

Base de Dados de Cliente: Modelo RF-Similis

Luís Cavique

ESCS, Instituto Politécnico de Lisboa, Portugal

lcavique@escs.ipl.pt

Resumo

Nos últimos anos tem-se assistido à fusão das Tecnologias de Informação e Comunicação com o Marketing Orientado para o Cliente, em aplicações que vão desde o simples "direct-mail", ao CRM, passando pelas Base de Dados de Clientes (*Database Marketing*). Nesta comunicação apresenta-se um novo algoritmo de segmentação de clientes, denominado RF, inspirado no algoritmo RFM, mas com a vantagem que permite definir diferentes estratégias de comunicação. O modelo RF-Similis resulta, da associação ao algoritmo Similis, que é utilizado na determinação de cabazes de compras.

Palavras chave: base dados clientes, "database marketing", RFM, "cross-selling", cabaz de compras.

1 Introdução

Os Sistemas de Informação SI, tradicionalmente, incluem a dicotomia Dados *versus* Processos – nesta abordagem iremos incluir o elemento Comunicação.

Até cerca de 1990 podemos classificar os SI orientados para os dados, para os processos e para a comunicação. No que diz respeito a SI orientados para os dados incluiremos os Sistemas Gestores de Base de Dados, para dados formatados e os sistemas de "Office Automation" para dados não formatados. Os sistemas orientados para os modelos podemos referir os "Decision Support System", os "Executive Information Systems" e os "Expert Systems".

A partir de 1990 assiste-se a uma conjunto de fusões de SI e sistemas de comunicação, a par da convergência da tecnologia digital (na electrónica, telecomunicações e computadores), o que veio tornar o vocabulário dos Sistemas de Informação mais rico [Amaral e Santos 1997]. Um desses novos sistemas é o CRM (*Customer Relationship Management*), que se tem revelado uma estratégia de gestão e extrema importância, patente em várias conferências, artigos e exposições [Reis 2000] [Brown 2000]. Contudo o número de implementações em Portugal é diminuto, pelo que julgamos com o interesse recuar um passo no tempo e analisar os sistemas que lhe deram origem – as Base de Dados de Clientes (*Database Marketing*).

A secção 2 abordamos a estrutura essencial de uma base de dados de clientes, também conhecida por "database marketing". As secções 3 e 4 são dedicadas aos algoritmos: na secção 3 é apresentado um novo algoritmo RF para segmentar clientes e na secção 4 é referido o inovador algoritmo Similis utilizado na determinação de cabazes de compras [Cavique e Themido 2001a] [Cavique 2002]. A associação dos dois algoritmos dá o nome ao modelo RF-Similis. Na secção 5 são exemplificadas algumas estratégias de comunicação com o cliente, tendo por base a segmentação realizada. Finalmente, na secção 6 são apresentadas as conclusões.

Em resumo, nesta comunicação pretende-se integrar técnicas conhecidas de Base de Dados de Clientes, como é o caso da análise RF, com algoritmos inovadores em Data Mining, quando nos referimos ao algoritmo Similis; resultando o modelo RF-Similis.

2 Base de Dados de Clientes

Na origem do CRM, ou Marketing Relacional, está o Marketing Directo, onde se pretende atingir o público alvo sem intermediários, utilizando uma comunicação personalizada.

Por forma a aumentar as baixas taxas de respostas *direct-mail* que orça, geralmente os 2%, as Bases de Dados de Clientes (*Database Marketing*) recorrem a técnicas de segmentação baseada no clientes mais recentes, mais frequentes e de maior valor monetário em compras, esta técnica é conhecida por RFM (recency, frequency, monetary) [Hughes 2000].

A venda de produtos para CRM apresenta duas abordagens distintas: o CRM operacional e o CRM analítico [Cavique e Themido 2001 b]. O CRM operacional vem acrescentar às Bases de Dados de Clientes a capacidade de integração de vários canais de comunicação, como os call-centers e os serviços de Internet. O CRM analítico substitui os métodos de segmentação por técnicas de Data Mining e valoriza o cliente através do "lifetime value".

