UNIFY - Sistema de Unificação de Cadastros: Uma visão de marketing

Fabiano Santana David, Professora Margrit Krug

Curso de Análise e Desenvolvimento de Sistemas
Faculdade de Tecnologia SENAC RS (FATEC/RS)
Porto Alegre – RS – Brasil

david.fabiano@gmail.com, margrit.senac@gmail.com

Resumo: Este trabalho apresenta uma ferramenta que utiliza processos e técnicas de *data mining*, a qual visa auxiliar as empresas a descobrir quem realmente são seus clientes, ao mesmo tempo em que define sua localização, favorecendo a equipe de marketing da empresa condições de criar e manter ações de relacionamento eficazes, obtendo melhores resultados e, otimizando os recursos da organização.

Abstract: *In this work a tool was developed utilizing Data Mining processes and techniques that helps enterprises find out who and where their potential customers are, and therefore the marketing team is able to create and maintain an efficient relationship action plan to achieve better results, optimizing the organization resources.*

1. Introdução

É comum encontrar no mercado empresas que se preocupam com seus sistemas apenas na área operacional e comercial, dando pouca atenção para a área de marketing, sem saber direito o que fazer com os dados de compras, e com o cadastro dos clientes, informações essas, valiosíssimas para tomada de decisões, e antecipação aos anseios dos seus clientes.

A fim de facilitar o conhecimento da empresa com relação ao histórico de consumo de seus clientes, um banco de dados bem estruturado, sem duplicidades e principalmente, no que se refere aos dados pessoais dos clientes, mantido o mais atualizado quanto possível e, além disso, ser único em todas as filiais da empresa deve ser adotado.

Ainda hoje, existem empresas que não possuem um cadastro de clientes unificado, como também é comum a fusão entre empresas, onde a real necessidade de unir seus cadastros é eminente. O que ocorre é que a empresa necessita conhecer seu consumidor, para torná-lo um cliente fiel, em contra partida só é um cliente fiel àquele que está satisfeito e para satisfazê-lo é necessário surpreendê-lo positivamente, para tanto é necessário conhecê-lo.

É comum ouvir do departamento de marketing das empresas que "é mais caro para a empresa trazer um cliente de volta, após uma experiência ruim, do que mantê-lo na empresa", mas poucas empresas sabem o porquê isso acontece. Em uma palestra de marketing, apresentada por João Stringhini[1], o palestrante mostrou o significado dessa frase: devido ao comportamento e a atitude das pessoas. Comportamento é o que as pessoas praticam, e atitude é o que pessoas pensam. Então, as pessoas podem ter comportamento positivo ou negativo e atitude positiva ou negativa em relação a uma empresa, produto ou serviço. Um exemplo prático:

- Uma pessoa vai sempre a um mesmo restaurante porque gosta dele, ou seja, está tendo um comportamento positivo e uma atitude positiva. Porém, um dia, nesse restaurante o garçom deixa cair refrigerante no seu colo, o cliente releva. Num outro dia a comida não está muito boa, essa pessoa começa a pensar que o restaurante não é mais o mesmo, mas continua freqüentando, ou seja, está tendo um comportamento positivo e uma atitude negativa, mas como ainda freqüenta o restaurante, o dono do estabelecimento não percebe essa atitude, e acha que está tudo normal. Certo dia abre um restaurante novo e esse cliente já insatisfeito, troca de restaurante e não aparece mais no antigo, ou seja, tem um comportamento negativo e uma atitude negativa, porém para o restaurante já é tarde, ele não percebeu essa mudança a tempo, e agora é muito caro trazer esse cliente de volta.

O gráfico abaixo (Figura 1) apresentado na palestra mostra a relação entre o comportamento e a atitude das pessoas e a sua relação com as vendas de uma empresa.

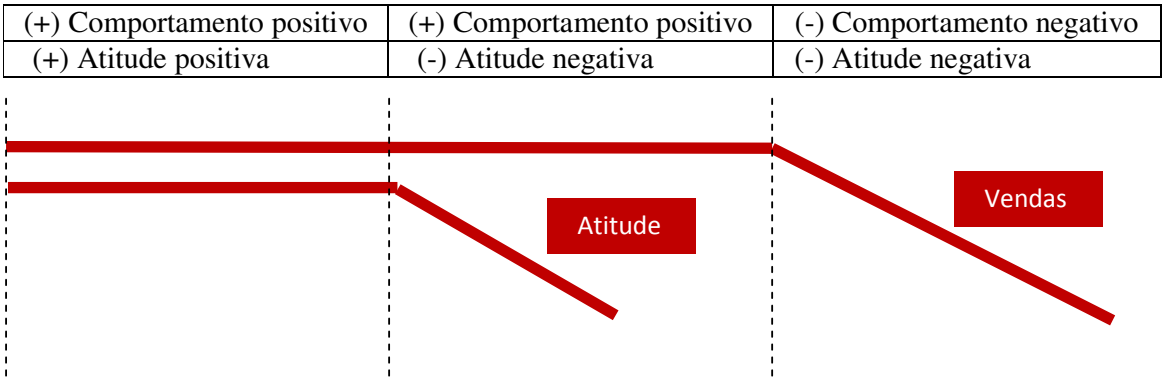

Figura 1 – Relação entre Comportamento e Atitude

Pode-se perceber que a atitude é um item que cai bem antes das vendas, portanto se a empresa não monitorar o comportamento dos seus clientes, no dia a dia, só se dará conta que o cliente estava insatisfeito quando as vendas caem, e aí é muito tarde e caro para trazê-lo de volta. Então, a máxima do marketing diz que: *"É muito mais barato, custando de três a cinco vezes menos manter um cliente do que obter um novo."* (Artigo: Marketing Viral: Tornando

