Análise Visual de Base de Dados – Um Estudo de caso

Carlos H. Timoteo[1], Edgar W. Almeida[1], Maxwell F. Queiroz[1]

[1]Departamento de Engenharia da Computação – Universidade de Pernambuco(UPE)
Caixa Postal 15.064 – 91.501-970 – Recife – PE – Brazil

{chmst, ewma, mqf}@ecomp.poli.br

Abstract. *The need for efficiency during organization decision process requires the use of solutions that generate consistent information. It is in this context that fall BI (Business Intelligence) tools. An initial question in any BI project is the choice of tooling support to be used in the development. Visual Data Analysis is a promising approach to transform a overloaded information into opportunity. This article presents a case study conducted on a database of scientific publications after the comparative analysis between two free tools BI: Tableau and Pentaho. As the first tool proved itself incapable of analyzing dynamic data, we focused on solving some simple questions that should be answered by visual tool Tableau. As a result, we present some results in a visual way.*

Resumo. *A necessidade de eficiência no processo decisório das organizações exige a utilização de soluções que gerem informações consistentes. É nesse contexto que se inserem as ferramentas de BI (Business Intelligence). Uma questão inicial em qualquer projeto de BI é a escolha do ferramental de apoio a ser utilizado no desenvolvimento. A análise visual de dados é uma abordagem promissora para transformar uma informação sobrecarregada em oportunidade .Este artigo apresenta um estudo de caso realizado numa base de dados de trabalhos científicos, após a análise comparativa entre duas ferramentas de BI livres: Pentaho e Tableau. Devido a deficiências da primeira, focou-se na resolução de algumas questões simples a serem respondidas pela ferramenta Tableau. Como resultado, alguns resultados são apresentados de forma visual.*

1. Introdução

Estamos vivendo em um mundo que enfrenta uma quantidade crescente de dados a ser tratados em uma base diária. Na última década, a melhoria constante de dispositivos de armazenamento de dados e meios para criação e coleta de dados influenciou nossa forma de lidar com a informação: A maior parte do tempo, os dados são armazenados, sem filtragem e refinamento para, assim, seu uso futuro. Praticamente todos os ramos da indústria ou de negócios, e qualquer atividade política ou pessoal geram grandes quantidades de dados. Para piorar a situação, as taxas de coleta e armazenamento de dados estão mais aceleradas do que a nossa capacidade de usá-lo para tomar decisões. No entanto, na maioria das aplicações, os dados brutos não tem valor por si mesmo, em vez disso, queremos extrair informações neles contidos [1].

Existe o conceito de que o sucesso na gestão do conhecimento depende da informação correta, estando disponível no momento certo, no lugar certo e da forma devida. Hoje em dia, a aquisição de dados brutos não é mais o problema de corrente, mas, é a capacidade de identificar métodos e modelos, que podem transformar os dados em conhecimento confiável e comprovável. Qualquer tecnologia, que alega superar o problema da sobrecarga da informação, tem de fornecer respostas para os seguintes problemas:

- Quem ou o que define a "relevância da informação" para uma determinada tarefa?

- Como procedimentos adequados em um processo de tomada de decisão complexa podem ser identificados?

- Como pode a informação resultante ser apresentada para suportar decisões ou orientar as tarefas?

- Que tipo de interação pode facilitar a resolução de problemas e a tomada de decisão?

A visão atual da análise visual é transformar a informação sobrecarregada em uma oportunidade. Logo, como visualização da informação mudou a nossa forma de ver bancos de dados, o objetivo da análise visual é fazer com que o nosso modo de processamento de dados e informações sejam transparentes do ponto de vista analítico. A análise visual promoverá a avaliação construtiva, correção e melhoria rápida dos nossos processos e modelos e - em última instância - a melhoria do nosso conhecimento e nossas decisões.

Aplicações da vida real necessitam de análise visual de dados, tais como (i) bases de dados biológicos armazenando genes em massa e conjuntos de dados de proteínas, (ii) sistemas de monitoramento em tempo real acumulando múltiplos conjuntos de dados, sob múltiplos fluxos de fontes, (iii) sistemas de inteligência de negócios (BI) avançados coletando dados de negócio para a tomada de decisão.