Como já referimos, não vamos desenvolver o tema de CRM, mas antes, concentrar este estudo nas Bases de Dados de Clientes e incluir as técnicas necessárias para atingir o objectivo seguinte:

- *Com o apoio das Base de Dados de Clientes pretende-se construir relações lucrativas e duradouras, ao comunicar com o cliente certo, utilizando o produto certo, com a mensagem certa (emitida no momento certo e através do canal certo).*

Do objectivo em cima distinguem claramente três tipos de conjuntos de dados: os dados do cliente, os dados da compra do cliente e os dados relativos à comunicação com o cliente [Brito 2000]. Para responder a esta solicitação existe uma estrutura de dados, como se apresenta em seguida:

- 1) Dados do Cliente: nome, contactos, canal preferencial, datas mágicas, dados derivados (próximo-produto, rfm, número de reclamações, *lifetime value*), dados demográficos, sociográficos e psicográficos.
- 2) Compras do Cliente: data, produto, quantidade, preço, tipo pagamento, vendedor, descontos, devoluções, ofertas, pontos.
- 3) Comunicações com o Cliente (reactiva e pró-activa): data-hora, origem, destino, assunto, conteúdo da mensagem, canal.

As Bases de Dados de Clientes para além de registarem os dados do cliente e das suas compras, como qualquer sistema de contabilidade, integra o registo das comunicações. As formas de comunicação classificam-se em reactivas e pró-activas, reactivas se têm origem no cliente e pró-activas de têm origem na empresa. A (in)satisfação do cliente pode ser medida através da comunicação reactiva: os pedidos de ajuda e as reclamações. Por outro lado, existe um novo tipo de comunicação programada que envia ao cliente, em determinadas datas e com uma certa frequência, diferentes tipos de contactos ("touch points") simulando uma relação directa.

A simulação do ambiente personalizado das Bases de Dados de Clientes é alcançado através de uma estratégia de comunicação e pelo efeito combinatório dos produtos.

3 Algoritmo de Segmentação RF

A conhecida técnica de segmentação RFM, onde R (recentidade ou qualidade de ser recente) é dada pela última data da visita à loja, F representa a frequência de compras na loja e M o valor monetário global do cliente [Miglausch 2000] [Miglausch 2002].

Se escolhermos para cada atributo RFM, 5 classes, com os números de 1 a 5, obtemos 125 classificações diferentes. Assim o cliente 555 é um cliente muito recente, muito frequente e com um alto volume de compras, enquanto que um cliente 111 é pouco recente, pouco frequente e com baixo volume de compras. Existe uma possibilidade de hierarquizar as classificações, seguindo a valorização que é comum dar aos números inteiros.

O algoritmo de segmentação irá ordenar a tabela de clientes de forma crescente pela data da última compra, num segundo passo classifica os primeiros $n/5$ clientes com o número 1, os segundos $n/5$ clientes com o número 2 e assim sucessivamente até ao número 5. O processo repete-se para os atributos da frequência e valor monetário. A concatenação dos três atributos resulta a classificação RFM de cada cliente.

No algoritmo RFM cada célula contém um número igual de elementos. Desta forma para a classificação segundo a recentidade é utilizado uma única chamada a um procedimento de ordenação (*sort*). Para ordenar segundo o critério de frequência, o algoritmo RFM faz a chamada ao procedimento de *sort*, k vezes, sendo k o número de classes. Para terminar o algoritmo, para classificar M , o procedimento de *sort* é chamado k^2 vezes, por forma a obter que cada célula RFM tenha o mesmo número de clientes. Desta forma o algoritmo RFM utiliza o procedimento de *sort*, $1+k+k^2$ vezes.

Com base nesta segmentação muitas variantes podem ser ensaiadas e testadas, com base na experiência as taxas de resposta são maiores para a segmentação RFM do que para a segmentação FRM ou qualquer outra combinação de atributos.