[1] João Stringhini é diretor da Stringhini Marketing, formado em Psicologia – PUCRS, tem pós-graduação em Administração de Empresas – Especialização em Marketing pela PUCRS e mestre em Manegement e Marketing Estratégico pela UCES de Buenos Aires.

os clientes a mídia mais eficiente para propagar a marca da empresa. Daniela Camarinha[2]) se dá por causa desse fenômeno.

Neste trabalho desenvolveu-se uma ferramenta que auxilia as empresas a descobrir quem é realmente seu cliente e onde ele está, para que a equipe de marketing tenha condições de criar e manter ações de relacionamento eficazes que tragam resultados, otimizando os recursos da organização.

Na seção 2 será descrito o problema que motivou o desenvolvimento deste trabalho. Na seção 3 o objetivo e o estudo das possíveis soluções serão apresentados. A subseção 3.2 conceituará *data mining*, pois utilizaram-se neste trabalho alguns métodos utilizados por esta área de conhecimento. Os algoritmos de comparação encontrados na bibliografia e utilizados na implementação deste trabalho serão apresentados na seção 4. A seção 5 apresentará os detalhes da ferramenta desenvolvida. Finalmente, na seção 6, as conclusões obtidas durante a execução deste trabalho.

2. O Problema

Uma empresa, qualquer, necessita que os registros do cadastro de clientes em seu banco de dados sejam comparados por semelhança, pois não existe a unificação do banco de dados entre as filiais. Esses dados do cadastro deverão ser unificados em uma outra tabela do seu banco de dados, a fim de manter uma visão única dos clientes.

Isso ocorre porque esta empresa recebe diariamente os registros de cadastro de clientes, provenientes de vendas ou atualizações cadastrais, ou simplesmente através de formulários de contato de seus web site, e estes são armazenados sem controle de duplicidade, ou padronização dos dados recebidos.

Esse processo, de certa forma descontrolado, inviabiliza uma visão de marketing sobre os clientes, impossibilitando uma implantação confiável de CRM[3] ou um sistema de apoio a decisões como BI[4] que utiliza OLAP[5]. Os sistemas em geral, apenas transformam dados em

---

[2] Daniela Camarinha é Administradora de Empresas, Pós-graduanda em Gestão Empresarial pelo Instituto Trevisan e em curso de MBA em Marketing de Serviços e Comunicação. Gerente Comercial e de Marketing do SAE Laboratório Médico e professora da pós-graduação da Faculdade São Camilo.

[3] **CRM Customer Relationship Management** é uma expressão em inglês que pode ser traduzida para a língua portuguesa como **Gestão de Relacionamento com o Cliente**. Foi criada para definir toda uma classe de ferramentas que automatizam as funções de contacto com o cliente, essas ferramentas compreendem sistemas informatizados e fundamentalmente uma mudança de atitude corporativa, que objetiva ajudar as companhias a criar e manter um bom relacionamento com seus clientes armazenando e inter-relacionando de forma inteligente, informações sobre suas atividades e interações com a empresa.

[4] O termo **BI** - **Business Intelligence**), pode ser traduzido como **Inteligência de negócios**, refere-se ao processo de coleta, organização, análise, compartilhamento e monitoramento de informações que oferecem suporte a gestão de negócios.

informações, e isso não é mais suficiente para tomada de decisões. As ferramentas de CRM e BI utilizam técnicas para transformar informações em conhecimento. O conhecimento permitirá que os gestores possam se antecipar aos fatos, surpreendendo seus clientes.

Uma outra situação comum, hoje em dia, são empresas em processo de aquisição de outras, onde necessitam incorporar em seu banco de dados atual, os dados dos cadastros de clientes existentes nos bancos de dados das empresas que estão sendo agregadas a esta. Como as futuras empresas que serão adquiridas, muito provavelmente, são do mesmo ramo de atuação, prevê-se um alto índice de duplicação de cadastro.

3. Estudo de Alternativas para a Solução do Problema

A solução proposta para o problema descrito anteriormente é possibilitar uma visão única dos dados dos clientes, assim, tem-se os objetivos específicos:

- comparar por semelhança os dados dos cadastrais dos clientes e unificá-los;
- permitir a parametrização dos campos a serem comparados;
- permitir parametrização do grau de confiabilidade da unificação dos registros;
- gerar uma tabela única de registros unificados;
- gerar uma tabela de relacionamento entre a tabela unificada e a tabela original;
- documentar as regras de unificação dos registros.