O objetivo desse trabalho é realizar uma análise de ferramentas para extração de informação visual dinâmica de base de dados para auxiliar a tomada de decisão. As duas principais ferramentas do mercado foram analisadas: Pentaho[2] e Tableau[3]. Após isso foi conduzido um estudo de caso real para a extração de informação de uma base de dados de publicações de trabalhos num congresso, algumas perguntas foram respondidas para auxiliar a tomada de decisão.

2. Revisão Bibliográfica

2.1. Análise Visual

Thomas e Cook[4] definem análise visual como a ciência do raciocínio analítico facilitada por interfaces visuais interativas. No entanto, existe uma definição mais específica: Análise visual combina técnicas de análise automatizada com visualizações interativas para uma efetiva compreensão, raciocínio e tomada de decisão sobre base de dados muito grandes e complexas. O objetivo da análise visual é a criação de ferramentas e técnicas que permitam as pessoas:

- Sintetizar informações e conseguir o discernimento de massa, dinâmico, ambíguo,

e muitas vezes entre dados conflitantes.

- Detectar o esperado e descobrir o inesperado.

- Fornecer avaliações oportunas, defensáveis, e compreensíveis.

- Comunicar avaliação efetiva para a ação.

As características citadas acima satisfazem as necessidades de uma ferramenta para análise de dados grandes e complexos, e portanto a ferramenta que estamos procurando necessita dessas características.

2.2. Extração, Transformação e Carga

Os processos de software que facilitam o preenchimento do armazém de dados (data warehouse) são comumente conhecidos como Extração - Transformação - Carga- (Extraction – Transformation - Loading - ETL).

Processos de ETL são responsáveis por extração dos dados apropriados das fontes; o seu transporte para uma área do armazém de dados com propósito especial, onde ele será processado; a transformação dos dados de origem e do cálculo de novos valores (e, possivelmente, registros), a fim de obedecer a estrutura da relação do armazém de dados para as quais são orientados; isolamento e limpeza de tuplas problemáticas, a fim de garantir que as regras de negócio e restrições do banco de dados sejam respeitadas; o carregamento dos dados limpos e transformados para a relação adequada no armazém, junto com a atualização de seus índices de acompanhamento e visualizações materializadas.

Como se pode observar, um processo de ETL é a síntese de tarefas individuais que executam extração, transformação, limpeza ou o carregamento de dados num gráfico de execução - também referido como um fluxo de trabalho. Além disso, devido à natureza dos artefatos de design e a interface do usuário das ferramentas ETL , um processo de ETL é acompanhado por um plano de execução [5].

Ferramentas para análise visual de dados trabalham de forma semelhante a um armazém de dados, portanto essas ferramentas necessitam que um processo de ETL seja implementado para que as visualizações sejam geradas.

2.3. VizQL

Linguagens de consulta convencionais, tais como SQL e MDX têm capacidades de formatação e visualização limitadas. Assim, embora consultas potentes possam ser construídas, uma nova camada de software é necessária para relatar ou apresentar os resultados de uma forma útil para a analista. VizQL é projetada para preencher essa lacuna. VizQL evoluiu a partir do Polaris na Universidade de Stanford, que combina consulta, análise e visualização em uma única estrutura [6].

VizQL é uma linguagem formal para descrever tabelas, gráficos, mapas, séries cronológicas e tabelas de visualizações. Esses diferentes tipos de representações visuais são unificados em um quadro, tornando mais fácil mudar de uma representação visual para outra (por exemplo, a partir de uma visão de lista para uma tabulação cruzada de um gráfico). Ao contrário dos pacotes gráficos atuais e semelhante a linguagens de consulta, VizQL permite um número ilimitado de expressões. Visualizações podem, assim, ser facilmente customizadas e controladas. VizQL é uma linguagem declarativa.

A imagem desejada é descrita, as operações de baixo nível necessárias para obter os resultados, para realizar cálculos analíticos, para mapear os resultados para uma representação visual, e para processar a imagem são geradas automaticamente pelo analisador de consulta.

