleware OLAP distribuído, implementado como um sistema multi-agente, que permitirá a automatização completa da utilização e manutenção das estruturas OLAP distribuídas, sob o controlo e fornecendo informação acerca do estado do sistema ao administrador do Data Warehouse (DW). Este sistema distribuído ficará responsável:

1. Por aceitar as interrogações colocadas pelos utilizadores OLAP, actuando no sentido de providenciar à sua resposta (seleccionado o par nó/subcubo capaz de melhor responder ao pedido);
2. Pelo processo de armazenamento da história das interrogações colocadas (possivelmente num DW de interrogações) e extracção de várias estatísticas;
3. Pela extracção de informação relacionada com a utilidade dos subcubos com uma possível abordagem especulativa, tentando adivinhar o seu uso futuro;
4. Pela geração da nova distribuição do M-OLAP cubo;
5. Pelo processo de manutenção das estruturas OLAP: geração dos subcubos ou deltas e a sua transmissão e integração nos nós previamente seleccionados, possivelmente, numa abordagem de processamento paralelo;
6. Pela difusão de informação acerca da distribuição global dos subcubos que permitam ao componente 1 decidir da melhor forma de responder a cada interrogação colocada.

B.3 Objectivos Estratégicos

Contribuir para a satisfação dos utilizadores das plataformas de processamento analítico, mormente os agentes de decisão, aumentando a sua produtividade e qualidade das suas

decisões, através da melhoria substancial do tempo de resposta às interrogações OLAP colocadas. Vai actuar-se a nível da optimização das estruturas multidimensionais e dados, com especial ênfase na sua vertente distribuída. Para este objectivo pretendem conceber-se, desenvolver e avaliar algoritmos de selecção e alocação de cubos para sistemas OLAP centralizados e distribuídos, utilizando modelos de custos lineares e não lineares capazes de suportar o cálculo de custos de interrogação e manutenção, usando algoritmos de cálculo de simulação de execução paralela (no caso de se tratar de uma arquitectura OLAP distribuída).

C. Contribuições

C.1 Principais contribuições técnico-científicas

1. Aplicação de algoritmo de enxame de partículas discreto ao problema de selecção de cubos para sistemas OLAP centralizados e sua avaliação comparativa face a um algoritmo genético standard.
2. Adaptação e concepção de modelos de custos linear e não linear para uma arquitectura OLAP multi-nó. Os modelos desenvolvidos permitiram derivar fórmulas de cálculo dos custos de interrogação e manutenção usando unidades temporais, considerando custos de busca/agregação de dados, custos de comunicação e de integração. Os modelos incluem parâmetros típicos reais definidores das características dos canais de comunicação e dos nós OLAP.
3. Desenvolvimento de algoritmos de estimação de custos de interrogação e manutenção, usando os modelos anteriores, que simulam a execução paralela das tarefas, aproveitando o paralelismo inerente da arquitectura M-OLAP.
4. Concepção e desenvolvimento de algoritmos de selecção (e alocação) de cubos para arquitecturas OLAP centralizada (e distribuída) utilizando algoritmos evolucionários, co-evolucionários e de enxame de partículas com o objectivo de minimizar os custos de interrogação (e manutenção), aplicando constrangimentos temporais de manutenção (e espaciais de materialização).

A descrição anterior surge numa perspectiva temporal de I&D: 1) iniciar com arquitectura OLAP centralizada e expandir a solução para uma abordagem OLAP distribuída; 2) conceber e desenvolver modelos e ferramentas e depois promover a sua aplicação. Em termos de relevância técnico-científica, especificaria a ordem (4./3./1./2).

C.2 Publicações

1. A Discrete Particle Swarm Algorithm for OLAP Data Cube Selection. To appear in: Proceedings of 8th International Conference on Enterprise Information Systems, Paphos, Cyprus, 23-27 May, 2006.

Abstract: Multidimensional analysis supported by Online Analytical Processing (OLAP) systems demands for many aggregation functions over enormous data volumes. In order to achieve query answering times compatible with the OLAP systems' users, and allowing all the business analytical views required, OLAP data is organized as a multidimensional model, known as data cube. The materialization of all the data cubes required for decision makers would allow fast and consistent answering times to OLAP queries. However, this also imply intolerable costs, concerning to storage space and time, even when a data warehouse had a medium size and dimensionality - this will be critical on refreshing operations. On the other hand, given a query profile, only a part of all subcubes are really interesting. Thus,

cube selection must be made aiming to minimize query (and maintenance) costs, keeping as a constraint the materializing space. That is a complex problem: its solution is NP-hard. Many algorithms and several heuristics, especially of greedy nature and evolutionary approaches, have been used to provide an approximate solution. To this problem, a new algorithm is proposed in this paper: particle swarm optimization (PSO). According to our experimental results, the solution achieved by the PSO algorithm showed a speed of execution, convergence capacity and consistence that allow electing it to use in data warehouse systems of medium dimensionalities.

2. Genetic and Swarm Algorithms for the Selection of OLAP Data Cubes. In Proceedings of the 4th WSEAS International Conference on Information Security, Communications and Computers (ISCOCO'05), Canary Islands, 16-18 Dec. 2005.