3.1 Algoritmos de comparação

Para o desenvolvimento de uma solução de software para esse problema, inicialmente pesquisou-se algoritmos de comparação, pois se verificou que se fosse desenvolvido um algoritmo, proprietário, esta seria uma tarefa complexa, além do fato de se estar "reinventando" a roda, pois existem na literatura vários algoritmos que já fazem isso, tais como: Levenshtein Metric, Smith Waterman, Stochastic Model, Jaro Metric, Hamming Distance, Soundex Distance Metric, Covington´s distance function, Autômato Não Determinístico (NFA), BLASTQ-gram. *(Todos esses algoritmos foram estudados e documentados em: GODIM, Flávio Melo, Recife 2006. Algoritmo de Comparação de Strings para Integração de Esquemas de Dados).*

Com isso, a solução encontrada para o problema foi o desenvolvimento de um sistema de Data Mining[6] (Mineração de Dados), que diante do cenário, onde já existem algoritmos de

---

[5] **OLAP - On-line Analytical Processing** - é a capacidade para manipular e analisar um largo volume de dados sob múltiplas perspectivas.As aplicações OLAP são usadas pelos gestores em qualquer nível da organização para lhes permitir análises comparativas que facilitem a sua tomada de decisões diária.

[6] **Data Mining -** "Refere-se à garimpagem ou descoberta de novas informações em termos de padrões ou regras oriundas de grandes quantidades de dados." (ELMASRI e NAVATHE, 2002)

comparação e técnicas de mineração de dados, a tarefa deste trabalho foi analisar e compreender os processos para isso, e qual a ferramenta de software se utilizar para o desenvolvimento da solução.

3.2 Data Mining

Data Mining é um termo genérico utilizado para todas as novas técnicas computacionais para a extração de informações úteis a partir de grandes conjuntos de dados armazenados, onde o objetivo é dar suporte à tomada de decisão com base em dados.

> Data Mining como parte do processo de descoberta de conhecimento, freqüentemente abreviada como KDD (*Knowledge Discovery in Databases*), compreende seis processos:

a) Seleção de dados;

b) Limpeza de dados;

c) Pré-processamento (enriquecimento);

d) Transformação ou codificação de dados;

e) Análise, assimilação, interpretação, avaliação, e;

f) Divulgação e exposição das informações descobertas.

3.3 Solução Proposta

Utilizar as técnicas e processos de Data Mining para encontrar registros iguais, comparados por semelhança. Os processos que serão adotados para atender os objetivos serão compostos por: Parametrização, Simulação**,** Unificação e Relatórios. Esses processos são detalhados na Figura 2.

---

[6] **Data Mining -** "A data mining ajuda na extração de novos padrões significativos que não podem ser encontrados por mera consulta ou processamento ou processamento de dados ou metadados no data warehouse." (ELMASRI e NAVATHE, 2002)

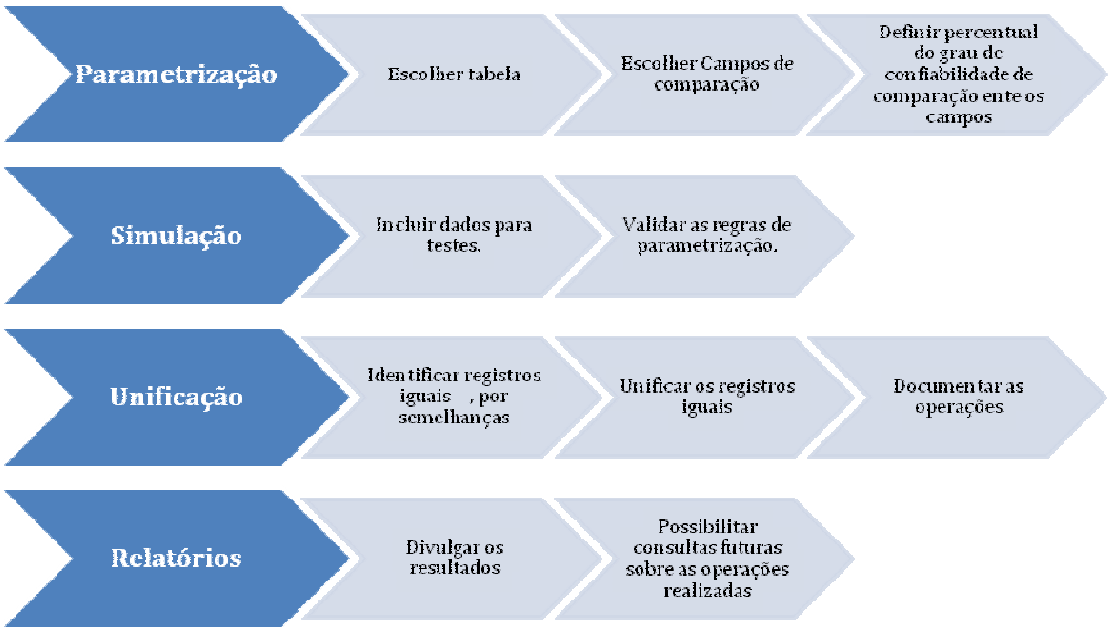

Figura 2 – Processos utilizados na Solução Proposta

3.4 Análise de Requisitos

Análise de Requisitos é o estudo das características que o sistema possui para atender às necessidades e expectativas de cada cliente.