O analisador de consulta compila expressões VizQL para SQL e MDX e, portanto, VizQL pode ser usado com banco de dados relacionais e cubos de dados. A implementação atual suporta Hyperion Essbase, Microsoft SQL Server, Microsoft Analysis Serviços, MySQL, Oracle, bem como fontes de dados, tais como CSV e Excel. Esta ferramenta de análise inclui muitas otimizações que permitem grandes bancos de dados serem pesquisados interativamente. VizQL possibilita uma nova geração de ferramentas de análise visual que casam consulta, análise e visualização [6].

2.4. Business Intelligence

De acordo com [7], *Business Intelligence* (BI) é um conjunto de tecnologias que tem como objetivo prover e oferecer suporte a um ambiente de informacão. A necessidade de eficiência e agilidade no processo decisório nas instituições exige delas a utilização de soluções que gerem informações consistentes e ao mesmo tempo sejam flexíveis de modo a se enquadrar nas suas necessidades e limitações. Dessa forma, é necessário efetuar uma análise dessas instituições e das ferramentas do mercado de modo a verificar quais delas são compatíveis.

Existe uma grande quantidade de ferramentas de BI no mercado atualmente, com uma ampla variedade de funcionalidades e valores. Além disso, têm ocorrido um grande crescimento e amadurecimento das soluções livres. No entanto, como não existe um padrão estrutural e funcional seguido por todos os processos de comparação entre essas ferramentas é dificultado, podendo levar a uma escolha demorada e não necessariamente correta da ferramenta. Nesse contexto, é inserido o risco decorrente da ferramenta não corresponder na realidade àquilo que está explicitado nos manuais.

2.5. Pentaho

Pentaho é um sistema de código aberto para business intelligence, desenvolvido em linguagem Java. O Pentaho consiste dos seguintes componentes: Pentaho Data Integration, Pentaho Analysis Services, Pehtaho Reporting, Pentaho Data Mining, Pentaho DashBoard, Pentaho for Apache Hadoop.

No estudo de caso, optou-se por utilizar a ferramenta Pentaho BI Platform. O Pentaho BI Platform é uma plataforma escalável, centrada a processo, habilitada a fluxo de trabalho para resolver problemas de business intelligence. Comumente conhecida como a plataforma de BI, e recentemente renomeada como Business Analysis Platform (BA Platform). Essa ferramenta apresenta as seguintes características:

- A plataforma de BI provê um framework e serviços que incluem registro, auditoria, segurança, programação, ETL, web services, repositório de atributos e motores de regras.

- As capacidades de usuário final de BI incluem relatório, análise, fluxo de trabalho, painel e mineração de dados.

- Pentaho Design Studio é um conjunto de ferramentas de projeto e administração que são integradas num ambiente Eclipse popular. Essas ferramentas permitem aos analistas de negócios criar relatórios, painéis, modelos de análise, regras de negócio e processos de BI [2].

2.6. Tableau

O Tableau constrói análises visuais que ajudam pessoas a perguntar e responder questões analíticas envolvendo dados armazenados em banco de dados e planilhas. Tableau permite que pessoas possam analisar dados de forma melhor usando sua habilidade natural de pensar visualmente. O Tableau combina gráficos, banco de dados e análise de dados em um framework de análise visual unificado. Esse framework é baseado em cinco princípios. Três princípios são apresentados a seguir.

O primeiro princípio, é uma interface fácil. Uma ferramenta de análise visual deve ter uma interface de usuário projetada cuidadosamente que realize a geração de consultas facilmente. E se você pudesse analisar informações somente criando a visualização que apresenta a resposta para sua pergunta? Se essa interface pudesse ser inventada, ela pode trazer para a análise exploratória uma nova classe de trabalhadores do conhecimento.

Segundo princípio, exploração de dados. Tableau ajuda o usuário a pensar visualmente. Não é somente criar visualizações de alta qualidade. É essencialmente um sistema de análise visual interativa para responder questões. Todas as vezes que você compõe uma figura, Tableau automaticamente compõe as consultas e computações analíticas necessárias para criar a figura.

O terceiro princípio é a expressividade. A maioria dos aplicativos de análise depende de modelos, assistentes, widgets e seus tipos de gráficos associados para fornecer visualizações de dados. Isto é verdade para o Excel, mas também é verdade de empresas de inteligência de negócios para aplicações Tableau baseia-se numa abordagem fundamentalmente diferente: Tableau é baseada em uma linguagem de consulta declarativa visual (VizQL) que é infinitamente expressivo e combinável. Esta é a tecnologia no coração da suíte de produtos da Tableau [3].