Abstract: The materialization of some aggregate views is a common technique for speeding up OLAP query processing. The huge amount of data usually stored in a data warehouse and the complexity of its schema implies that only a few of the total aggregated views may be materialized, specially due to the refresh update time strain required when the base relations are modified in result of changes in the operational world. The correct selection of the materialized views is a basic condition for performance and its study motivated the panoply of proposals to solve the so called cube view selection problem, which solution is NP-hard. Several heuristics were proposed as base to the design of the algorithm, being the most relevant the greedy and evolutionary ones. In this paper, we study the performance of two biological motivated algorithms to this problem: a genetic and a discrete particle swarm algorithm, based on a linear cost model, considering both query and maintenance costs and a space constraint. According to our experimental results, both algorithms showed a speed of execution, convergence capacity and consistence that allow electing them to use in data warehouse systems of medium and moderated dimensionalities.

3. Estimating Querying and Maintenance Costs for Restructuring Data Cubes. To appear in Proceedings of the IASTED International Conference on Databases and Applications (DBA'2006), Innsbruck, Austria, 14-16 February, 2006.

Abstract: In their daily routine, enterprise decision makers use to analyze huge amounts of information in order to sustain their decisions and, consequently, ensuring success of enterprises business activities. Through time, on-line analytical processing systems have contributed decisively to the decision making process improvement, not only by granting extremely flexible data manipulation mechanisms, but also allowing the materialization of the analysis indexes required. However, that analytical "power" uses to exhaust the computational resources, especially disk space and processing time, especially materializing specialized views. Besides, as time goes by, multidimensional databases become very large, being its management very difficult. Aiming to optimize maintenance and operationality of such databases, we design a system that is able to restructure them in useful time and reduce multidimensional query processing time, according to the exploitation trends of knowledge workers. In this paper, we present the system's structure, its correspondent cost model, query and maintenance algorithms, restructuring strategies, and, finally, its distribution through several processing OLAP nodes.

4. Life Inspired Algorithms for the Selection of OLAP Data Cubes. In WSEAS Transactions on Computers, Issue 1, Volume 5, January 2006, pp. 8-14.

Abstract: The use of materialized views is a common technique to speed up on-line analytical processing, as the on-demand scan and aggregation of detailed data would imply, mostly, large query response times, which would hurt hardly the

efficacy of decision making process. But the huge amount of data usually stored in a data warehouse and the complexity of its schema implies that only a few of the total aggregated views may be materialized, specially due to the refresh update time strain required when the base relations are modified in result of changes in the operational world. The correct selection of the materialized views is a basic condition for performance, but it is a recognized NP-hard problem and so, has triggered a huge variety of approximate solution' proposals. Several heuristics were proposed to the designing layer of the algorithms, being the most relevant the greedy and evolutionary ones. In this paper, we study the performance of two biological inspired algorithms applied to the cube selection problem: a genetic and a discrete particle swarm, based on a linear cost model, considering both query and maintenance costs and a space constraint. According to our experimental results, both algorithms showed a speed of execution, convergence capacity and consistence that allow electing them to use in data warehouse systems of medium and moderated dimensionalities, being the swarm solution the one with better overall performance.

5. A Non-Linear Cost Model for Multi-Node OLAP Cubes. Accepted to the CAiSE'06 Forum, 5-9 June, 2006, Luxembourg.

Abstract: User satisfaction depends heavily on the proper selection of cube views to materialize in OLAP systems. Frequently, any OLAP system administrator has to recalibrate periodically the OLAP structures, trying to follow the better he could the changes occurred on users' queries profiles. Any automatic system that will be able to propose the most efficient set of views to materialize in a particular moment would be a valuable help increasing the quality of service of the OLAP system. This paper introduces a distributed non-linear generalized cost model that addresses the estimation of query and maintenance cost of an extended distributed OLAP platform with n nodes in a centralized local facility or in a high geographically spread on. This architecture extends the centralized OLAP structures to real distributed cube views structures using a number of OLAP server nodes interconnected by a heterogeneous communication network.

D. Enquadramento

D.1 Enquadramento Científico

Conforme descrito em B. 2, o problema da selecção de vistas ou subcubos pode ser classificado, na sua generalidade, segundo uma perspectiva tridimensional, aqui repetida:

1) espacial, relativa à distribuição das estruturas,
2) lógica de selecção (heurísticas e outras lógicas de optimização);
3) tempo (periodicidade de recalibração).

Ainda referindo B.2, outras dimensões caracterizadores das soluções do problema de selecção de vistas ou subcubos podem ser:

4) os constrangimentos aplicados ao processo;
5) a natureza do modelo de custos subjacente ao cálculo de custos (de interrogação e manutenção).

Enquadrando este trabalho de doutoramento na caracterização atrás especificada, teremos, correspondentemente, a cada dimensão, que:

1) propõe soluções para o problema da selecção (e alocação) de cubos em sistemas OLAP centralizados e distribuídos (quando aplicado a arquitecturas M-OLAP);
2) usa algoritmos evolucionários e de enxame de partículas discretos (além dos greedy) para a solução do problema da selecção e alocação de cubos;
3) pode considerar-se que os algoritmos propostos são maioritariamente de aplicabilidade à selecção estática das estruturas, mas que podem também alargar-se a uma perspectiva dinâmica (se considerarmos a recalibração fina das estruturas mais "leves").