Cada funcionalidade demandada pelo cliente deve ser analisada a fim de verificar os possíveis impactos no desenvolvimento das demais funcionalidades do sistema, e verificado em conjunto com a equipe de desenvolvimento se as necessidades tecnológicas para a implementação estão disponíveis, e atenderão as necessidades do cliente.

3.4.1 Requisitos diretos

Para atender as restrições do sistema e atender as regras de negócio da empresa, determinou-se que os requisitos diretos desta ferramenta seriam:

a) Gerar histórico de regras;
b) Gerar histórico de regras aplicadas a cada registro unificado;
c) Escolher os campos a serem unificados;
d) Escolher o percentual de comparação dos registros;
e) Gerar uma tabela de registros unificados;
f) Trabalhar com tabelas com uma ou mais chaves, e;
g) Fornecer relatórios, com:
    1. Quantidade de registros da tabela origem.
    2. Quantidade de registros unificados, e;
    3. Visualização dos registros relacionados.

3.2.2 Requisitos Indiretos

Além dos requisitos diretos o sistema possui como requisitos indiretos, principalmente:

a) Padronização e normalização prévia de registros de endereço;
b) Utilização da linguagem de programação PHP 5.2.5 para o desenvolvimento do software em questão, e;
c) Adoção do sistema gerenciado de banco de dados Microsoft SQL Server 2000.

4. Implementação

Antes de iniciar a implementação realizou-se pesquisas a fim de identificar os recursos disponíveis no banco de dados utilizado pela organização, a qual utiliza Microsoft SQL Server[7] versão 2000. As pesquisas iniciaram-se pelas funções nativas do banco de dados, no contexto deste trabalho duas, as quais poderiam ser aplicadas nessa solução: SOUNDEX[8] e DIFFERENCE.

A função SOUNDEX tem a seguinte sintaxe, SOUNDEX (*character_expression*) e retorna um código de quatro letras, que poderá ser comparado com o código de outra palavra. O algoritmo dessa função baseia-se no som produzido pela expressão, e não pela escrita da expressão.

A função DIFFERENCE tem a seguinte sintaxe, DIFFERENCE (*character_expression*, *character_expression*) e retorna um valor inteiro que é a diferença entre o valor SOUNDEX das duas expressões enviadas como parâmetro para a função.

O problema de evoluir nessa solução era a questão de que a função SOUNDEX baseia-se no som produzido na língua Inglesa, durante os testes se mostrou ineficiente com dados em Português.

Como os recursos do banco de dados não foram suficientes para atender os objetivos, começou-se, então, a pesquisar sobre as funções disponíveis no PHP[9], pelo fato de ser a linguagem de desenvolvimento preferencial da organização. No PHP versão 5.2.5, versão utilizada para pesquisa, descobriu-se duas funções possíveis de aplicação: levenshtein()[10] e

---

[7] Microsoft SQL Server - http://www.microsoft.com/brasil/servidores/sql/default.mspx

[8] SOUNDEX – Algoritmo fonético - http://en.wikipedia.org/wiki/Soundex

[9] PHP (um acrónimo recursivo para "PHP: Hypertext Preprocessor") é uma linguagem de programação de computadores interpretada, livre e muito utilizada para gerar conteúdo dinâmico na World Wide Web (Internet). http://pt.wikipedia.org/wiki/PHP

[10] Levenshtein — Compara duas expressões e informa um valor de similaridade entre elas. (http://www.php.net/manual/pt_BR/function.levenshtein.php)

similar_text()[11].

A função levenshtein(), na sua forma mais simples, tem a sintaxe *int levenshtein (string str1, string str2),* e retorna apenas o número de operações de inserção, substituição e deleção necessárias para transformar str1 em str2. Esta função nos testes realizados mostrou-se difícil de utilizar sob o aspecto da formulação de critério para parametrização de grau de similaridade.

A função similar_text() tem a seguinte sintaxe *int similar_text (string first, string second [, float percent]).* A similaridade entre duas strings como descrito em Oliver (1993), utiliza chamadas recursivas as quais podem ou não tornar todo o processo mais rápido. Passando por referência o terceiro argumento, similar_text() irá retornar a similaridade em porcentagem entre a primeira String e a Segunda String. Não é escopo deste trabalho, mas a título de curiosidade a complexidade deste algoritmo é O(N**3), onde N é o tamanho da maior string.

5. A Ferramenta

Esta seção tem como objetivo expor as funcionalidades da ferramenta, bem como as restrições impostas pela regra do negócio e pela própria, possibilitando seu funcionamento.

Tendo em vista o que foi exposto anteriormente, o sistema desenvolvido permite ao usuário do sistema informar qual a tabela que contém os dados de cadastro dos clientes, permite, também, que o usuário configure os campos que deverão ser comparados por semelhança definindo ainda o percentual de confiabilidade da comparação entre os registros. O sistema desenvolvido foi denominado como ***Unify*** e utiliza processos demonstrados na Figura 2.

Ao final do processo de unificação o sistema gera outra tabela no banco de dados com os registros unificados, devidamente relacionados com a tabela original. O sistema também gera relatórios do processo de unificação, informando a quantidade de registros originais na tabela, quantidade de registros unificados, além de permitir a visualização dos registros relacionados.

