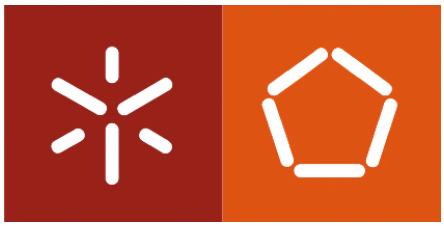


Universidade do Minho
Escola de Engenharia
Departamento de Informática

SIMPÓSIO DOUTORAL 2006

Relatório do Doutorando

Paulo Jorge Machado Oliveira

A. IDENTIFICAÇÃO

A.1 Doutorando: Paulo Jorge Machado Oliveira
pjo@isep.ipp.pt

A.2 Título da Tese: Detecção e Correcção de Problemas de Qualidade nos Dados

A.3 Orientadores: Prof. Doutor Pedro Rangel Henriques
Universidade do Minho
Departamento de Informática
prh@di.uminho.pt

Prof. Doutora Maria de Fátima Rodrigues
Instituto Superior de Engenharia – Instituto Politécnico do Porto
Departamento de Engenharia Informática
fr@dei.isep.ipp.pt

A.4 Data Início: 07/2003 (oficial)
02/2004 (efectiva)

A.5 Data Término: 07/2007 (prevista)
02/2008 (efectiva)

B. RESUMO

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

Qualidade dos dados/limpeza dos dados (dados segundo o modelo relacional).

B.2 Resumo

As organizações públicas e privadas começam, finalmente, a perceber o valor dos dados que têm à sua disposição, e a considerá-los como um bem importante no aumento da produtividade, eficiência e competitividade. Como consequência, a exploração de enormes volumes de dados assume um papel cada vez mais importante na sociedade actual. No entanto, constata-se que a generalidade das fontes apresenta Problemas de Qualidade dos Dados (PQD). Na literatura, estes problemas também são designados de erros, anomalias, ou mesmo "sujidade". Entre outras possibilidades, estes problemas correspondem a valores de atributos em falta, valores de atributos errados, ou representações diferentes dos mesmos dados. Os problemas de qualidade nos dados criam problemas à sua efectiva utilização, influenciando negativamente a validade dos resultados e conclusões obtidas.

Assim sendo, antecedendo a aplicação de qualquer ferramenta orientada à análise, os dados devem ser "limpos" com o intuito de detectar e corrigir quaisquer problemas de qualidade que possam existir. Neste contexto, a qualidade dos dados tem sido objecto de um interesse crescente ao longo dos últimos anos.

As actuais soluções que visam a melhoria da qualidade dos dados evidenciam uma cobertura deficiente, em amplitude e qualidade de manipulação, quando comparadas com o vasto leque de problemas susceptíveis de ocorrerem. Assim, o principal objectivo deste trabalho consiste no desenvolvimento de uma arquitectura para a melhoria da qualidade dos dados que suporte de uma forma global e integrada todos esses problemas. Pretende-se que a automatização seja um princípio orientador neste trabalho. Procura-se eliminar ou, pelo menos, reduzir ao máximo a intervenção humana. Na procura da automatização, métodos "inteligentes" são incluídos na arquitectura. A solução preconizada para atingir estes vários objectivos baseia-se na definição de uma linguagem formal para descrever cada PQD e respectiva manipulação, a nível de detecção e correcção.

B.3 Objectivos Estratégicos

- Identificar de forma sistemática e rigorosa todos os problemas de qualidade que podem ser encontrados nos dados ao nível das instâncias (i.e., valores dos dados).
- Propor uma arquitectura para a melhoria da qualidade dos dados que:
    - Cubra os problemas identificados.
    - Suporte não só uma mas ambas as vertentes envolvidas no processo, isto é, detecção e correcção dos problemas existentes.
    - Incorpore mecanismos de aprendizagem que permitam reduzir a necessidade de intervenção por parte do utilizador na definição das operações de detecção e correcção a efectuar.
- Materializar a arquitectura proposta num sistema computacional.