Ainda quanto às referidas dimensões adicionais:

4) nos algoritmos propostos, são aplicados constrangimentos espaciais e temporais de manutenção ao processo de selecção;
5) sendo usado um modelo linear e também não linear de base à estimação dos custos de interrogação e manutenção; estes custos são calculados usando algoritmos especificamente desenvolvidos para o efeito que, no caso de estarmos em presença de arquitecturas OLAP multi-nó, simulam o processamento paralelo das tarefas, utilizando assim o paralelismo inerente da arquitectura.

D.2 Motivação

A motivação para este trabalho de doutoramento surge na sequência dos trabalhos conducentes à escrita da dissertação de mestrado e interesse despertado pela problemática da selecção das vistas e soluções propostas. Também o estudo de algoritmos de data mining, com espacial destaque nas heurísticas construtivas greedy e algoritmos genéticos, abriu a

porta à sua aplicação ao problema já conhecido. O aprofundar do estudo de outros algoritmos inspirados na vida e sua aplicação em outros problemas de optimização, conduziu à sua aplicação ao problema da selecção das vistas, no início, a uma arquitectura OLAP centralizada, rapidamente estendida a uma perspectiva distribuída.

D.3 Objectivos Detalhados

1. Conceber e desenvolver algoritmos de estimação de custos de interrogação e manutenção baseados em modelos de custo lineares;
2. Aplicar algoritmos de optimização discreta por enxame de partículas na selecção de cubos numa arquitectura OLAP centralizada utilizando como funções de optimização os algoritmos desenvolvidos no ponto anterior; efectuar a avaliação comparativa do algoritmo desenvolvido face a um algoritmo genético standard;
3. Adaptar e estender modelos de custos lineares e não-lineares a uma arquitectura OLAP distribuída (M-OLAP – MultiNode OLAP), de forma a permitir conceber e desenvolver algoritmos capazes de estimar os custos de interrogação e manutenção, utilizando simulação de execução paralela de tarefas, aproveitando assim o paralelismo inerente da arquitectura;
4. Estender a aplicação de algoritmos genéricos de optimização, especialmente no domínio dos algoritmos evolucionários e de enxame ao problema da selecção e alocação de cubos numa arquitectura M-OLAP, utilizando como funções de adaptação os algoritmos de cálculo de custos referidos no objectivo anterior;
5. Efectuar testes experimentais simulados dos algoritmos desenvolvidos, avaliando o seu desempenho e escalaridade e o impacto do valor especificado para alguns parâmetros de funcionamento e também uma avaliação comparativa de algumas variantes desses algoritmos;
6. Efectuar uma avaliação comparativa dos algoritmos propostos face a algumas soluções existentes.

D.4 Trabalhos Alternativos

O problema clássico da selecção de cubos foi abordado pela primeira vez em [Harinarayan et al. 1996], tendo sido, desde então, alvo de grande esforço de investigação. Embora seja patente uma enorme diversidade de propostas, para descrever os trabalhos alternativos, vou, mais uma vez, socorrer-me da classificação tridimensional descrita em B.2 e que dá origem à esquematização do trabalho relacionado mostrada na Fig. 1, que refere igualmente algumas das propostas referenciais para cada classe.

Olhando a figura, percebe-se que, quanto à primeira dimensão, há um sobrepovoamento de soluções endereçadas a estruturas centralizadas. Quanto à segunda dimensão, há a

considerar uma maioria de propostas no domínio dos algoritmos greedy. A estas há a juntar soluções baseadas em heurísticas específicas, em algoritmos genéticos e selecção aleatória com pesquisa iterativa, por têmpera simulada e uma combinação de ambas. A optimização por enxame de partículas, surge já no âmbito deste trabalho de doutoramento. Quanto à terceira dimensão, teremos as soluções estáticas (onde o intervalo de recalibração é alargado e assim $t >> 0$), dinâmicas (ajuste fino, $t \rightarrow 0$) ou pró-activas (a preparação das estruturas, especulando as necessidades futuras, implica um $t < 0$). Importa tecer algumas considerações, ainda que breves, quanto a cada uma destas categorias.

Consideremos, em primeiro lugar a distribuição das estruturas. A maioria das soluções é focada na perspectiva centralizada, já que é a mais utilizada e também, até há muito pouco tempo, a única ser implementada. Só muito recentemente, a solução distribuída veio a terreiro [Bauer & Lehner 2003], nomeadamente em organizações que operam à escala global. Trata-se de transpor para o domínio do DW as vantagens já efectivadas com a distribuição das bases de dados nos sistemas transaccionais. Trata-se de efectuar uma fusão das soluções endereçadas ao DW centralizado com as soluções genéricas de distribuição de bases de dados. Às soluções tradicionais são agora acometidos problemas como optimização da distribuição da alocação dos dados e distribuição de processamento, além de terem de ser levados em conta os problemas da gestão de comunicações e controlo de congestão. Acrescente-se ainda a própria distribuição espacial das interrogações e temos os componentes fundamentais a ser levados em conta para o problema de selecção e alocação de cubos distribuídos, que pode ser definido como a selecção dos subcubos apropriados em cada um dos nós OLAP de uma arquitectura M-OLAP, considerando uma distribuição espacial de interrogações, e constrangimentos como capacidade de armazenamento e processamento de cada nó, características dos links de comunicação, de forma a minimizar os custos totais, um somatório dos custos relativos à resposta das interrogações colocadas e da manutenção das próprias estruturas multidimensionais, já que estas, periodicamente, necessitam de ser refrescadas.