O sistema mantém um histórico das regras utilizadas e quais registros foram unificados através desses critérios. Como regra de negócio o sistema ao utilizar uma regra em um processo de unificação, não permite sua alteração.

Para realizar a comparação dos campos de endereço utiliza-se como pré-requisito que os registros estejam normalizados e padronizados, pois a falta de padrão poderá dificultar a precisão na comparação dos registros, mas não inviabiliza a utilização o sistema. O sistema ***Unify*** não faz a limpeza dos dados e/ou qualquer tratamento de padronização, pressupõe-se que essas tarefas já foram efetuadas em processo anterior. Por tanto, é importante que ao utilizar o sistema o usuário saiba desta restrição que está devidamente documentada e é conhecimento de todos que utilizarão este processo.

[11] Similar_text — Compara duas expressões e informa o valor percentual de similaridade entre elas. (http://www.php.net/similar_text)

Para realizar a comparação dos registros foi adotado um critério onde há a necessidade de pelo menos 02 (dois) campos do cadastro de cliente. Isso porque com apenas um campo, seja ele qual for não é possível afirmar, com alto índice de certeza, que os dois registros são de uma mesma pessoa, especialmente quando os dados não estão completos, por exemplo:

Um registro onde encontra-se cadastrado o nome como "Fabiano" e um outro registro onde o nome está cadastrado como "Fabiano David", neste exemplo não pode-se afirmar que referem-se a uma mesma pessoa, precisa-se de mais dados, tais como CPF, endereço ou data de nascimento.

5.1 Técnica

Para cada campo, escolhido pelo usuário, do cadastro de clientes após analisado pelo processo de Aplicação de Critérios (vide Processo 3.2 Aplicação de Critérios do anexo 3), foi atribuído uma pontuação. Para isso foi criada uma matriz de Critérios de Pontuação (vide anexo 5), que serve como base para o sistema utilizá-la no desenvolvimento dos algoritmos que irão realizar a comparação dos registros, um a um, e na identificação dos registros semelhantes, agrupá-los e tomar a decisão de quais são os dados que devem formar os registros unificados. O sistema além de possuir um algoritmo proprietário de comparação e unificação, com base na função similiar_text(), levará em consideração o conhecimento do usuário sobre a sua base de dados e sua área de atuação, possibilitando a parametrização do percentual de semelhança.

Para isso, os processos: seleção de dados, limpeza de dados, pré-processamento, transformação, análise e divulgação das informações, são atendidos conforme o detalhamento a seguir.

5.1.1 Seleção de dados

O Usuário escolherá a tabela que será unificada, os campos de comparação e o percentual de aceitação de igualdade (Figura 3).

O sistema irá comparar todos os registros que estejam exclusivamente na tabela de origem, utilizando o processo 3.1 Separação de Registros (conforme DFD nível 1, anexo 3).

5.1.2 Limpeza de dados

O sistema assume que os dados já passaram por tratamento e padronização por algum procedimento externo. Por exemplo, a padronização do endereço através do DNE® - Diretório Nacional de Endereços[12].

5.1.3 Pré-processamento (enriquecimento)

O Sistema após a comparação de registros agregará na tabela origem uma

[12] DNE® - Diretório Nacional de Endereços é um banco de dados de abrangência nacional constituído de elementos de endereçamento até nível de secção de logradouro e Códigos de Endereçamento Postal – CEP.

identificação de quais registros ele considera iguais. Para isso utilizará o processo “Aplicação de Critérios” (vide Processo 3.2 Aplicação de Critérios do anexo 3).

Este processo recebe como entrada dois registros para comparação e os critérios, definidos pelo usuário, contidos na regra de comparação. O processo “Aplicação de Critérios” é composto de sub-processos, onde cada um é responsável por analisar uma determinada variável e cálculo da pontuação final. Cada variável analisada é pontuada, enviando ao sub-processo “Calcular Pontuação Final” as notas obtidas, e este processo define se os registros são iguais ou não. A informação final de igualdade ou não entre os registros é gravada na tabela origem, como saída do processo “Aplicação de Critérios”.