C. CONTRIBUIÇÕES

C.1 Principais Contribuições Técnico-científicas

As principais contribuições técnico-científicas são o resultado natural dos objectivos estratégicos traçados para este trabalho de doutoramento, ou seja:

- Definição de uma taxionomia dos problemas de qualidade que ocorrem ao nível da instância.

- Formalização do significado de cada problema de qualidade dos dados.
- Especificação de uma arquitectura que:
    - Cobre um maior leque de problemas do que as soluções actualmente existentes.
    - Integra em simultâneo operações de detecção e correcção dos PQD num único ambiente.
    - Incorpora mecanismos "inteligentes" que reutilizam conhecimento sobre as operações de detecção e correcção previamente realizadas, diminuindo a necessidade de intervenção por parte do utilizador.
- Desenvolvimento de um sistema computacional que visa a melhoria da qualidade dos dados baseado na arquitectura especificada.

C.2 Publicações

Publicação 1: Paulo Oliveira, Fátima Rodrigues, e Pedro Henriques – "A Formal Definition of Data Quality Problems". In *Proceedings of the 10th International Conference on Information Quality*, MIT, Boston, EUA, Novembro de 2005. p. 13-26.

Abstract. The exploration of data to extract information or knowledge to support decision making is a critical success factor for an organization in today's society. However, several problems can affect data quality. These problems have a negative effect in the results extracted from data, affecting their usefulness and correctness. In this context, it is quite important to know and understand the data problems. This paper presents a taxonomy of data quality problems, organizing them by granularity levels of occurrence. A formal definition is presented for each problem included. The taxonomy provides rigorous definitions, which are information-richer than the textual definitions used in previous works. These definitions are useful to the development of a data quality tool that automatically detects the identified problems.

Publicação 2: Paulo Oliveira, Fátima Rodrigues, Pedro Henriques, e Helena Galhardas – "A Taxonomy of Data Quality Problems". In *Proceedings of the 2nd International Workshop on Data and Information Quality* (em conjunto com a conferência CAiSE'05), Porto, Portugal, Junho de 2005. p. 219-233.

Abstract. In today's society the exploration of one or more databases to extract information or knowledge to support management is a critical success factor for an organization. However, it is well known that several problems can affect data quality.

These problems have a negative effect in the results extracted from data, influencing their correction and validity. In this context, it is quite important to understand theoretically and in practice these data problems. This paper presents a taxonomy of data quality problems, derived from real-world databases. The taxonomy organizes the problems at different levels of complexity. Methods to detect data quality problems represented as binary trees are also proposed for each complexity level. The paper also compares this taxonomy with others already proposed in the literature.

Publicação 3: Paulo Oliveira, Fátima Rodrigues, e Pedro Henriques – "A Framework for Detection and Correction of Data Quality Problems". In *Proceedings of Data Gadgets 2005 International Workshop - Bringing Up Emerging Solutions for Data Warehousing Systems*, Granada, Espanha, Setembro de 2005. p. 59-74.

Abstract. The exploration of data to extract information or knowledge to support management is a critical success factor for an organization in today's society. However, several problems can affect data quality. These problems have a negative effect in the results extracted from data, affecting their usefulness and correctness. In this context, it is quite important to detect and correct these data problems. This paper presents a modular framework, tightly integrating the detection and correction of data quality problems. The purpose and operation mode of each module is explained in detail. The framework has extensibility capabilities which allow domain-specific algorithms to be plugged-in. The framework also has an intelligent behavior, trying to use in the current case the available knowledge about the previous ones.

Publicação 4: Paulo Oliveira, Fátima Rodrigues, e Pedro Henriques – "Data Profiling versus Data Quality Problems". In *Proceedings of the 2nd International Workshop on Data and Knowledge Quality* (em conjunto com a conferência EGC'06), Lille, França, Janeiro de 2006. p. 9-15.

Abstract. Data suffers from several quality problems. Before starting a data-driven project, data must be assessed to check whether the required quality is assured. Data profiling is a new emerging technology whose intention is to detect the data quality problems. In this paper, we present an overview about data profiling and assess the coverage that is given to data problems. The assessment is based on our proposed taxonomy of data quality problems.