Já quanto à segunda dimensão, temporalidade de recalibração, as denominadas propostas estáticas, baseadas em heurísticas greedy várias [Harinarayan et al. 1996], [Shukla et al. 1998],[Gupta & Mumick 1999],[Liang et al. 2001], ou em algoritmos genéticos [Horng et al. 1999],[Zhang et al. 2001],[Lin & Kuo 2004], actuam normalmente a nível de estruturas de grandes dimensões. Não permitem uma reconfiguração a intervalos curtos, já que de custos muito elevados, daí a sua denominação “estática”, pois que relativamente invariável no tempo. Já as soluções que podem ser englobadas no que e denominou como abordagem dinâmica [Scheuermann et al. 1996], [Kotidis & Roussopoulos 1999], actuam, em regra a nível de caches, que são normalmente de pequena dimensão e, além disso, não implicam, em regra, custos adicionais para a sua materialização, já que armazenam as respostas a interrogações colocadas. As últimas, pró-activas, tal como o seu nome indica, introduzem a perspectiva especulativa, procurando saber antecipadamente as necessidades e, dessa forma, prepará-las em tempo útil. Caches com prefetching ou restruturação (com recalculo dinâmico dos subcubos

futuramente necessários) são formas possíveis de soluções desta classe de propostas [Belo 2000],[Sapia 2000]. Também a inclusão de avaliação de utilidade futura de subcubos ou fragmentos na política de admissão e substituição de cache, constitui uma proposta englobada nesta categoria de soluções [Park et al. 2003].

Já no que respeita à terceira dimensão, a primeira família, que poderemos denominar de "heurísticas greedy" é aquela que um maior número de propostas de solução mereceu. Logo em [Harinarayan et al. 1996], foi iniciada, sob a forma de um algoritmo GSC (Greedy under Space Constraint), que introduz a heurística base de toda a família: o conceito de benefício e iniciar com um conjunto de subcubos vazio, a que vai adicionando aquele que, em cada fase, se revelar como o mais benéfico em termos da diminuição do custo de resposta ao conjunto de interrogações colocadas. Muitas outras propostas poderiam ser referidas, que foram acrescentando heurísticas várias à heurística base descrita, alargando a aplicabilidade e introduzindo novos constrangimentos (como é o caso do tempo de manutenção [Gupta & Mumick 1999]). Na actualidade, há nesta família, quatro propostas dominantes: os algoritmos greedy de árvore invertida e A* [Gupta & Mumick 1999] e os algoritmos greedy de duas fases e greedy integrado [Liang et al. 2001]. Uma análise comparativa e descrição pormenorizada pode ser encontrada em [Yu et al. 2004].

A segunda família, traz para o domínio da selecção das vistas ou subcubos os algoritmos genéticos [Holland 1992]. Na sua essência, um algoritmo genético procura capitalizar directamente a forma como se processa a evolução, sendo conhecido que esta é um método robusto e com sucesso para adaptação em sistemas biológicos. Podem pesquisar espaços de hipóteses, contendo partes inter-actuantes complexas, onde o impacto de cada parte nas hipóteses genéricas de elegância possam ser difíceis de modelar. Começam com uma população aleatória (grupos de cromossomas); em cada geração são seleccionados pais e gera-se uma descendência, utilizando operações análogas aos processos biológicos, tipicamente, o cruzamento e mutação. Adoptando o princípio da sobrevivência do mais apto, todos os cromossomas são avaliados utilizando uma função de adaptação para determinar a qualidade de cada solução, que é então utilizada para decidir que indivíduos (e respectivos cromossomas) vão ser eliminados ou propagados. Quão mais elevada for a adaptação, maior a probabilidade do cromossoma respectivo ser utilizado na geração da nova população. O processo é repetido sucessivamente até que um determinado critério de finalização seja satisfeito. O cromossoma com o maior valor de adaptação na última população dá a solução. Uma descrição geral dum algoritmo genético simples pode ser consultada em [Goldberg 1989]. Ao fim de um número de gerações reduzido, conseguem chegar a soluções, senão melhores do que os algoritmos greedy, pelo menos algo comparáveis [Horng et al. 1999], [Zhang et al. 2001], [Lin & Kuo 2004].

Já as propostas no domínio da pesquisa aleatória, procuram dar resposta ao tempo de execução muito longo dos algoritmos greedy, inaplicáveis em DW de elevada dimensionalidade. Em [Kalnis et al. 2002], é proposta uma heurística sob a forma de uma pool

(com um tamanho correspondente ao constrangimento espacial) à qual são adicionadas vistas (previamente ordenadas) até que o constrangimento temporal seja violado. A selecção e eliminação das vistas pode ser efectuada através de três tipos de pesquisa: iterative improvement (II), simulated annealing (SA) e two-phase optimization (2PO) uma combinação das duas anteriores, dando origem a três algoritmos. Estes são aplicados ao problema da selecção de vistas sob constrangimento espacial num DW centralizado, mostrando o último conseguir soluções de qualidade comparável ao algoritmo GSC em tempos três ordens de grandeza inferiores (para um DW com 15 dimensões).

Finalmente, há a referir a optimização por enxame de partículas, já que proposta no âmbito deste trabalho de doutoramento. O conceito do enxame de partículas teve origem como uma simulação de um sistema social simplificado. A sua aplicação em optimização é proposta em [Kennedy & Eberhart 1995] e [Eberhart & Kennedy 1995]. O algoritmo de optimização por enxame de partículas (PSO) simula a capacidade de processar conhecimento das sociedades animais mais evoluídas. Há duas versões principais do algoritmo PSO: a versão inicial, que usa valores reais, podendo denominar-se contínua e a versão binária ou discreta, proposta em [Kennedy & Eberhart 1997].