3. Análise de ferramentas

A metodologia utilizada nesse trabalho segue os seguintes passos:

- Pesquisa de ferramentas;

- Análise de viabilidade das ferramentas para satisfação dos requisitos do trabalho;

- Pré-tratamento dos dados e executar processo de ETL para as ferramentas;

- Implementar as visualizações que respondem as perguntas pré-definidas;

Durante a análise de viabilidade das ferramentas para satisfação dos requisitos, verificamos que a versão utilizada do Pentaho BI Platform não permite uma análise dos dados de forma dinâmica. Portanto, essa ferramenta foi descartada da abordagem do nosso estudo. De contra partida, a ferramenta Tableau mostrou-se uma abordagem bastante promissora e com um nível elevado de qualidade.

Logo, a nossa metodologia decidiu focar completamente na ferramenta Tableau para que pudéssemos abordar os conhecimentos que envolvem essa tecnologia e estudar a sua viabilidade.

2.6. Apresentação da base de dados

Para esse estudo, utilizamos como estudo de caso, a base de dados gerada a partir do processamento digital de todos os trabalhos científicos apresentados e publicados no Simpósio Brasileiro de Telecomunicação (SBRT) [8].

A base de dados analisada contém dez tabelas, as condições de normalidade, e as premissas para a realização do ETL também foram satisfeitas. No entanto, os dados armazenados estão duplicados ou inconsistentes. O modelo relacional é apresentado na Figura 1.