Publicação 5: Paulo Oliveira, Fátima Rodrigues, e Pedro Henriques – "Limpeza de Dados: Uma Visão Geral". In *Proceedings of Data Gadgets 2004 International Workshop - Bringing Up Emerging Solutions for Data Warehousing Systems*, Málaga, Espanha, Novembro de 2004. p. 39-51.

Resumo. No contexto da actual necessidade de explorar bases de dados, para delas extrair informação/conhecimento para apoio à gestão, é fundamental a correcção/validade dos dados para a qualidade dos resultados extraídos. Sendo certo que são várias as soluções parciais para a resolução dos problemas nos dados, tornou-se necessário fazer uma sistematização de todos os erros que podem ocorrer no sentido de identificar aqueles ainda não resolvidos e preconizar uma abordagem global. Este artigo descreve precisamente o referido estudo, e as ilações que se extraem quando se comparam os erros com as abordagens de limpeza de dados actualmente existentes, perspectivando-se a concepção de uma nova aproximação global à limpeza de dados como trabalho futuro, em consequência das conclusões obtidas.

D. ENQUADRAMENTO

D.1 Enquadramento Científico

Nos pontos seguintes, contextualiza-se a importância dos objectivos estratégicos deste trabalho de doutoramento para a área da qualidade dos dados/limpeza dos dados.

- Identificar de forma sistemática e rigorosa todos os problemas de qualidade que podem ser encontrados nos dados ao nível das instâncias – A criação de uma taxionomia de PQD é importante por dois motivos: (1) permite avaliar qual a cobertura dada por uma determinada solução informática à detecção e correcção dos PQD; (2) auxilia a orientar os esforços de I&D para os problemas que não têm suporte, a nível de detecção ou correcção.
- Propor uma arquitectura para a melhoria da qualidade dos dados que:
    - Cubra os problemas identificados – Do estudo realizado foi possível constatar que há um conjunto de problemas da taxionomia que actualmente não é suportado (a nível de detecção e/ou correcção) por qualquer solução existente (por exemplo: detecção da utilização de unidades de medida diferentes). Isto significa que nenhuma solução cobre o leque de problemas incluídos na taxionomia desenvolvida. Na realidade, o conjunto de

problemas suportado encontra-se fragmentado pelas múltiplas soluções existentes.

- Integre em simultâneo operações de detecção e correcção dos PQD num único ambiente – Ainda que intimamente relacionadas, estas operações têm sido efectuadas de forma não integrada por diferentes ferramentas. A nível comercial, a detecção é efectuada por ferramentas de *Data Profiling*, enquanto que a correcção é efectuada por ferramentas de limpeza de dados e de Extracção, Transformação e Carregamento (ETC). A realização destas operações de uma forma integrada torna a manipulação dos problemas mais eficiente.
- Incorpore mecanismos de aprendizagem que permitam reduzir a necessidade de intervenção por parte do utilizador na definição das operações de detecção e correcção a efectuar – As soluções actualmente existentes implicam que o utilizador especifique as operações de detecção e correcção a efectuar. Este é um processo manual, complexo e fastidioso para o utilizador. Neste trabalho pretende-se reduzir essa intervenção, propondo ao utilizador automaticamente operações de detecção e correcção, com base nas operações anteriormente efectuadas em casos similares.

- Materializar a arquitectura proposta num sistema computacional – Desta forma, a arquitectura proposta poderá ser validada e demonstrado o seu interesse prático.

D.2 Motivação

Aumentar a cobertura, em amplitude e qualidade de manipulação, dada aos PQD que ocorrem ao nível da instância, tendo como referencial as soluções (comerciais e protótipos de investigação) actualmente existentes.

D.3 Objectivos Detalhados

Os principais objectivos específicos deste trabalho de doutoramento são:

- Identificar e classificar os PQD que podem ser encontrados ao nível da instância, organizando-os numa taxionomia.
- Formalizar o significado de cada problema de qualidade dos dados. O conjunto de definições matemáticas serve para exprimir de forma rigorosa cada problema. Estas definições são usadas para a detecção dos PQD.