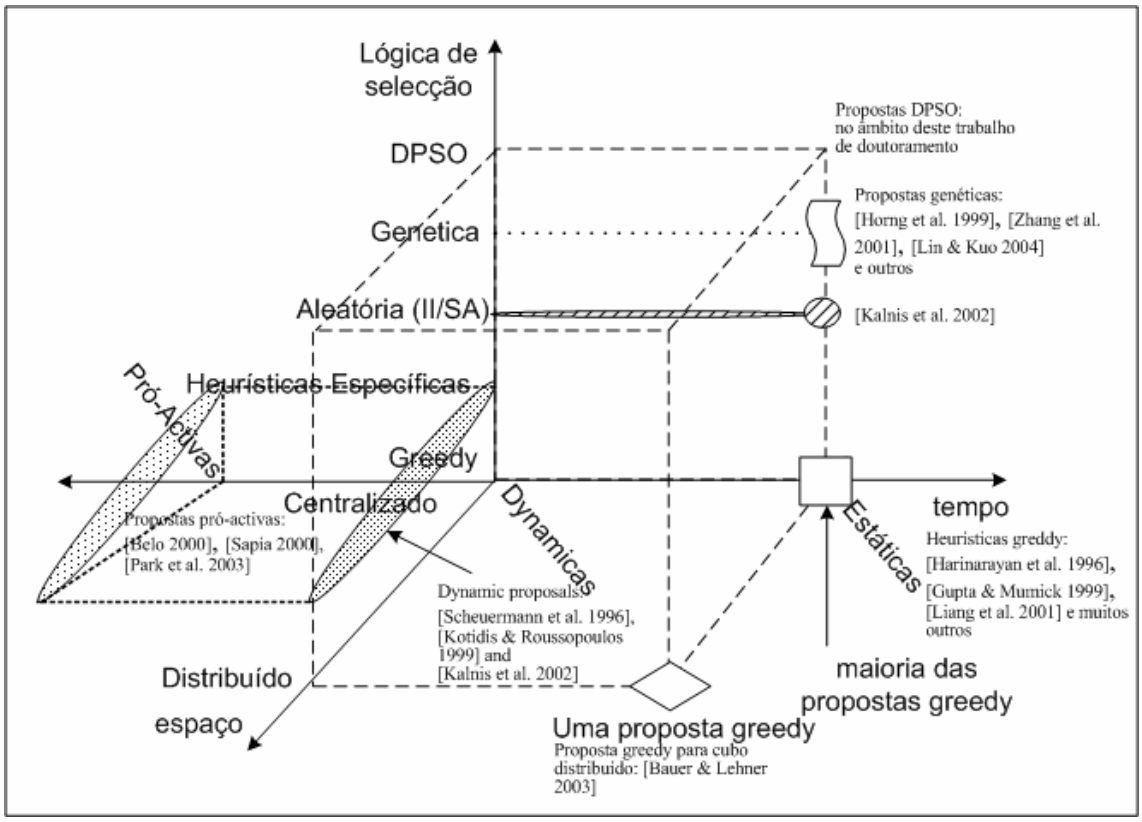

Fig. 1. Caracterização tridimensional das soluções propostas para o problema de selecção de cubos.

D.5 Bibliografia Principal

[Bauer & Lehner 2003] Bauer, A. and Lehner, W. "*On Solving the View Selection Problem in Distributed Data Warehouse Architectures*". Proceedings of the 15th International Conference on Scientific and Statistical Database Management (SSDBM'03), IEEE, 2003, 43-51.

[Belo 2000] Belo, O. "*Putting Intelligent Personal Assistants Working on Dynamic Hypercube Views Updating*". Proceedings of 2nd International Symposium on Robotics and Automation (ISRA'2000), Monterrey, México, November, 2000.

[Eberhart & Kennedy 1995] Eberhart, R.C., and Kennedy, J. "*A new Optimizer Using Particle Swarm Theory*". In Proc. Os the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, IEEE Service Center, Piscataway, NJ, pp. 39-43.

[Goldberg 1989] Goldberg, D.E. "*Genetic Algorithms in Search, Optimization, and Machine Learning*". Addison-Wesley, Reading, MA, 1989.

[Gupta & Mumick 1999] Gupta, H. and Mumick, I.S. "*Selection of Views to Materialize under a Maintenance-Time Constraint*". In: Proc. of the International Conference on Database Theory, 1999.

[Harinarayan et al. 1996] Harinarayan, V., Rajaraman, A., and Ullman, J. "*Implementing Data Cubes Efficiently*", Proc. of ACM SIGMOD, Montreal, Canada, June 1996, 205-216.

[Holland 1992] Holland, J.H. "*Adaptation in Natural and Artificial Systems*". MIT Press, Cambridge, MA, (2nd edition), 1992.

[Horng et al. 1999] Horng, J.T., Chang, Y.J., Liu, B.J., and Kao, C.Y. "*Materialized View Selection Using Genetic Algorithms in a Data Warehouse*". In Proceedings of World Congress on Evolutionary Computation, Washington D.C., July, 1999.