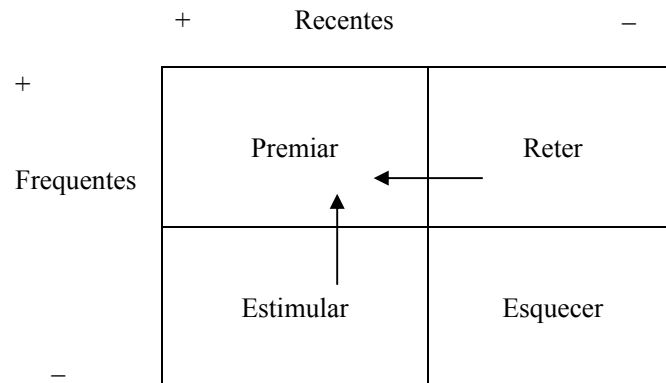


Figura 1 - Segmentação RF

Os clientes a seleccionar para cada campanha (ou micro-campanha) são os n -ésimos primeiros clientes com maior valor de RFM, permitindo uma única estratégia.

Nesta comunicação apresentamos um novo algoritmo inspirado na análise RFM, o algoritmo RF, que permite definir quatro estratégias diferentes de comunicação com o cliente. Dado que o atributo do valor monetário é função da frequência de compras vamos desprezar esta variável, mantendo os atributos R e F.

Para obter a classificação RF, basta ordenar pela data mais recente e afectar R como foi definido anteriormente. O processo repete-se para classificar F, sendo necessário recorrer ao procedimento de *sort* unicamente duas vezes. O resultado é uma matriz RF, cujas células têm valores diferentes de clientes.

Depois de classificar os clientes com base na recentidade e na frequência, podemos realizar a análise segundo as duas variáveis em simultâneo, dividindo os clientes em quatro quadrantes por forma a adoptar estratégias diferentes para cada grupo, conforme está representado na Figura 1.

Os clientes R^+F^+ são mais recentes, com maior frequência, as opções mais caras e com menos custos no processo de venda. Este tipo de clientes para além de serem os mais lucrativos, têm uma importância acrescida, recomendam a empresa a potenciais clientes, fazendo referência aos seus produtos/serviços originam novas aquisições. A estratégia a utilizar é de premiar a fidelização, que não se esgotam numa campanha, mas pelo contrário, têm um carácter permanente. Por outro lado aos clientes R^-F^- não deve recair qualquer esforço, dado que são clientes muito pouco rentáveis ou mesmo prejudiciais à empresa.

Para os clientes R^+F^- , recentes e pouco frequentes, que correspondem a novas aquisições da empresa, pretende-se estimular a compra por forma a migrá-los para R^+F^+ . O mesmo objectivo de migração existe para os clientes R^-F^+ , muito frequentes mas que não visitam a loja há muito tempo; neste caso devem ser implementadas estratégias de retenção. Para qualquer dos clientes R^+F^- e R^-F^+ a estratégia de cross-selling aparece com extrema importância.

4 Algoritmo Similis para determinação Cabaz de Compras

As estratégias de cross-selling são suportadas por algoritmos de análise do Cabaz de Compras.

Quando se fala em Cabaz de Compras compreende-se um conjunto de N produtos adquiridos por um cliente numa visita a um loja. Dado que cada cliente compra um conjunto diferente de produtos, em quantidades diferentes e em diferentes instantes de tempo, torna a determinação da Cabaz de Compras um problema complexo. Com base nos atributos escolhidos <cliente, produto> iremos definir o problema do Cabaz de Compras como a procura dos grupos mais frequentes de N produtos que são comprados em conjunto.

Definido o cabaz de compras tipo com N produtos, que represente um conjunto grande de clientes, podemos partir para as estratégias de ‘cross-selling’ e ‘up-selling’. Com a primeira estratégia, depois de determinado o cabaz, vamos procurar todos os clientes que tenham adquirido N-1 produtos desse cabaz e sugerir a compra do produto que falta - o Próximo Produto.

Para a determinação do cabaz de compras, numa das primeiras aproximações, destaca-se o programa Collaborative Filtering disponibilizado pela NetPerception, que determina para cada cliente a sua ‘alma gémea’. Entende-se por ‘alma gémea’ um outro cliente que tenham os mesmos gostos, i.e. que tenha um cabaz de compras idêntico. Este programa está instalado em centenas de empresas, tendo contudo a desvantagem de estarmos a comparar simplesmente dois indivíduos, não tendo a visão do conjunto. Ex: Se o indivíduo X comprou 10 livros sobre Internet e um livro de Culinária e o indivíduo Y comprou 8 livros sobre Internet, tal que $Compras(Y) \subseteq Compras(X)$, o programa irá sugerir a compra do livro de Culinária ao indivíduo Y.

