UNIVERSIDADE FEDERAL DO PARANÁ

DEPARTAMENTO DE CIÊNCIA E GESTÃO DA INFORMAÇÃO

MAYARA LETÍCIA FABIANO SILVA

# MINERAÇÃO DE DADOS EM REDES SOCIAIS: UM ESTUDO DE CASO NA FERRAMENTA TWITTER SOBRE A ORGANIZAÇÃO LITTLE, BROWN BOOK COMPANY

CURITIBA
2011

MAYARA LETÍCIA FABIANO SILVA

# MINERAÇÃO DE DADOS EM REDES SOCIAIS:
# UM ESTUDO DE CASO NA FERRAMENTA TWITTER
# SOBRE A ORGANIZAÇÃO LITTLE, BROWN BOOK COMPANY

Monografia apresentada à disciplina de Pesquisa em Informação, como requisito parcial à conclusão do curso de Gestão da Informação, do Departamento de Ciência e Gestão da Informação do setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná.

Orientadora: Profª Drª Denise Fukumi Tsunoda

CURITIBA
2011

“O único lugar onde sucesso vem antes do trabalho é no dicionário.”

Albert Einstein

## AGRADECIMENTOS

A Deus, pela força espiritual para a realização desse trabalho e de toda a minha jornada dentro da UFPR**.**

Aos meus pais, pelo eterno orgulho de nossa caminhada, pelo apoio, compreensão, ajuda, e em especial, por todo carinho ao longo deste percurso.

Aos meus irmãos, pelo carinho, compreensão pelas ausências em decorrência dessa jornada e pela grande ajuda sempre.

Às minhas amigas e colegas de curso Ana Keli Fonseca, Andressa Hudzinski, Bruna Regina Pellizzari, Ruth G. de Lima e Tania Carina de Melo, pela cumplicidade, ajuda e amizade.

À professora Denise Fukumi Tsunoda, pela orientação deste trabalho de forma paciente e sempre me incentivando a fazer o meu melhor.

A todos que de alguma forma me ajudaram nesse trabalho, em especial ao Frank Coelho de Alcantara que com dedicação e empenho me ajudou a estudar e a entender as ferramentas fundamentais para que esse trabalho pudesse ser realizado.

À Universidade Federal do Paraná e aos professores, pelo ensino de qualidade e pela oportunidade de crescimento pessoal, acadêmico e profissional.

Agradeço aos demais amigos que souberam entender minhas ausências e me apoiaram sempre.

## RESUMO

Este trabalho tem por objetivo avaliar os usuários da rede social Twitter que tenham mencionado algum termo relacionado à organização Little,Brown Book Group por meio do método de mineração de dados sumarização, para descobrir a opinião desses usuários sobre a organização. Busca-se também verificar se esse método de mineração de dados é válido para as organizações utilizarem como meio de pesquisa de satisfação dos clientes, novos ramos de atuação e até novos clientes. Utiliza-se um ambiente de pesquisa de dados disponíveis na web, por meio da linguagem de programação R. Usa-se também o programa Revolutions R para o processamento desse código. Para que os objetivos possam ser alcançados utiliza-se a metodologia descritiva com abordagem qualitativa e como técnica, aplica-se um estudo de caso. A partir dos testes realizados foi obtido como resultado uma imagem positiva da organização no ambiente virtual do Twitter. Além disso, pode-se perceber que o método de mineração de dados web é válido para as organizações encontrarem dados que podem se tornar fontes de informações uteis para um estudo de mercado, pesquisa para um novo produto ou até mesmo um novo ramo de atuação.



Palavras-chaves: Mineração de dados. Redes sociais. Twitter. Linguagem R. Web 2.0. Mineração de dados web. Sumarização.

## LISTA DE ILUSTRAÇÕES

**SUMÁRIO**

# 1 INTRODUÇÃO

A sociedade vem mudando a cada geração, trazendo novas características e conceitos para a população. Vive-se a chamada “sociedade da informação” que pode ser definida como:

> A Sociedade da informação está baseada nas tecnologias de informação e comunicação que envolve a aquisição, o armazenamento, o processamento e a distribuição da informação por meios eletrônicos, como o rádio, a televisão, telefone e computadores, entre outros. Estas tecnologias não transformam a sociedade por si só, mas são utilizadas pelas pessoas em seus contextos sociais, econômicos e políticos, criando uma nova comunidade local e global: a Sociedade da Informação. (GOUVEIA, 2004, p.?).