[Kalnis et al. 2002] Kalnis, P., Mamoulis, N., and D. Papadias. "*View Selection Using Randomized Search*". In: Data Knowledge Engineering, vol. 42, number 1, pp. 89-111, 2002.

[Kennedy & Eberhart 1995] Kennedy, J., and Eberhart, R.C. "*Particle Swarm Optimization*". In Proc. Of IEEE Intl. Conference on Neural Networks (Perth, Australia), IEEE Service Center, Piscataway, NJ, IV:1942-1948.

[Kennedy & Eberhart 1997] Kennedy, J., and Eberhart, R.C. "*A Discrete Binary Version of the Particle Swarm Optimization Algorithm.*". In Proc. of the 1997 Conference on Systems, Man and Cybernetics (SMC'97), pp. 4104-41098, 1997.

[Kotidis & Roussopoulos 1999] Kotidis, Y., and Roussopoulos, N. Dynamat: A Dynamic View Management System for Data Warehouses. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (Philadelphia, Pennsylvania, June 1999), pp.371-382.

[Liang et al. 2001] Liang, W., Wang, H., and Orlowska, M.E. "*Materialized View Selection Under the Maintenance Cost Constraint*". In: Data and Knowledge Engineering, 37(2), pp. 203-216, 2001.

[Lin & Kuo 2004] Lin, W.-Y., and Kuo, I-C. "*A Genetic Selection Algorithm for OLAP Data Cubes*". In Knowledge and Information Systems, Volume 6, Number 1, pp. 83-102, Springer-Verlag London Ltd., 2004.

[Park et al. 2003] Park, C.S., Kim, M.H., and Lee, Y.J. "*Usability-based Caching of Query Results in OLAP Systems*". In Journal of Systems and Software, Vol. 68, Issue 2, November 2003, 103-119.

[Sapia 2000] Sapia, C. "*PROMISE – Modeling and Predicting User Query Behavior in Online Analytical Processing Environments*". Proceedings of the 2nd International Conference on Data Warehousing and Knowledge Discovery (DAWAK'00), London, UK, Springer LNCS, September, 2000.

[Scheuermann et al. 1996] Scheuermann, P., Shim, J., and Vingralek, R. "*WATCHMAN: A Data Warehouse Intelligent Cache Manager*". Proc. of the 22th International Conference on Very Large Data Bases VLDB'96, Sept. 3-6, Bombay, 1996, 51-62.

[Shukla et al. 1998] Shukla, A., Deshpande, P.M., and Naughton, J.F. "*Materialized View Selection for Multidimensional Datasets*". In: Proc. of VLDB, 1998.

[Yu et al. 2004] Yu, J.X., Choi, C-H, Gou, G., and Lu, H. "*Selecting Views with maintenance Cost Constraints: Issues, Heuristics and Performance*". In Journal of Research and Practice in Information Technology, Viol. 36, No. 2, May 2004.

[Zhang et al. 2001] Zhang, C., Yao, X., and Yang, J. "*An Evolutionary Approach to Materialized Views Selection in a Data Warehouse Environment*". In: IEEE Trans. on Systems, Man and Cybernetics, Part C, Vol. 31, N.º 3, Sept. 2001.

E. Desenvolvimento

E.1 Macro-Planeamento das Actividades

A calendarização e plano de trabalhos inicialmente previsto (mostrado na Tabela 1) sofreu alterações, pelo facto de a dispensa formal de serviço docente só ter sido concedida em Janeiro de 2004 e, ainda, devido à impossibilidade de interrupção abrupta da prestação desse serviço, por motivos pedagógicos. Adicionalmente, a vertente de execução prática de um sistema completo e funcional foi preterida em lugar de uma mais profunda investigação na vertente de concepção e desenvolvimento de algoritmos de selecção e também na extensão do âmbito espacial do sistema de processamento analítico.

Tabela 1. Actividades efectivas do presente trabalho de doutoramento.

Fase	Duração	Actividades
F1	8 meses	• Estudo aprofundado das áreas de data warehousing, processamento analítico de dados, computação baseada em agentes e mecanismos de aprendizagem.
F2	6 meses	• Identificação de um caso de aplicação real para o desenvolvimento de um sistema de processamento analítico. • Definição e povoamento das estruturas multidimensionais de dados necessárias. • Análise, planeamento e implementação de um assistente de monitorização de processos de interacção em sistemas multidimensionais de dados para o sistema implementado. • Documentação do sistema desenvolvido.
F3	6 meses	• Definição e implementação de um assistente de reestruturação de estruturas multidimensionais de dados, com as características apresentadas na secção anterior. • Estudo e implementação de mecanismos de descoberta de conhecimento em hipercubos e sua integração no assistente desenvolvido. • Documentação do sistema desenvolvido.
F4	9 meses	• Definição e implementação de uma comunidade especial de "wrappers" com a capacidade de estabelecer novos planos de povoamento de hipercubos. • Estudo de mecanismos para a comunicação e coordenação de acções inter-wrappers. • Projecto e implementação das infra-estruturas de suporte do ambiente de integração de todas as entidades - assistentes e wrappers - do sistema. • Documentação do sistema desenvolvido..
F5	7 meses	• Escrita da tese de doutoramento.

Com base no exposto, só (F1) foi realizada de acordo com o planeado, tendo F3 sido executada a um nível bastante mais profundo, ainda que num sentido algo diverso do previsto inicialmente. Das restantes fases, F2 foi parcialmente executada, F5 foi repetida para cada modelo e algoritmo implementado e F6 vai ser iniciada brevemente. Assim também a duração prevista do trabalho de doutoramento foi estendida de 36 para 42 meses.