Uma forma de representar o cabaz de compras é através de regras do tipo: “se <condição> então <acção>”. Esta abordagem da análise do Cabaz de Compras foi inicialmente utilizada na descoberta novos padrões, tendo sido popularizado pela descoberta do padrão que referia que “às 5ª feiras nos super-mercados, fraldas e cervejas são compradas em conjunto”. Com base neste padrão podemos gerar as seguintes regras: $\{fraldas\} \Rightarrow \{cerveja\}$ e $\{cerveja\} \Rightarrow \{fraldas\}$ [Berry e Linoff 1997]. Num dos pacotes de Data Mining mais conhecidos, o Enterprise Miner da SAS Institute, a análise do cabaz de compras está implementada utilizando com este tipo de gerador de regras, que recorre ao algoritmo Apriori [Agrawal et al. 1996].

Dada a elevada complexidade do algoritmo anterior, foi recentemente desenvolvido um algoritmo de mais baixa complexidade temporal. O algoritmo Similis [Cavique e Themido 2001a] [Cavique 2002] que permite resolver um maior número de problemas reais. O algoritmo Similis estende a aplicação da análise do cabaz de compras até às centenas de produtos, permitindo a sua aplicação em ambientes como hiper-mercados ou livrarias virtuais.

O algoritmo desenvolve-se em três passos. Dada uma tabela T, com os atributos <cliente, produto> é gerada uma matriz binária M, onde para cada elemento $M(i,j)$ de M, $M(i,j)=1$ significa que o cliente i comprou o produto j . Num segundo passo é gerado um grafo ponderado $G=(V,A)$ com base nas semelhanças dos produtos. O conjunto de vértices V corresponde ao conjunto de produtos do cabaz. O peso na aresta $(v_i, v_j) \in A$ é dado pela semelhança entre a produto v_i e o produto v_j . Finalmente para encontrar as cliques de maior peso no grafo $G=(V,A)$, que correspondem às soluções S de cabazes de compras mais frequentes, foi utilizada uma abordagem meta-heurística baseado em trabalho desenvolvido em Cavique, Rego e Themido [2002a] e em Cavique, Rego e Themido [2002b].

Em seguida apresentamos o procedimento para o Algoritmo SIMILIS, identificando as seguintes variáveis, onde:

- Ttrans é a tabela de transacções
- $[M(i,j)]$ é a matriz binária
- G é o grafo ponderado
- S^* é a clique de maior peso
- k é o número de produtos adquiridos em conjunto

O Algoritmo SIMILIS

1. como base em Ttrans gerar a matriz binária $[M(i,j)]$;
2. com base em $[M(i,j)]$, para cada aresta, calcular o peso para cada par de produtos (i,j) e gerar um grafo ponderado $G=(V,A)$;
3. encontrar no grafo ponderado $G=(V,A)$ a clique de maior peso, S^* , de dimensão k, que corresponde ao cabaz de compras mais frequente;

Nesta secção, depois de enquadrar e descrever o problema do cabaz de compras foi apresentado o algoritmo Apriori, implementado em pacotes comerciais como o Enterprise Miner da SAS Institute. Dada a elevada complexidade algorítmica do Apriori desenvolvemos o algoritmo Similis, de mais baixa complexidade computacional, permitindo resolver um maior número de problemas reais.

Os dois algoritmos apresentam características notoriamente diferentes nos seguintes pontos:

- Relativamente à estrutura de dados utilizada o algoritmo Apriori utiliza a tabela original enquanto que o algoritmo Similis organiza os dados num grafo ponderado.
- No procedimento para obtenção das soluções os algoritmos também divergem: o algoritmo Apriori utiliza um método de contagem exacto com complexidade exponencial, enquanto que o algoritmo Similis utiliza uma abordagem que recorre a meta-heurísticas.