- Explorar as lacunas e oportunidades de investigação encontradas resultantes do cruzamento entre os problemas incluídos na taxionomia e as soluções actualmente existentes para esses problemas.
- Especificar uma arquitectura que:
    - Suporte de uma forma global e integrada a detecção e a correcção dos PQD incluídos na taxionomia
    - Use uma linguagem declarativa para especificação das operações de detecção e correcção a efectuar.
    - Tenha um comportamento "inteligente", reconhecendo casos anteriores similares ao actual, propondo automaticamente ao utilizador as operações de detecção e correcção a efectuar.
- Desenvolver um sistema informático que materialize a arquitectura proposta para o problema em questão, tendo especial atenção à interface gráfica de diálogo com o utilizador.
- Validar a arquitectura proposta e o sistema informático subjacente, utilizando para o efeito dados reais em que se pretenda aumentar o seu nível de qualidade.

D.4 Trabalhos Alternativos

Nos parágrafos seguintes são apresentados trabalhos de I&D que, de uma forma directa ou indirecta, visam a melhoria da qualidade dos dados. Estes trabalhos seguem abordagens distintas daquela que é defendida neste trabalho de doutoramento.

Na base do sistema *AJAX* [1, 2] encontra-se uma arquitectura flexível e extensível que procura separar o nível lógico (especificação das operações de limpeza de dados a realizar) do físico (aspectos relacionados com a implementação) de um processo de limpeza de dados. O seu principal objectivo consiste em transformar dados de uma ou várias fontes num determinado esquema alvo, manipulando uma série de problemas típicos de qualidade dos dados (por exemplo: eliminação de duplicados) durante o processo. A lógica de um processo de limpeza de dados é modelada como um grafo dirigido de transformações sobre fluxos de dados. O processo de limpeza de dados recebe um conjunto de fluxos de dados, possivelmente errados e/ou inconsistentes e origina um conjunto de fluxos de dados formatados, correctos e consistentes. Uma linguagem declarativa e extensível, baseada em declarações SQL devidamente enriquecidas com um conjunto de primitivas de transformação, possibilita a especificação das transformações de dados (programas de

limpeza de dados) de uma forma compacta e de manutenção simplificada. A semântica de cada uma destas transformações envolve a geração de excepções sobre situações anormais que possam ocorrer (erros ou inconsistências). Estas excepções são o alicerce de um ambiente interactivo com o utilizador. O último aspecto relevante consiste na existência de um mecanismo genealógico dos dados, o que possibilita a obtenção de explicações. Para cada transformação de dados, o utilizador pode obter informação sobre quais os tuplos que estiveram na base da geração de um determinado tuplo.

O *ARKTOS* [3] é uma ferramenta que possibilita a modelação e execução de cenários de ETC para a criação de armazéns de dados, com base num conjunto de primitivas que permitem a realização de tarefas usuais. Entre estas, além de operações de transformação, encontram-se também operações de limpeza de dados, consideradas como uma parte integrante do processo de ETC. Este processo consiste numa sequência de passos que extraem dados relevantes das fontes, efectuam a sua transformação para o formato pretendido, procedem à sua limpeza e, por último, executam o seu carregamento para o armazém de dados. As operações durante o processo de ETC são denominadas de actividades. Uma actividade constitui uma unidade atómica de trabalho, apresentando-se como um passo na sequência de operações do processo. Uma vez que a finalidade de uma actividade é efectuar processamento sob um fluxo de dados, cada uma destas encontra-se conectada a tabelas de entrada e de saída, de uma ou mais bases de dados. A cada actividade encontra-se também associada um tipo de erro particular e uma política. A lógica subjacente a uma actividade é descrita declarativamente através de uma instrução SQL. No entanto, não é obrigatório que a sua execução seja efectuada como tal.

