

Este relatório é apenas para dar a conhecer o nosso trabalho até agora e para onde a nossa pesquisa e realização do trabalho se encaminha. Iremos fazer uma breve descrição do nosso modelo de dados, e dare-mos a conhecer o algoritmo escolhido para a implementação, e o seu funcionamento de uma forma sucinta. Expressamos desde já que se algum ponto não tiver sido focado com muita clareza, para entrarem em contacto conosco para qualquer informação.

O nosso modelo de dados é constituído por 4 tabelas, sendo uma delas (VendasArtigos) apenas uma tabela de resolução de uma relação de N:M. Como um artigo pode estar associado a várias vendas e como uma venda pode ter mais do que um artigo, a relação entre a tabela Artigos e Vendas, seria uma relação N:M, e daí tivemos que proceder à criação de uma tabela extra.

Tal como o Alexandre nos disse, pusemos 13 registos de Clientes, 25 registos de Artigos, na tabela Vendas fizemos a cada cliente ter apenas uma venda, mas se for preciso mais acrescentaremos. A tabela VendasArtigos apresenta 63 registos, e optamos por distribuir os artigos pelos clientes, isto é, os artigos comprados pelos clientes variam dos 3 aos 8.

Fizemos vários *print screen* do modelo de dados e dos dados, e exportamos os dados para um ficheiro *txt* de modo a que seja possível observarem o trabalho até aqui feito. Envia-mos então uma imagem da base de dados geral que será a imagem "Modelo de dados", e envia-mos também uma imagem e um ficheiro *txt* sobre os dados das 4 tabelas existentes, em que são facilmente identificadas pelo nome dado aos ficheiros e imagens (nome de respectiva tabela).

O algoritmo escolhido por nós foi o *Mondrian*, que foi apresentado em 2006, por *Kristen le Fevre*, *Davit deWitt* e *Raghu Ramakrishnam*. Manda-mos também, junto deste relatório o *paper* de apresentação do *Mondrian*.

A partir dessa escolha, a nossa pesquisa têm-se focado exclusivamente no algoritmo escolhido. Através dessa pesquisa encontramos, além de diversos *papers*, a tese de doutoramento da *Kristen le Fevre* que tem uma secção dedicada inteiramente ao *Mondrian* que tem sido exaustivamente lida por nós.

Pelo que lemos, tiramos a conclusão que o algoritmo funciona através da recursividade sobre uma função chamada *anonymize*. Para isso teremos que escolher quais os nossos atributos *quasi-identifier* (no nosso caso consideramos como sendo a idade, o sexo e a localidade), e a partir daí, temos que escolher o melhor atributo candidato e parti-lo em 2 grupos de atributos de valores semelhantes e de pelo menos tamanho k e assim sucessivamente, até não ser possível partir em mais grupos de atributos. Isto é, para um *2-anonymize*, só se poderá partir em grupos em que contenham mais de 2 atributos. Por exemplo, a figura abaixo mostra isso perfeitamente, na figura do lado esquerdo não se pode partir pois tem $k-1$ atributos, mas se se puser mais um elemento, já se poderá partir em grupos.

	a1	a2	a3
b1		k-1	
b2	k-1	m	k-1
b3		k-1	

	a1	a2	a3
b1		k-1	
b2	k-1	m	k-1
b3		k-1	1

Figure 1: *k-anonimize*

A nossa dúvida, é descobrir qual a melhor implementação para descobrir o atributo que se deverá partir, e é nesta etapa de pesquisa que nos encontramos agora, enquanto vamos continuando a implementação das outras funções, como a *anonymize*.

Outra das nossas dúvidas que esquecemo-nos de mencionar hoje ao professor Manuel Barbosa, foi que ainda não temos muito ideia como podemos conciliar o modelo de dados com a nossa implementação.

E por fim, gostaríamos de saber se no próximo relatório, que será apresentado na próxima semana, é necessário já ter a implementação do algoritmo completa e se é necessário a execução de testes.