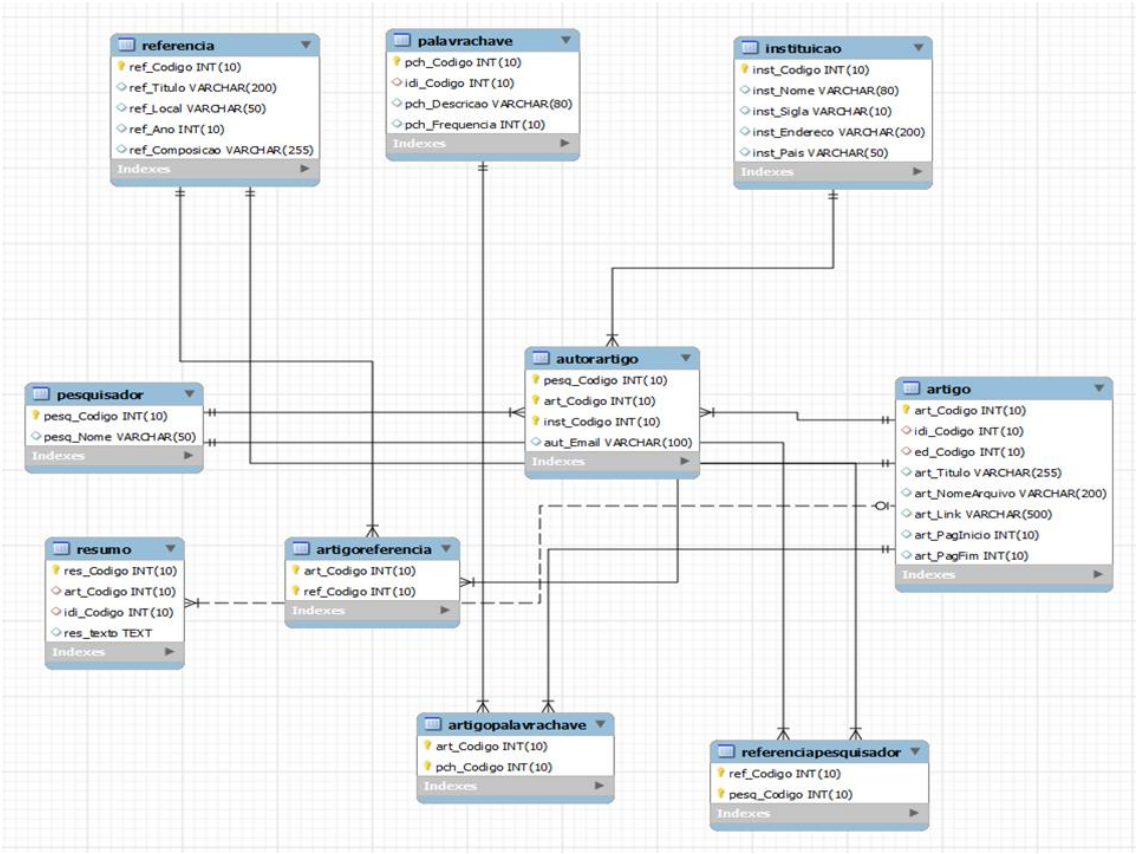

Figura 1. Modelo relacional do banco de dados

4. Análise de Resultados

Para o estudo de caso, foram elaboradas perguntas específicas referente a base de dados para

1. Qual autor mais publicou na SBRT?

2. Qual referência da SBRT foi mais citada?

3. Qual referência de todos os eventos foi mais citada?

4. Quais as palavras-chaves mais citadas?

5. Quem mais se auto referencia?

Após realizer o processo de ETL, os resultados para as perguntas estabelecidas acima foram obtidos. A Figura 2 apresenta de forma visual qual o autor que mais publicou na SBRT. A Figura 3 apresenta a referência mais citada. A Figura 4 a referência de todos os eventos mais citada. A Figura 5 apresenta as palavras-chaves mais citadas. Por fim, a Figura 6 apresenta quem mais se auto referencia.