O *IntelliClean* [4, 5] assenta numa arquitectura genérica baseada em conhecimento para a limpeza inteligente dos dados, com especial ênfase na eliminação de duplicados. A arquitectura pode ser aplicada sobre qualquer base de dados, permitindo a implementação de qualquer estratégia de limpeza de dados actualmente existente. Estas estratégias, traduzindo conhecimento sobre o domínio, são representadas sob a forma de regras, sendo a sua aplicação efectuada através de um motor de inferência de um sistema pericial. A arquitectura especifica três fases distintas para o processo de limpeza de dados: pré-processamento, processamento e verificação e validação humana. Na fase de pré-processamento, os tuplos são analisados e todas as anomalias sintácticas susceptíveis de serem detectadas nesta fase são corrigidas. Entre estas encontram-se verificações ao tipo de dados, uniformização de formatos e adopção de uma representação consistente para as

abreviaturas. A fase de processamento envolve a avaliação das regras de limpeza sobre os tuplos pré-processados, que alimentam o mecanismo de inferência do sistema pericial (os tuplos são os factos). As regras especificam a realização de determinadas acções mediante a ocorrência de determinadas situações nos tuplos, podendo conter predicados complexos e referências a funções externas tanto no antecedente como no seu consequente. Na fase de verificação e validação humana, como o nome o indica, é necessária a intervenção humana para analisar o ficheiro de registo. Este ficheiro permite verificar a consistência e precisão das acções efectuadas e, eventualmente, efectuar a sua correcção.

O *Potter's Wheel* [6] é um protótipo de investigação baseado numa arquitectura interactiva simples mas poderosa para a transformação e limpeza de dados. Permite aos utilizadores gradualmente procederem à composição e análise do efeito das transformações num interface gráfico e intuitivo do tipo folha de cálculo. O sistema permite a especificação gráfica de variadas transformações de dados. A especificação do processo de limpeza de dados é efectuada interactivamente sob a forma de um conjunto de transformações simples aplicados sobre uma amostra de dados. O efeito de uma transformação pode ser observado de imediato, pelo utilizador, nos tuplos visíveis no ecrã. Não é necessário aguardar pela transformação de todo o conjunto de dados para se analisar as suas consequências. Em simultâneo, algoritmos de exploração de dados e algoritmos específicos do domínio pesquisam incrementalmente, em segundo plano, a existência de problemas na última versão transformada dos dados, assinalando-os à medida que estes são encontrados.

O *FraQL* [7, 8] define uma arquitectura para as tarefas envolvidas na preparação de dados (integração, transformação, limpeza e redução de dados), tendo por base uma linguagem declarativa que permite o acesso e manipulação de tuplos armazenados em múltiplas fontes. Tendo por base um modelo de dados objecto-relacional, a linguagem é uma extensão ao SQL, com características que permitem cobrir as necessidades particulares inerentes à preparação de dados. A implementação das extensões como primitivas de base de dados permite tirar partido das potencialidades intrínsecas dos sistemas de gestão de base de dados, produzindo um efeito sinergético para ambas. A principal vantagem da utilização de uma linguagem deste tipo, que combina mecanismos de preparação dos dados e potencialidades poderosas de interrogação a várias fontes heterogéneas, consiste numa integração virtual, na qual é possível executar operações de transformação e limpeza, sem afectar o conjunto de dados original. Desta forma, é possível ensaiar e avaliar diferentes estratégias de integração e limpeza, sem ser necessário proceder ao

carregamento e à materialização explícita dos dados, o que resulta num esforço computacional reduzido. Uma atenção especial é concedida à limpeza de dados, materializada na possibilidade de realização de qualquer uma das seguintes operações: detecção e reconciliação de duplicados; preenchimento de valores em falta; manipulação de ruído nos dados; e detecção e remoção de desvios.

D.5 Bibliografia Principal

[1] Galhardas, H.; Florescu, D.; Shasha, D. e Simon E. – "AJAX: An Extensible Data Cleaning Tool". In *Proceedings of the ACM SIGMOD on Management of Data*. Dallas, EUA. 2000.