- Quanto à informação devolvida, o algoritmo Apriori apresenta as frequências exactas dos cabazes de compras, enquanto que o algoritmo Similis, ao trabalhar com o problema transformado, utiliza a informação “imperfeita” do peso das cliques e apresenta os quase-exactos cabazes mais frequentes.
- Quanto ao volume de dados que cada um pode tratar, o algoritmo Apriori só trabalha com um pequeno volume de dados, enquanto que o algoritmo Similis pode tratar grandes volumes de informação.
- A complexidade temporal do algoritmo Apriori é dependente do número de produtos, do número de registos da tabela e da dimensão do cabaz; para o algoritmo Similis essa complexidade é independente do número de registos e da dimensão do cabaz, sendo crescente relativamente ao número de produtos que corresponde ao número de vértices do grafo.

Para validação do algoritmo Similis foram utilizados dois problemas reais [Cavique 2002]: o primeiro relativo às Pousadas de Portugal com 43 produtos e o segundo referente aos ultra-congelados da Nestlé com 158 produtos. O algoritmo Similis apresenta um bom desempenho tanto a nível da qualidade de resultados, com um índice de qualidade médio sempre superior a 94%, com tempos computacionais reduzidos. Foram ainda testadas várias medidas de semelhança por forma a validar o atributo mais sensível do algoritmo Similis, que é a noção de peso da clique, não se tendo encontrado diferenças significativas.

5 Estratégias de Comunicação

A orientação do Marketing centrada no produto, tem vindo a se deslocar para a orientação centrada no cliente. A preocupação já não está na qualidade do produto, nem mesmo do serviço; a aposta está na criação de relações lucrativas de duradouras com o cliente.

Os programas de fidelização das companhias de aviação provaram ser um excelente método para adquirir e reter clientes fieis. A melhor forma de recompensar a sua lealdade é dar em troca pontos, cupões ou outros benefícios. O desconto não deve ser usado, dado que pode ser copiado pela concorrência e provoca no cliente uma busca constante por um preço menor, indo no sentido inverso ao da fidelização. Para além dos benefícios na trocas de pontos e de referir os prémios de referência quando da aquisição de novos clientes.

A par da compensação com benefícios, devem ser combinadas dois tipos de contactos: mensagens orientadas para o cliente, sem qualquer motivação ao consumo, e mensagens relativas aos produtos. Como mensagens orientadas para o cliente teremos: uma simples carta de agradecimento por se tornar cliente, um cartão de boas-festas, uma cartão aniversário, mensagens enviadas em "datas mágicas" ou um contacto do gestor de clientes. O primeiro tipo de mensagem deve ser combinado com elementos orientados para o produto; tal como uma “newsletter”, um questionário de opinião acerca de um produto, um questionário de retenção depois de uma reclamação, uma proposta de um próximo produto ("cross-selling") ou um convite exclusivo à loja para apresentação de uma nova colecção, irá com certeza criar um ambiente favorável com o cliente.

A utilização destas formas de comunicação deve ser gerida de forma diferenciada para cada segmento de clientes. Utilizando a segmentação RF, apresenta-se na Figura 2 um exemplo de estratégia de comunicação. O ambiente de comunicação pró-activa com o cliente é previamente programado. Cada "ponto de contacto" corresponde a uma mensagem associada uma data precisa de emissão, simulando assim a personalização.

Toda a comunicação vai no sentido de tornar a relação com os clientes mais lucrativa e duradoura, forçando-os para a classificação de R^+F^+ . Para implementar a estratégia de cross-

selling é de toda a conveniência utilizar o algoritmo Similis, resultando esta combinação no modelo RF-Similis.

		+	Recentes	–
+				
Frequentes		<div>Troca de pontos</div> <div>Convites à loja</div> <div>Prémio referência</div>	<div>Questionário retenção</div> <div>Newsletter</div> <div>Cross-selling</div>	
–		<div>Agradecimento</div> <div>Newsletter</div> <div>Cross-selling</div>		

Figura 2 - Estratégias de comunicação

6 Conclusões

Esta comunicação divide-se em três grupos, que correspondem às formas de ver as Bases de Dados de Clientes: os dados, os modelos e a comunicação. Com a abordagem seguida concretiza-se, de uma forma integrada, os objectivos definidos para uma Bases de Dados de Clientes, onde se pretende construir relações lucrativas e duradouras, ao comunicar com o cliente certo, utilizando o produto certo, com a mensagem certa (emitida no momento certo e através do canal certo). O cliente certo associado à mensagem adequada é dado pela segmentação RF e o produto certo (ou próximo produto) pelo algoritmo Similis. Como procedimento geral podemos seguir a seguinte sequência:

armazenar dados dos clientes → segmentar os clientes → comunicar com os clientes

A re-alimentação do sistema é feita através das aquisições dos clientes, permitindo a actualização dos seus dados de compra, que conforme vimos são essenciais. No armazenamento de dados para além dos dados do cliente e das compras as Bases de Dados de Clientes incluem os dados relativos à comunicação.