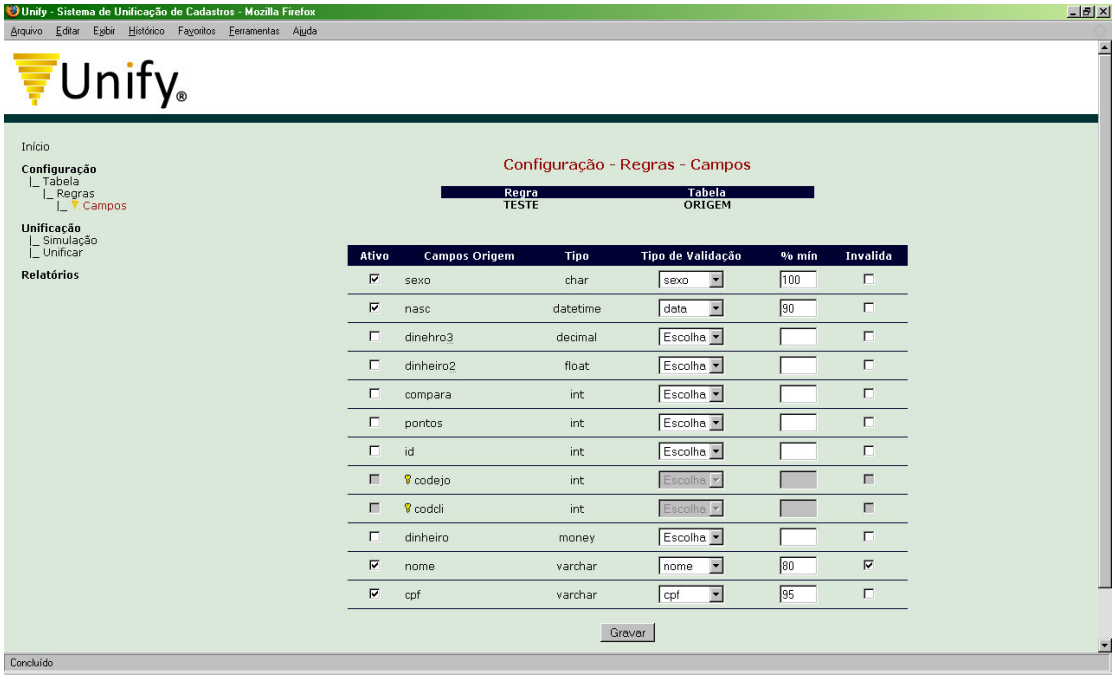

Figura 3 - Tela de Parametrização de Campos da tabela a ser unificada

5.1.4 Transformação ou codificação de dados

Os dados analisados e unificados são gravados em uma tabela independente do cadastro original. Na seqüência, é gerado o relacionamento, entre a tabela origem e a unificada, seguindo os critérios definidos pelo usuário. Cada registro da tabela é comparado com os registros que ainda não foram associados, até que se esgotem os registros da tabela original. A escolha dos dados para formar o registro final é atendida no caso de uso *Escolher dados para registro final* (demonstrado na Figura 4).

5.1.5 Análise, assimilação, interpretação e avaliação

A função similar_text() do PHP foi escolhida para fazer a análise de cada bloco de comparação, onde utilizou-se mais uma tabela de pontuação para cada tipo de valor analisado (nome, CPF, data, etc.) conforme o caso de uso *Aplicação de Critérios*. Cada caso de uso

definido na parametrização do sistema fornecerá uma pontuação, então se somam todos os pontos de cada bloco e têm-se a pontuação final.

Para determinar a pontuação necessária que definirá se dois ou mais registros são iguais por semelhança, somam-se os pontos atingidos por cada campo no processo de aplicação de critérios, de acordo com a tabela de critérios pontuação (vide anexo 5). Como a pontuação máxima de um campo é 3 e, como dito na seção 5, que são necessários pelo menos dois campos, chegou-se a pontuação necessária igual ou superior a 04 (quatro), para considerar que dois registros são iguais. Essa função foi a que obteve melhores resultados durante os testes de comparação de dados, e aderiu perfeitamente aos critérios estipulados na matriz de Critérios de Pontuação (vide anexo 5).

Depois de selecionados os registros o sistema identifica aqueles que são semelhantes, para então realizar a Unificação dos Registros, utilizando os critérios para cada campo (nome, datas, endereço, etc.) para assim gerar um único registro.