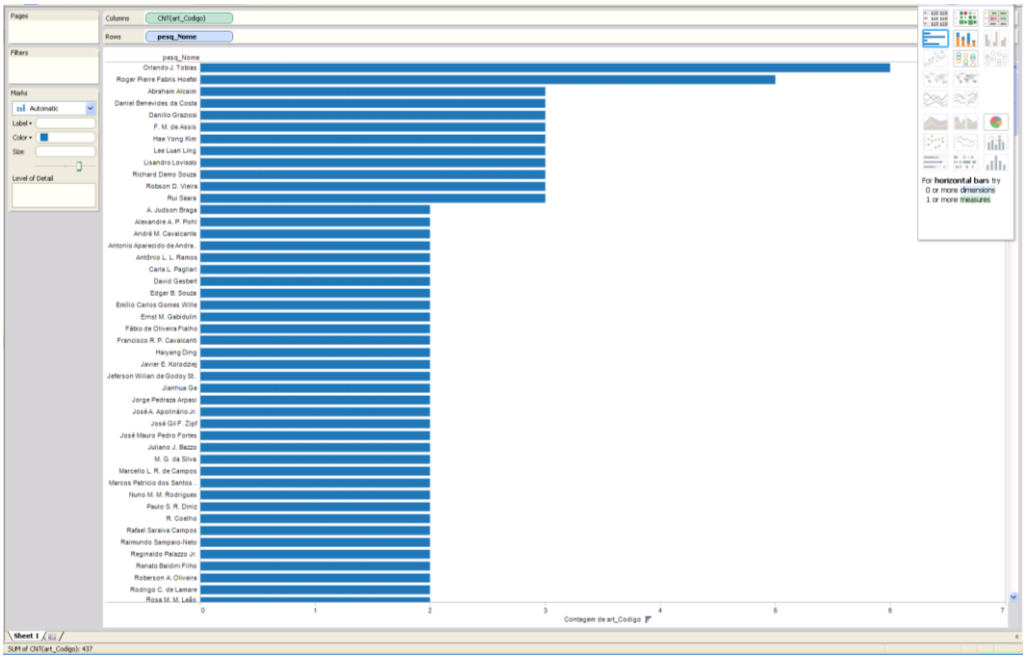

Figura 2. Qual autor mais publicou na SBRT

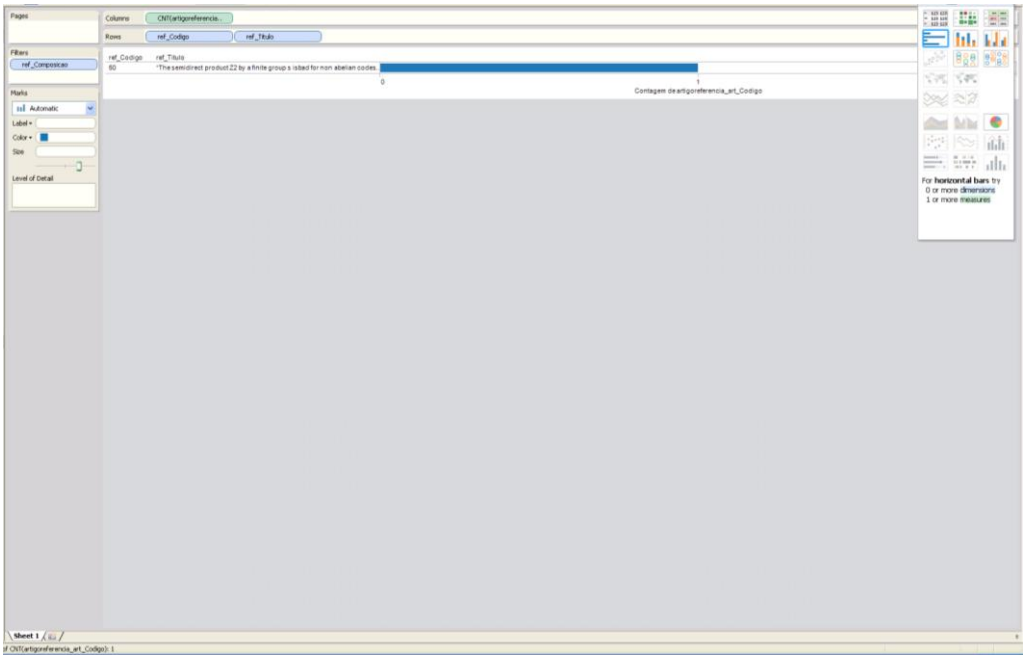

Figura 3. Qual referência da SBRT foi mais citada

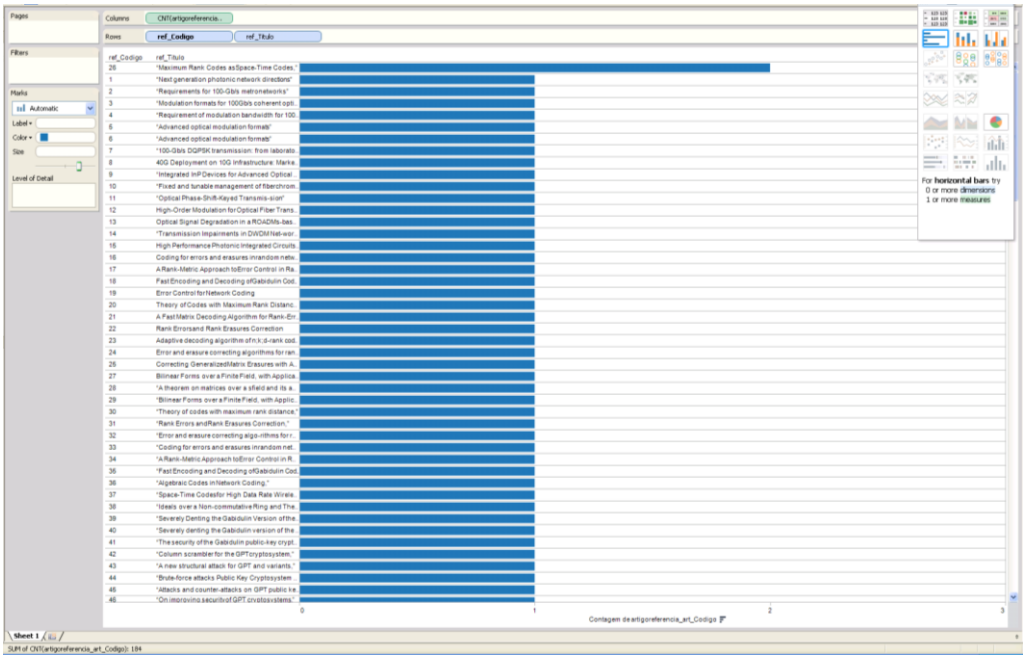

Figura 4. Qual referência de todos os eventos foi mais citada

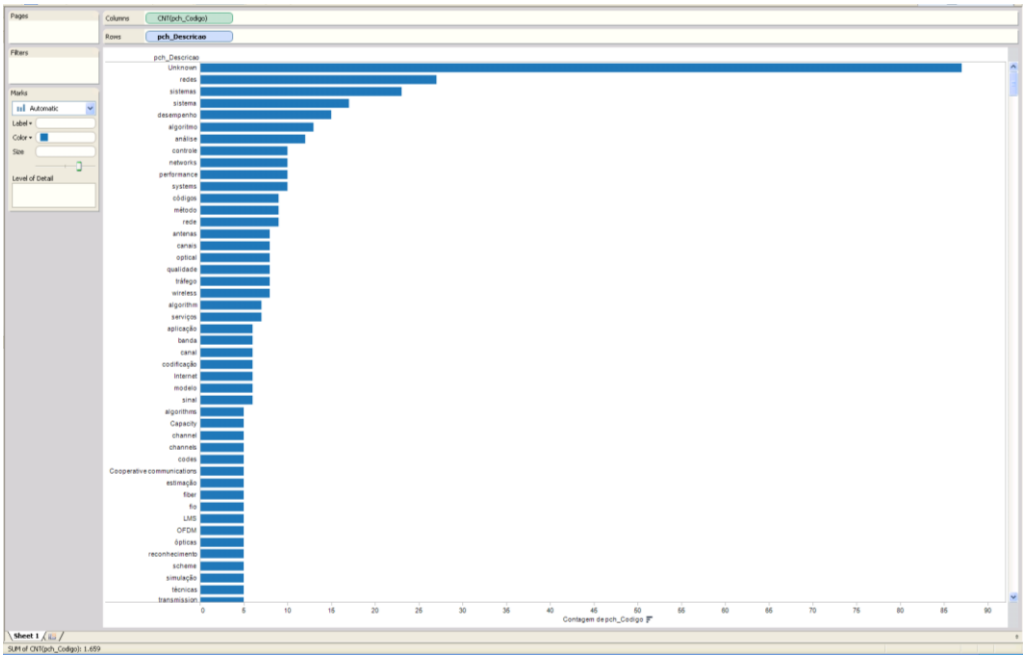

Figura 5. Quais as palavras-chaves mais citadas

5. Conclusão

Podemos concluir que a ferramenta Tableau é uma alternativa promissora e bastante eficaz para a análise visual de banco de dados. A análise dos resultados demonstram que a ferramenta mostrou-se bastante fácil, ágil e com bons recursos gráficos para representação dos resultados gerados pelas consultas criadas pelo analista de negócios.

Referências

[1] Keim, D.A., Andrienko, G., Fekete, JD., Gorg, C., Kohlhammer, J., Melançon, G: Visual Analytics: Definition, Process, and Challenges. A. Kerren et al. (Eds.): Information Visualization, LNCS 4950, pp. 154–175, 2008.

[2] Pentaho (2012). Pentaho BI. http://www.pentaho.com/ - Acessado em 02/12/2012.

[3] Tableau (2012). Tableau Software. http://www.tableausoftware.com/ - Acessado em 02/12/2012.

[4] Thomas, J.J., Cook, K.A.: Illuminating the Path. IEEE Computer Society Press, Los Alamitos (2005).

[5] Vassiliadis, P. A Survey of Extract-Transform-Load Technology. International Journal of Data Warehousing & Mining, 5(3), 1-27, July-September 2009.

[6] Hanrahan, P. VizQL: A Language for Query, Analysis and Visualization. SIGMOD 2006, June 27-29, 2006, Chicago, Illinois, USA.

[7] Petrini, M., Freitas, M. T., e Pozzebon, M. (2006). Inteligência de negócios ou inteligência competitiva: noivo neurótico, noiva nervosa. Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Administração (EnANPAD).

[8] Alves, N., Lencastre, M., Lins, R. Improving Requirements Quality in Digital Libraries: The case of Scientific Proceedings. In Proceedings of the Quatic 2012-8th International Conference on the Quality of Information and Communucations Technology. Lisbon, Portugal, September 3-6, 2012.