[2] Galhardas, H.; Florescu, D.; Shasha, D.; Simon E. e Saita, C.A. – "Declarative Data Cleaning: Language, Model and Algorithms". In *Proceeding of the 27th Very Large Databases Conference*. Roma. Itália. 2001.

[3] Vassiliadis, P.; Vagena, Z.; Skiadopoulos, S.; Karayannidis, N. e Sellis, T. – "ARKTOS: Towards the Modeling, Design, Control and Execution of ETL Processes". *Information Systems*, 26: 537-561. 2001

[4] Lee, M. L.; Ling, T. W. e Low, W. L. – "IntelliClean: A Knowledge-Based Intelligent Data Cleaner". In *Proceedings of the ACM SIGKDD*. Boston, EUA. 2000.

[5] Low, W. L.; Lee, M. L. e Ling, T. W. – "A Knowledge-Based Approach for Duplicate Elimination in Data Cleaning". *Information Systems*, 26: 585-606. 2001.

[6] Raman, V. e Hellerstein, J. M. – "Potter's Wheel: An Interactive Framework for Data Transformation and Cleaning". In *Proceeding of the 27th Very Large Databases Conference*. Roma. Itália. 2001.

[7] Sattler, K. U.; Conrad, S. e Saake, G. – "Adding Conflict Resolution Features to a Query Language for Database Federations". In *Proceedings of the 3rd International Workshop on Engineering Federated Information Systems*. Dublin, Irlanda. 2000.

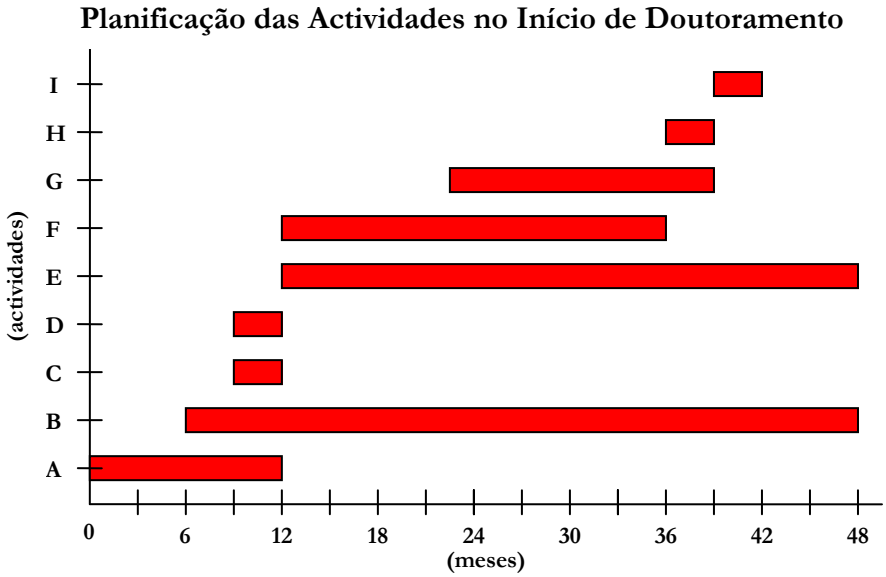
[8] Sattler, K. U. e Schallehn E. – "A Data Preparation Framework Based on a Multidatabase Language". *International Database Engineering Applications Symposium*. Grenoble, France. 2001.

[9] Rahm, E. e Do, H. H. – "Data Cleaning: Problems and Current Approaches". *IEEE Bulletin of the Technical Committee on Data Engineering*, 24(4). 2000.