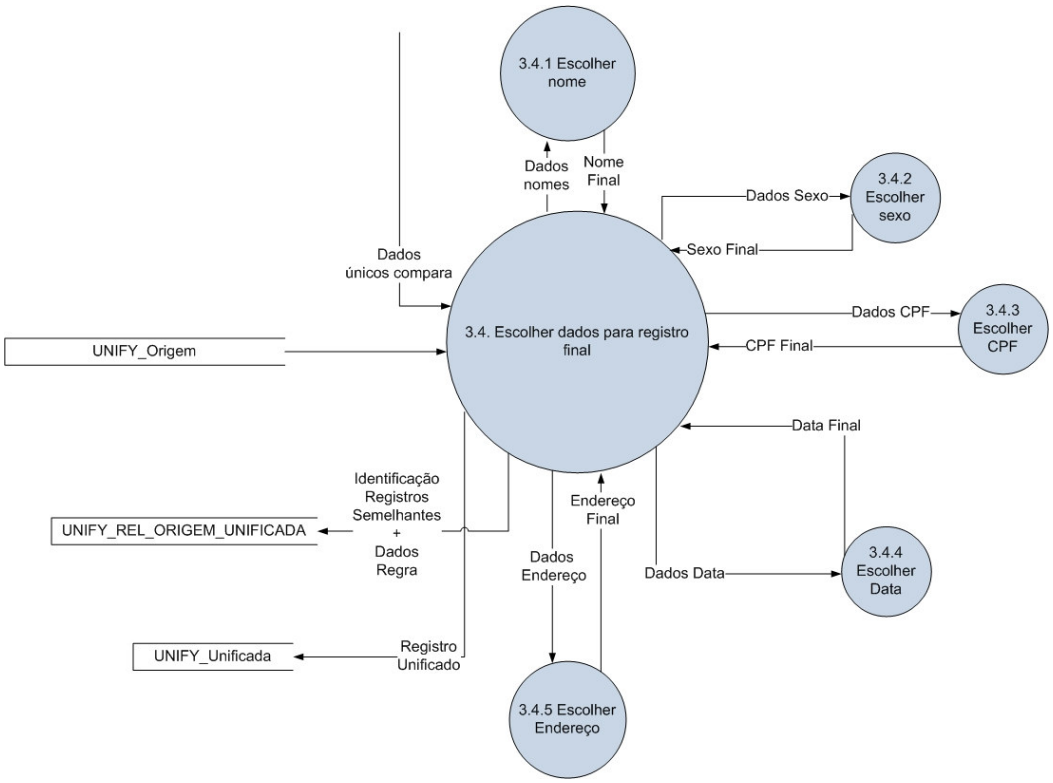

Figura 4 - Escolha dos dados para a geração do registro final

5.1.5 Divulgação e exposição das informações descobertas

O sistema desenvolvido apresenta e disponibiliza relatórios com os resultados gerados após a unificação dos registros. Os dados unificados estão disponíveis em uma tabela independente da tabela original, mantendo o relacionamento e histórico da unificação. O sistema também mantém o histórico das regras utilizadas em cada processo de unificação.

Conclusões e Trabalhos Futuros

Inicialmente, havia-se previsto que o sistema deveria trabalhar apenas com uma tabela e uma empresa específica. Porém, verificou-se que seria mais interessante desenvolver um sistema mais flexível, que atendesse qualquer empresa e qualquer estrutura de tabela, e assim foi desenvolvido esse trabalho.

O sistema tem como objetivo unificar uma tabela com dados cadastrais, evitando a duplicação de informações, especialmente quando bases de dados são aproveitadas de outras empresas que foram incorporadas à uma instituição, especialmente porque, geralmente as empresas individualmente, possuem clientes em comum, e quando agregadas estes tornam-se clientes de uma empresa apenas (isso porque estas se tornaram uma única empresa). O Software desenvolvido pode ser utilizado por uma empresa que presta o serviço de duplicação e unificação de dados para várias empresas, para isto o sistema cria uma estrutura de trabalho independente para cada empresa, facilitando a exportação dos dados e integração com outros aplicativos.

Como base para os testes foram utilizadas três amostra reais, que chamaremos de Amostra "A", "B" e "C" de empresas distintas.

A amostra "A" é composta de registros de endereços normalizados, e as amostras "B" e "C" são compostas por registros de endereços não normalizados de acordo com a base do Diretório Nacional de Endereços dos correios.

Os resultados obtidos estão demonstrados na Tabela 1 e os critérios de comparação utilizados estão demonstrados no Anexo 6 – Regras e critérios utilizados nas amostras de teste.

Tabela 1 - Tabela de Resultados dos Testes

Amostras	Reg. Origem	Reg. Unificados	Redução	Tempo
A	2.135	1.294	39%	3 min.
B	6.369	2.272	64%	10 min.
C	12.705	5.563	56%	30 min.

A Tabela 2 mostra a projeção de economia em uma campanha de acionamento de clientes via mala-direta, após a utilização do sistema ***Unify*** utilizando as amostras dos testes acima mencionados.

Tabela 2 – Simulação Prática Aplicada

ACIONAMENTO MALA-DIRETA				
Amostras	Custo unit.	Antes	Depois	Economia
A	R$ 2,00	R$ 4.270,00	R$ 2.588,00	R$ 1.682,00
B	R$ 2,00	R$ 12.738,00	R$ 4.544,00	R$ 8.194,00
C	R$ 2,00	R$ 25.410,00	R$ 11.126,00	R$ 14.284,00

Mesmo com as amostras “B” e “C”, onde os registros não estavam com os dados de endereçamento normalizados, e, como dito anteriormente, esse é um pré-requisito para se obter melhores resultados, o sistema obteve alto grau de confiabilidade nos registros unificados, comprovados nos relatórios apresentados no próprio sistema, e como mostra a tabela acima (Tabela 1).

Cabe ressaltar que nenhuma grande disparidade foi identificada, e os registros com dados insuficientes para unificação ficaram em torno de 0,007%, ou seja, esses registros com análise manual e com conhecimento das informações poderiam ser unificados com outros registros contidos da amostra.

Esses resultados se mostraram amplamente satisfatórios, devido ao fato que a não padronização dos campos de endereço nas amostras “B” e “C” não comprometeram o trabalho e também pelo fato de nessas amostras tinha-se apenas o nome e o endereço como dados de comparação. Com isso, pode-se tirar a conclusão de que se em uma tabela onde se utilize apenas o nome e os dados de endereço como variáveis de comparação, não será possível determinar se o cliente mudou seu endereço, para isso seria necessário mais dados.

Com o algoritmo atual o sistema ganha desempenho à medida que os registros são identificados. Isso ocorre porque após um registro ser comparado com todos os outros, este não é mais analisado, pois seus “pares” já foram encontrados. Isso diminui o escopo de procura para o próximo registro a ser analisado. Outro algoritmo poderá ser implementado, e facilmente ajustado no sistema atual, para que todos os registros sejam comparados entre todos, isso faria que um registro pudesse ser igualado com outro através de um terceiro. Na prática esse último algoritmo diminuirá o desempenho, e teoricamente poderá ter mais informações para a tomada de decisão.