Neste artigo apresenta-se um novo algoritmo para segmentação de clientes, o algoritmo RF e faz-se referência a um algoritmo recente para implementar estratégias de cross-selling, o algoritmo Similis. O primeiro, o algoritmo RF, apesar da sua simplicidade é de grande utilidade ao diferenciar quatro importantes estratégias. O algoritmo RF mostra-se superior ao algoritmo RFM, não só pela sua eficiência computacional mas também pela sua capacidade de definir um maior número de estratégias. O algoritmo Similis, apresenta-se como um algoritmo muito eficiente para a determinação de cabazes de compras. Da associação dos dois algoritmos é criado o modelo RF-Similis.

Com base na segmentação de clientes, pode-se partir para estratégias diversificadas de comunicação, que suportam a relação com o cliente. Na comunicação pró-activa (da empresa para o cliente) a existência de "pontos de contactos", a par da atribuição de benefícios, é a base do processo de fidelização.

Como comentário final, diremos que o desenvolvimento do modelo RF-Similis, permite a potenciação do algoritmo Similis [Cavique e Themido 2001a] [Cavique 2002]; em vez de ser

um algoritmo isolado de Data Mining, permite a integração numa Base de Dados de Clientes (*Database Marketing*) e na vida das empresas.

7 Referências

- Agrawal R., H. Mannila, R. Srikant, H. Toivonen e A. Verkamo "Fast discovery of association rules" in *Advances in Knowledge and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth e R. Uthurusamy (eds.), MIT Press (1996).
- Amaral L. e M. Santos, "Modelos de Estádios de Crescimento", *Revista de Sistemas de Informação*, pp.41-60 (1997).
- Berry M. e G. Linoff, *Data Mining Techniques, for Marketing, Sales and Customer Support*, John Wiley and Sons (1997).
- Brito C.M., "O Marketing Relacional" in *Os Horizontes do Marketing*, ed. Brito C.M. e Lencastre P., Editorial Verbo, Lisboa (2000).
- Brown S.A., *CRM- Customer Relashionshio Management*, John Wiley and Sons Canada Ltd (2000).
- Cavique L. e I.Themido, "A New Algorithm for the Market Basket Analysis", publicação Cesur-IST (2001a).
- Cavique L. e I.Themido, "Estratégias de Comunicação em CRM", in *E-Portugal*, L Valadares Tavares e M J Pereira, eds (aceite para publicação 2001b) .
- Cavique L., "Meta-heurísticas na Resolução do Problema da Clique Máxima e Aplicação na Determinação do Cabaz de Compras", dissertação de Doutoramento em Engenharia de Sistemas no Instituto Superior Técnico da Universidade Técnica de Lisboa (2002).
- Cavique L., C.Rego e I.Themido, "A Scatter Search Algorithm for the Maximum Clique Problem" in *Essays and Surveys in Metaheuristics* (Ribeiro C. e Hansen P. Editors), Kluwer Academic Publishers, pp. 227-244 (2002a).
- Cavique L., C.Rego e I.Themido, "Estruturas de vizinhança e algoritmos de procura local para o problema da clique máxima", *Revista de Investigação Operacional*, vol. 22, pp. 1-18 (2002b).
- Hughes A.M. *Database Marketing*, McGraw-Hill Companies (2000).
- Miglautsch J.R., "Application of RFM principles: What to do with 1–1–1 customers?", *The Journal of Database Marketing*, vol. 9, no. 4, pp. 319-324(6) (July 2002).
- Miglautsch J.R., "Thoughts on RFM scoring", *The Journal of Database Marketing*, vol. 8, no. 1, pp. 67-72(6) (August 2000).
- Reis J.L. *O Marketing Personalizado e as Tecnologias de Informação*, Centro Atlântico, Lda, Colecção Sociedade de Informação, Matosinhos (2000).