A partir dessa afirmação pode-se perceber que, as organizações estão inseridas na sociedade da informação, e também seguem a evolução das tecnologias, a velocidade e a quantidade de dados e informações que estão disponíveis a todo o momento.

Um novo tipo de fonte de informação que vem ganhando mais espaço dentro e fora das organizações são as mídias sociais. Nelas estão os consumidores finais dos produtos das empresas, futuros funcionários e novos possíveis ramos de atuação.

Porém a quantidade de dados disponível é imensa. Muitos desses dados não são úteis para as organizações e, para que esses dados se tornem informações relevantes para as empresas, é necessário realizar um estudo aprofundado destes para a construção de um conhecimento que agregue valor estratégico à empresa.

Uma atividade relevante para as organizações conseguirem as informações importantes é a mineração de dados. Esta agrega técnicas e métodos para que as empresas possam extrair informação e conhecimento das bases de dados. Por meio de diversos métodos pode-se conseguir padrões entre as informações da base de dados estudada e aplicar o resultado do estudo para alcançar novos consumidores, desenvolver produtos e serviços, dentre outras ações.

Essa pesquisa pretende aplicar o método de mineração de dados sumarização em um perfil de uma organização em uma rede social para, posteriormente, analisar o resultado final e concluir o que essa mineração de dados pode agregar para o mundo dos negócios.

## 1.1 PROBLEMÁTICA

Diante de muitos acontecimentos históricos, tais como guerras mundiais e avanços da tecnologia, as gerações de pessoas foram mudando, se adequando ao seu tempo e buscando melhorar o que se acreditava estar errado. Hoje, com a chamada geração Y, a tecnologia influencia o cotidiano de todos seja profissionalmente ou na vida particular. Porém, o mundo dos negócios não é formado apenas por profissionais de gerações que nasceram com a tecnologia e sabem todos os caminhos para dominá-la, na verdade, a maioria dos gerentes e diretores pertence a gerações anteriores e não conseguem acompanhar o ritmo da evolução das ferramentas digitais.

As redes sociais são um bom exemplo de tecnologia que não é bem explorada pelos profissionais. Por evoluir constantemente, ser uma tecnologia de compartilhamento de informações que tem muitas particularidades e por ter um perfil mais compatível com os novos profissionais que estão entrando agora no mercado, elas não vem sendo aproveitadas pelas organizações de forma relevante.

Uma pesquisa sobre Uso e Aplicação de Redes Sociais e Tecnologias Envolvidas, realizada pela youDb em parceria com o Núcleo de Transferência de Tecnologia (NTT) da Universidade Federal do Rio de Janeiro (UFRJ) mostra que poucas empresas estão usando as redes sociais como espaço de relacionamento, como reforça um levantamento com 67 executivos de 48 empresas de ramos diversificados, onde apenas 21% afirmam se relacionar com clientes e consumidores nesses espaços (Mundo do Marketing, 2009).

Outra pesquisa realizada em 2009 pelo Altimer Group e Wetpaint para a revista Business Week com as 100 empresas mais valiosas ao redor do globo mostraram que os empreendimentos que investem em mídias sociais apresentam melhores resultados e receitas finais mais recheadas. Em média, empresas que investiram em mídias sociais cresceram 18% em um ano, enquanto aquelas que investiram pouco nas redes tiveram queda de 6%, em média, em suas receitas no mesmo período (Pequenas Empresas Grandes Negócios, 2011).

A partir dessas pesquisas pode-se perceber que as redes sociais podem vir a representar uma vantagem competitiva. Porém, para que os resultados finais venham a ser satisfatórios e não ocorra perda de dinheiro e de tempo, é importante conhecer a ferramenta mais adequada aos objetivos da organização e qual delas melhor se encaixa aos padrões, cultura e necessidades da mesma.

Essa pesquisa pode ser definida na seguinte questão: É possível utilizar os dados disponíveis nas redes sociais, por meio de um método de mineração de dados, nas organizações como fontes de informação?