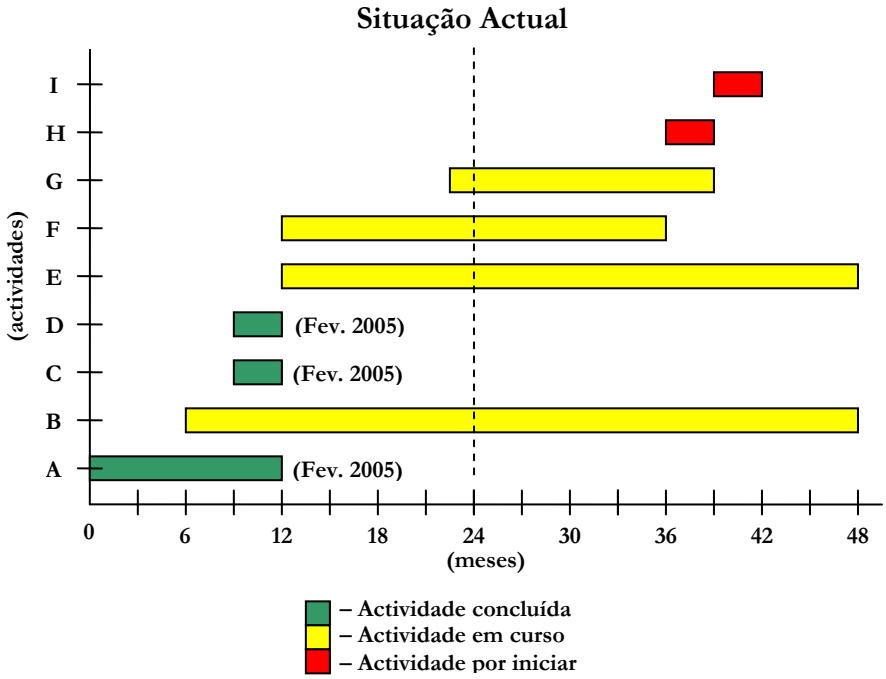
[10] Olson, J. – "Data Profiling: The Accuracy Dimension". Morgan Kaufmann Publishers, 2003.

E. DESENVOLVIMENTO

E.1 Macro-planeamento das Actividades



A – Estudo do estado da arte
B – Escrita da tese de doutoramento
C – Identificacão da motivacão + definicão de um enquadramento
D – Definicão dos objectivos estratégicos
E – Escrita e submissão de artigos em revistas e conferências
F – Desenvolvimento de actividades de investigacão original
G – Desenvolvimento de software para experimentacão
H – Realizacão de experiências utilizando o software desenvolvido
I – Análise dos resultados



E.2 Recursos Necessários

- 7744 horas de trabalho individual
- Hardware: PC com as seguintes características mínimas: Processador 2.5 GHz, 512 MB de memória; 60 GB de disco; Drive de CD/DVD
- Software de carácter genérico: Windows; Microsoft Office; Java; SQL Server; MySQL; Oracle; Acrobat Reader;
- *Demos* de software especifico (comercial e protótipos de investigação) da área da qualidade dos dados (por exemplo: ferramentas de limpeza de dados; ferramentas de *data profiling*) para experimentação
- Acesso à Internet
- Acesso a recursos bibliográficos on-line que possibilitem o acesso ao texto integral das publicações (por exemplo: ACM, IEEE, Elsevier)
- Alguns livros da área da qualidade dos dados

E.3 Recursos Disponibilizados

Todos os recursos necessários à realização das actividades já concluídas foram disponibilizados. Igualmente, prevê-se que todos os recursos necessários, à conclusão das actividades em curso e à execução das actividades ainda por iniciar, estarão também disponíveis.

F. AVALIAÇÃO

F.1 Análise Comparativa

A solução proposta neste trabalho distingue-se dos trabalhos alternativos apresentados, fundamentalmente pelo seguinte:

- A nível de detecção, os trabalhos alternativos cobrem um conjunto reduzido de PQD (por exemplo: detecção de duplicados). Isto significa que na generalidade dos problemas estes trabalhos não integram simultaneamente detecção e correcção, mas apenas correcção.
- Nenhum dos trabalhos alternativos cobre na íntegra o universo de problemas incluídos na taxionomia produzida. Na realidade, o conjunto de problemas coberto encontra-se fragmentado pelos diversos trabalhos alternativos.