O sistema atual não trabalha com variáveis de e-mail e telefone, porque não foi avaliado, até esse momento, qual é o peso dessas variáveis e o impacto nos critérios de pontuação. Isto porque são variáveis complexas, pois o e-mail e o telefone, em uma empresa, são dados que podem ser compartilhados por mais de uma pessoa, neste caso o sistema deverá classifica-los em dados pessoais ou comerciais.

Atualmente com a possibilidade de portabilidade do número de telefone, esse passa a ser uma variável importante para identificação de uma pessoa. Ao implementar esses requisitos, o algoritmo de pontuação deverá ser revisto, pois seu peso influenciará na nota final dependendo da combinação escolhida.

O sistema desenvolvido atenderá totalmente os objetivos propostos, com desempenho adequado e custo viável. E com o uso e avaliação do sistema no estado em que se encontra identificou-se novas funcionalidades, que serão logo implementadas, sendo elas: possibilidade de copiar uma regra e simulação direta na tabela origem; comparação e unificação do telefone e e-mail; processos para unificação sistemática comparando com a tabela já unificada; estudo de viabilidade de inclusão de regras de comparação fonética em português; possibilidade de o usuário escolher o também o banco de dados; comparação de registros entre uma ou mais tabelas; implementar módulos normalização de endereços, cleasing (limpeza de registros), pois são processos relacionados à Data Mining.

Bibliografia

CASTAGNETTO, Jesus *et al*. **Professional PHP Programando**. São Paulo: Makron Books, 2001

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de Banco de Dados.** 3.ed. Rio de Janeiro: LTC Editora, 2002.

GARCIA-MOLINA**,** Hector. **Implementação de sistemas de bancos de dados.** Rio de Janeiro: Campus, 2001.

GONDIM, Flávio Melo. **Algoritmo de Comparação de Strings para Integração de Esquemas de Dados**. Trabalho de Graduação. UNIVERSIDADE FEDERAL DE PERNAMBUCO, GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, CENTRO DE INFORMÁTICA. Recife, Fev/2006.

GONDIM, Flávio Melo. **Integração de Informações em Ambientes Heterogêneos: arquiteturas, modelos e implementações. Sub-Projeto: Enriquecimento Léxico de Esquemas para um Sistema de Integração de Dados**. RELATÓRIO FINAL DE ATIVIDADES DO ALUNO DE INICIAÇÃO CIENTÍFICA. PIBIC/UFPE/CNPq. Recife, Set/2005.

GUNDERLOY, Mike; JORDEN Joseph L. **Dominando SQL Server 2000 "A Bíblia".** São Paulo: Makron Books Ltda., 2001.

MAXIMILIANO**,** Luiz Souza Guimarães.**Técnicas de Descoberta do Conhecimento em Sistemas de Apoio à Decisão**. Artigo técnico do Unicentro Newton Paiva. (sem data)

SANTOS, Rafael. **Princípios e Aplicações de Mineração de Dados.** Ministério da Ciência e Tecnologia – Instituto Nacional de Pesquisas Espaciais. Trabalho de pesquisa apresentado na disciplina de "Princípios e Aplicações de Mineração de Dados". (sem data).

SCHNEIDER**,** Luís Felipe**. Mineração de Dados (*Data Mining*)** – Trabalho de pesquisa apresentado na disciplina "Tópicos Avançados em Modelos de Banco de Dados" da UFRGS. (sem data)

SILBERCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. **Sistema de Banco de Dados.** 5.ed. Rio de Janeiro: Elsevier, 2006.

SOARES, Walace. **PHP 5: Conceitos, Programação e Integração com Banco de Dados**. 1. Ed. São Paulo: Érica, 2004.

STRINGHINI**,** João. **Dicionário Stringhini de Marketing -** Dicionário de Termos e Expressões em Marketing. Porto Alegre: Sul Editores, 2007.

Outras Referências

Wikipédia:

BI

http://pt.wikipedia.org/wiki/Business_inteligence

CRM

http://pt.wikipedia.org/wiki/Customer_relationship_management

LEVENSHTEIN()

http://www.php.net/manual/pt_BR/function.levenshtein.php

Mineração de dados

http://pt.wikipedia.org/wiki/Minera%C3%A7%C3%A3o_de_dados

OLAP

http://pt.wikipedia.org/wiki/Cubo_OLAP

http://pt.wikipedia.org/wiki/OLAP

SIMILAR_TEXT()

http://www.php.net/similar_text

SOUNDEX

http://en.wikipedia.org/wiki/Soundex

Sites:

http://www.php.net/

http://www.phpmania.org/modules.php?name=php_how_to&page=function.levenshtein.html

http://www.php-welt.net/handbuecher/brazilian-portuguese/function.levenshtein.html

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-89102008000100008&lng=e&nrm=iso&tlng=e

http://www.phpgratis.com.br/manual/php/function.similar-text.html

http://www.php.net/similar_text

http://br.geocities.com/dugimenes/