1 Introdução

Com a evolução do mundo digital, cada vez mais o volume de dados armazenados e conseqüentemente analisados cresce de forma extremamente acelerada. Esses dados são encontrados nos mais diversos formatos, línguas, unidades de medida, e graus de confiança, gerando um problema crítico na tomada de decisão já que não há certeza sobre sua qualidade. Segundo Ted Friedman analista do Gartner Group, a maioria das empresas não se dá conta do impacto que os problemas na qualidade dos dados podem representar. Como demonstração desta subavaliação temos que mais de 50% dos projetos de *data warehouse* teriam seu sucesso comprometido devido à qualidade dos dados em 2007.

Nos últimos 25 anos pudemos acompanhar seguidas tentativas de implantar projetos ambiciosos de CRM[1], BI[2], ou ainda outras técnicas de suporte a decisão. Porém, a falta de consistência e robustez nos dados fez com que diversas dessas tentativas fossem frustradas. Ainda segundo o Gartner Group, a baixa qualidade nos dados pode elevar os custos e, nos casos mais acentuados, até mesmo impossibilitar a conclusão de projetos desperdiçando enormes volumes de recursos.

Como exemplo dessa inconsistência nos dados, temos o relatório do Nielsen Media Research indicando que de outubro de 2006 até setembro de 2007 aproximadamente 21% das correspondências foram endereçadas erroneamente somente no Reino Unido.

O mercado potencial de sistemas que melhorem ou mantenham a qualidade dos dados vem crescendo muito na esteira do crescimento do CRM. De acordo o Gartner Group, o mercado mundial de CRM cresceu a uma taxa anual de 23% em 2007 impulsionado pelo crescimento dos países emergentes.

No que tange ao mercado brasileiro de CRM, o IDC Brasil reportou em 2004 um crescimento de 7,1% frente a uma média mundial de 6,9%.

---

[1] CRM: *Customer Relationship Management* – Gestão de Relacionamento com o Cliente.

[2] BI: *Business Intelligence* – Inteligência de Negócio.

O IDC afirma também que no Brasil o mercado de *software* movimentou US$ 1,74 bilhão em 2004 e que o país tem 46% de participação no setor na América Latina.

A fatia do mercado brasileiro que utiliza e aplica conceitos de CRM é composta por empresas dos setores industrial, financeiro, petróleo e mineração, saúde e educação, telecomunicações e mídia, entre outros. A Figura 1 a seguir mostra a segmentação do mercado brasileiro de CRM.

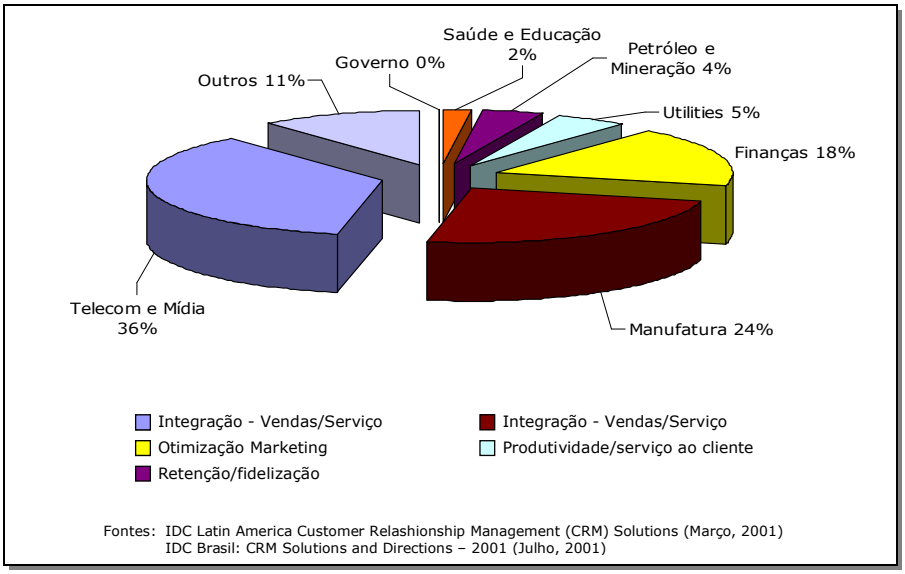

Figura 1: CRM no mercado brasileiro

Ainda segundo o IDC Brasil, para os próximos anos, espera-se um crescimento no total de receitas geradas mercado de *software* brasileiro a taxas de 23% até 2009.

Para garantir, melhorar ou manter a qualidade dos dados, uma série de técnicas é largamente utilizada pelos diversos sistemas disponíveis no mercado. Esse trabalho encara o agrupamento de palavras semelhantes como uma ferramenta utilizada para auxiliar a desambiguação de dados ou ainda eliminar possibilidades de duplicatas.

Com o aumento constante no volume de dados para aplicação de processos que avaliam ou melhoram qualidade, a velocidade na resolução de cada parte do processo tem se tornado ainda mais importante. Sendo assim, o objetivo principal é encontrar um método que possua uma relação custo/benefício mais interessante que aquelas usualmente utilizadas.

Neste trabalho tratamos especificamente de acelerar o processo de deduplicação através da divisão do conjunto de dados em subgrupos cujos itens sejam o mais parecidos possível. No caso ótimo cada subgrupo deve conter todas as duplicatas de cada registro reduzindo o erro do agrupamento à zero. Estabelecemos, porém, uma tolerância aceitável de 5% de erro após os agrupamentos.

Hoje, a identificação de duplicatas de registros textuais em bases de dados muito grandes representa um grande desafio. Em casos onde a base é relativamente pequena, esse problema pode ser endereçado de diversas formas, inclusive através da confrontação dos pares e submissão dos mesmos às regras que os identificariam como duplicatas. Seguindo essa linha de raciocínio, o agrupamento dos registros textuais é utilizado como redutor do tamanho do problema original deixando-o manuseável sem registrar perda de qualidade significativa.

Em um mundo sem restrições computacionais poderíamos comparar todos os registros dois-a-dois e, ao final do processo, teríamos o resultado esperado, ou seja, todas as duplicatas devidamente identificadas. Qualquer comparação dessa natureza em uma base de dados é **$O(n^2)$**, sendo *n* a quantidade de registros da base. Para os atuais padrões computacionais teríamos a relação ilustrada pela Tabela 1.

Quantidade de registros	Tempo
1.000	2 segundos
10.000	3 minutos
100.000	5 horas
1.000.000	23 dias
10.000.000	6,3 anos
Os dados acima são empíricos obtidos através de testes realizados supondo a comparação de registros textuais com aproximadamente 25 caracteres	

Tabela 1: Tempos para comparação dois-a-dois

Ao reduzir um problema de 10.000.000 registros como o da tabela acima em 1.000 problemas de 10.000 registros é possível reduzir o tempo de execução de 6,3 anos para menos de 60 horas. O agrupamento, portanto, torna o problema computável em um tempo que permite que esse tipo de solução seja aplicado no dia-a-dia.

Esta drástica redução dos tempos de processamento e conseqüente desoneração do processo é um benefício com impacto direto na qualidade de dados.

Para estabelecer uma base para a posterior discussão acerca da deduplicação e de seus complicadores, o documento segue com uma explicação de todo o processo de qualidade de dados e as atividades envolvidas. Após a descrição do processo e de seus passos é apresentado um conjunto de conceitos básicos necessário para completo entendimento do problema.

Em seguida são descritos os experimentos que relatam precisamente os resultados obtidos os quais mostram que, através das estratégias adotadas, os objetivos são completamente atingidos, indicando que existem várias possibilidades de melhoria e evolução para este trabalho.