- Os trabalhos alternativos não incluem capacidades de aprendizagem, dependendo sempre do utilizador para a especificação das operações de detecção e correcção a realizar.
- Essencialmente, a filosofia dos trabalhos alternativos é de se apresentarem como ferramentas de ETC, realizando operações de limpeza para melhorar a qualidade dos dados durante o processo. A filosofia deste trabalho é a de proceder à detecção e correcção dos problemas na própria fonte dos dados, sem realizar esse processo.

F.2 Auto-avaliação do Trabalho Realizado

Todo o trabalho realizado até ao momento visa a persecução dos objectivos delineados para este trabalho de doutoramento. Nos parágrafos seguintes encontra-se a respectiva argumentação:

- A revisão bibliográfica efectuada serviu para criar contexto e fundamentar a escolha e a direcção de investigação.
- A elaboração de uma taxionomia sistematiza os PQD que é possível encontrar ao nível da instância. O tratamento destes, a nível de detecção e correcção, é o cerne deste trabalho, daí a importância da sua identificação.
- A formalização de cada PQD incluído na taxionomia fornece definições rigorosas, que serão usadas na detecção e correcção dos problemas.
- A arquitectura proposta para detecção e correcção dos PQD, ainda que possa ser objecto de alguns refinamentos, define já com bastante exactidão a solução preconizada para o problema sobre o qual se debruça este trabalho de doutoramento.
- A definição de uma linguagem declarativa para especificação das operações de detecção e correcção a realizar visa ser o interface de alto nível entre o utilizador e o sistema computacional.
- O desenvolvimento do protótipo, entretanto já iniciado, visa materializar a arquitectura defendida numa ferramenta informática.

F.3 Auto-avaliação da Documentação Produzida

O trabalho de doutoramento realizado até ao momento originou cinco publicações. Todas estas publicações foram efectuadas em eventos internacionais, tendo sido aceites após um processo de revisão com *referees*. As publicações foram efectuadas em eventos específicos da área, tendo-se evitado aqueles que são de carácter generalista. Isto atesta a receptividade

e avaliação positiva que tem sido concedida ao trabalho em curso, entre a comunidade científica da área. Pela sua reconhecida relevância, destaca-se a publicação efectuada no MIT (publicação 1 na secção C.2) naquela que, provavelmente, é a conferência de maior importância nesta área da qualidade dos dados/informação.

As publicações (ver secção C.2) cobrem integralmente o trabalho efectuado até ao momento. As publicações 4 e 5 surgem como um resultado natural da revisão bibliográfica efectuada. A identificação e classificação de todos os problemas de qualidade susceptíveis de serem encontrados nos dados resultaram na produção de uma taxionomia, o que originou a publicação 2. Cada PQD foi depois devidamente formalizado para lhe dar um significado rigoroso. Este trabalho resultou na publicação 1. Por último, foi proposta uma arquitectura de um sistema computacional que visa suportar, tão automaticamente quanto o possível, a detecção e correcção de todos estes PQD. Este trabalho originou a publicação 3. Esta arquitectura está em fase de refinamento/detalhe, tendo este trabalho já resultado na produção de um novo artigo, já submetido a uma conferência internacional. Neste momento, decorre o processo de revisão aguardando-se o resultado da submissão.

G. DIFICULDADES

G.1 Dificuldades Técnico-Científicas

O acesso às *demos* das ferramentas comerciais existentes nesta área da qualidade dos dados com o objectivo de experimentação efectiva das suas potencialidades foi praticamente impossível. Na generalidade dos casos, as empresas não disponibilizam *demos* das suas ferramentas. Apenas em dois casos tal foi possível. Assim, a análise das potencialidades ficou restrita à informação técnica que foi possível recolher sobre as ferramentas nos respectivos *sites*. Apesar de ter faltado o carácter experimental, tal não teve consequências de maior no planeamento e no andamento dos trabalhos de doutoramento em curso. Os resultados que se pretendiam obter acabaram por se alcançados, isto é, as potencialidades e lacunas destas ferramentas foram identificadas.

G.2 Outras Dificuldades

Não foram sentidas dificuldades não técnicas ou científicas na realização dos trabalhos efectuados até ao momento.