Mostra-se na Tabela 2 o plano e calendarização efectiva das actividades que é esquematizada temporalmente na Fig. 2..

Tabela 2. Actividades efectivas do presente trabalho de doutoramento.

Fase	Duração	Actividades
F1	10 meses	• Estudo aprofundado das áreas de data warehousing, processamento analítico de dados, computação baseada em agentes e mecanismos de aprendizagem. • Produção de relatórios técnicos correspondentes a todas as áreas estudadas (6).
F2	4 meses	• Focalização na problemática da selecção de estruturas multidimensionais e definição dos objectivos estratégicos. • Concepção e implementação de modelo de custos linear e algoritmos de cálculo de custos para OLAP centralizado. • Concepção e implementação de um algoritmo greedy e genético para a selecção de cubos em OLAP centralizado. • Construção de uma bancada de trabalho para possibilitar a posterior investigação no domínio da selecção de estruturas multidimensionais. • Documentação do sistema desenvolvido.
F3	3 meses	• Estudo de outros métodos de optimização, especialmente os inspirados na vida. • Investigação acerca da sua aplicabilidade ao problema da selecção de cubos. • Concepção, desenvolvimento e teste do algoritmo proposto: DPSO. • Documentação do sistema desenvolvido, com produção de um relatório técnico.
F4	6 meses	• Alargamento do âmbito espacial do sistema de processamento analítico: M-OLAP. • Concepção e desenvolvimento do respectivo modelo de custos e algoritmos de estimação de custos. • Sua utilização em algoritmos evolucionários de selecção e alocação de cubos em arquitecturas OLAP distribuídas. • Documentação do sistema desenvolvido com produção de relatórios técnicos.
F5	5 meses	• Produção e submissão de artigos relativos aos trabalhos já realizados.
F6	3 meses	• Refinamento do modelo de custos (inclusão de não-linearidades e suporte a manutenção integral e incremental). • Correspondente refinamento dos algoritmos de estimação de custos. • Produção e submissão de artigo científico.
F7	4 meses	• Adaptação dos algoritmos desenvolvidos ao novo modelo e algoritmos de cálculo de custos implementados. • Teste experimental simulado comparativo. • Produção e submissão de artigo científico com descrição do sistema e resultados obtidos.
F8	7 meses	• Escrita da tese de doutoramento.

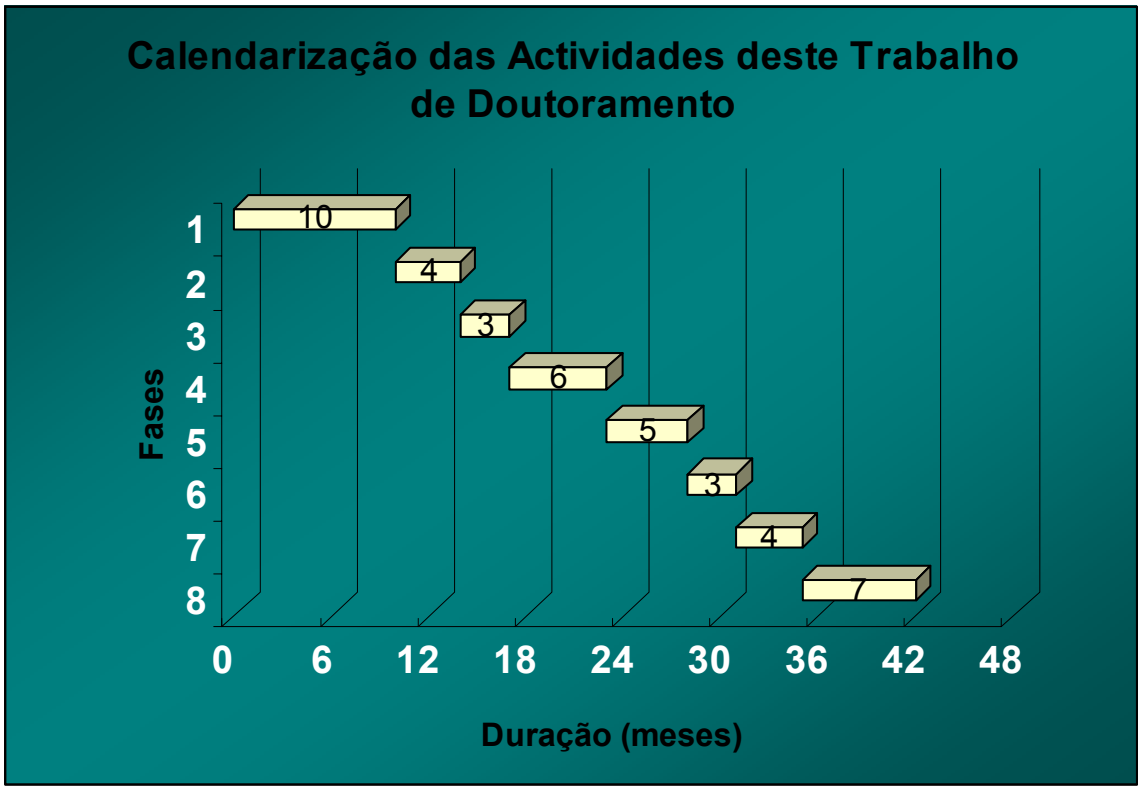

Fig. 2. Calendarização temporal das actividades realizadas e a realizar no presente trabalho de doutoramento.

E.2 Recursos Necessários

Tratando-se de um doutorando externo ao DI, este ponto não é aplicável.

E.3 Recursos Disponibilizados

Tratando-se de um doutorando externo ao DI, este ponto não é aplicável.

F. Avaliação

F.1 Análise Comparativa

Pelo exposto em B.2, D.1 e D.4 se percebe facilmente que os trabalhos do presente doutoramento se entrosam nas demais propostas da área. No entanto, enumeram-se aqui as principais contribuições, mostrando, sempre que haja propostas anteriores que possam servir de base para uma avaliação comparativa, o respectivo resultado:

- Proposta de uma nova lógica de selecção ao problema da selecção de cubos (optimização por enxame de partículas, versão discreta), que analisada comparativamente face a propostas greedy e genéticas existentes, mostrou ter um desempenho superior.
- Alargamento do modelo de custos linear para arquitecturas OLAP distribuídas (M-OLAP) com inclusão de grandezas reais (modelo estendido linear) e uso da métrica tempo para contabilização dos custos.
- Concepção de desenvolvimento de algoritmos de estimação de custos (de interrogação e manutenção) que incluem simulação de execução de tarefas em paralelo (para arquitecturas M-OLAP).
- Utilização de algoritmos genéticos normais e na sua variante co-evolucionária na solução do mesmo problema, utilizando o modelo anterior; a análise comparativa com algoritmo greedy (já existente, embora estendido agora para inclusão de custos de manutenção), revelou que a proposta com algoritmos genéticos e especialmente a sua variante co-evolucionária apresentaram desempenho claramente superiores. Estes algoritmos usaram para estimação de custos os algoritmos concebidos e desenvolvidos no âmbito do ponto anterior.
- Alargamento do modelo de custos para arquitectura M-OLAP para incluir não linearidades e suporte de o cálculo simultâneo de manutenção integral e incremental.
- Concepção e proposta de novos algoritmos de cálculo de custos, com simulação de execução paralela tipo *pipelining*, baseados no modelo proposto no ponto anterior.
- Uso da optimização por enxame de partículas, versão discreta na selecção e alocação de subcubos para selecção e alocação de subcubos numa arquitectura OLAP, ainda não realizado, mas que uma versão baseada num modelo de custos linear anterior, mostrou ter um bom desempenho.

F.2 Auto-avaliação do Trabalho Realizado

Os testes conduzidos até ao momento apontam no sentido da efectividade dos algoritmos implementados. O desenvolvimento da totalidade do sistema inicialmente suposto e descrito no final de B.2, foi relegada para um trabalho posterior, tendo sido dada prioridade à investigação de base.

F.3 Auto-avaliação da Documentação Produzida

A documentação produzida não se cinge apenas a artigos submetidos ou aceites em conferências, mas há toda uma produção de relatórios técnicos (só na fase 1 deste trabalho de doutoramento formam escritos 6) que importa referir. Considero que a documentação produzida abarca a totalidade das fases do trabalho, sendo um espelho fiel das actividades que foram sendo empreendidas.

O esforço de publicação teve início prático em Setembro de 2005, com uma elevada taxa de aceitação e está ainda a decorrer. Estão em fase de finalização dois novos artigos a submeter à DEXA'2006 e DaWaK'2006, respectivamente "Comparative Study of Maintenance Cost Computing Algorithms for Multi-Node Multidimensional Data Bases" e "Selecting and Allocating Distributed OLAP Cubes: An Evolutionary Approach". Versões alargadas da descrição de alguns dos trabalhos irão ser propostas para publicação em revistas da especialidade, a saber: "DATA & KNOWLEDGE ENGINEERING", "IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans", "ACM Transactions on Database Systems (TODS)". Planeia-se ainda submissão de artigos a talvez duas outras conferências: "JISBD'2006" e "SMC 2006".

Em resumo, considero que a produção de documentos decorreu a um ritmo elevado, tendo a sua qualidade melhorado ao longo do tempo (especialmente a escrita em inglês). A produção em curso pode vir a ser prejudicada pela exiguidade do tempo ainda disponível até ao *terminus* inadiável do presente trabalho e a necessária paralelização com a escrita da tese.

G. Dificuldades

G.1 Dificuldades Técnico-Científicas

As principais dificuldades sentidas prenderam-se, numa fase inicial, com o estudo aprofundado de algumas áreas apenas antes vislumbradas. Posteriormente, na elaboração do modelo de custos distribuído, um suporte científico na área de redes, especialmente no que concerne à definição de custos de comunicação em bases de dados distribuídas teria sido uma enorme ajuda e, porventura, uma muito maior validade científica dos modelos desenvolvidos.

G.2 Outras Dificuldades

A implementação real de um sistema M-OLAP, permitiria não só, refinar os modelos de custos propostos e a sua validação como possibilitaria também a validação experimental dos algoritmos de estimação de custos desenvolvidos. Uma proposta neste sentido foi ainda intentada (sob a forma de um projecto a lançar aos alunos de licenciatura), mas foi, infelizmente, inviabilizado, pois que não aceite. Futuramente, este trabalho poderá constituir um projecto de investigação a adicionar à implementação do middleware, já referido.