Esse estudo pretende analisar os dados dos usuários da rede social Twitter, mais especificamente dos usuários que comentem sobre a organização Little, Brown Book Group, por meio de um método de mineração de dados para buscar um possível padrão da opinião desses usuários sobre a organização ou até mesmo alguns de seus produtos que possa vir a ser útil para a organização estudada para alcançar novos clientes.

## 1.2 JUSTIFICATIVA

O mundo das organizações está cada vez mais competitivo. Quando uma empresa encontra uma solução que pode vir a se tornar uma vantagem competitiva ela deve pesquisar a melhor maneira de implantá-la e monitorar as atividades para que o objetivo final seja alcançado. As redes sociais trazem novas potencialidades em prol das organizações, ainda que estas não estejam totalmente definidas e sedimentadas.

Mesmo as organizações que já aderiram a essa ferramenta no seu cotidiano, muitas vezes, não tem conhecimento de como aproveitar a maioria das funcionalidades e informações que estão disponíveis nas mídias sociais.

A tecnologia das redes sociais faz parte do conceito de compartilhamento de informações na internet e da customização de um espaço para os usuários. A busca de dados nas redes sociais também não possui uma pesquisa aprofundada devido a novidade do tema. Atualmente, com o perfil dos novos profissionais e consumidores evoluindo para uma postura mais informada, crítica e que buscam meios de representar a sua individualidade, o estudo de como utilizar as redes sociais se torna

um tema atual e importante no desenvolvimento da sociedade da informação.

Um gestor da informação deve estar atento as fontes de informação que podem vir a ser úteis no dia a dia das organizações, conhecer as mudanças da sociedade da informação e aprender diversas formas de obter conhecimento. Novos métodos de recuperação das informações também representam novos conhecimentos e fontes de novos estudos.

Para que esse estudo apresente um resultado mais concreto foi escolhida a organização editora de livros Little,Brown Book Group para ter o seu perfil na rede social Twitter analisado por meio de um método de mineração de dados. A empresa foi escolhida pela razão de o mercado consumidor de livros impressos sofrer ameaças de novos concorrentes, como livros digitais, audiobooks, dentre outros. A organização também apresenta, no ramo de editoração de livros, uma das maiores quantidades de usuários “seguidores” que ajudará nas conclusões finais deste estudo.

## 1.3 OBJETIVOS

Como objetivo geral definiu-se: analisar os usuários da rede social Twitter, particularmente os que estão envolvidos com a empresa editora de livros Little, Brown Book Group para encontrar um possível padrão de comportamento ou características entre seus seguidores.

Do objetivo geral foram derivados os específicos:

- desenvolver uma breve descrição da empresa Little, Brown Book Group;
- aplicar o método de mineração de dados sumarização com a linguagem R, por meio do programa Revolutions R;
- analisar esses dados para verificar se esse método de coleta pode ser relevante ou não para as organizações.

## 1.4 ESTRUTURA DO DOCUMENTO

O trabalho está estruturado da seguinte forma: o primeiro capítulo trata da introdução ao projeto. Nele está contido a problemática do estudo, a justificativa, os

objetivos e a descrição da estrutura do projeto para que se compreenda como o problema de pesquisa teve seu início, o interesse da autora pelo tema e o que se pretende alcançar com o projeto.

O segundo capítulo aborda a literatura pertinente ao tema estudado. Serão levantadas na literatura as informações pertinentes a área de pesquisa nos temas: evolução das gerações da sociedade, Web 2.0, Redes Sociais, Twitter, Análise de sentimento, Mineração de Dados e linguagem R. Os temas foram escolhidos para fornecer o suporte necessário para auxiliar na pesquisa.

O terceiro capítulo apresenta a metodologia do trabalho. Nessa etapa estão descritos todos os procedimentos metodológicos que foram seguidos para que a análise final do estudo seja válida. Dentre esses métodos estão o tipo de metodologia de pesquisa utilizada, a descrição da organização e a descrição do método utilizado para realizar a mineração de dados.

O quarto capítulo mostra os resultados obtidos a partir da mineração de dados realizada no programa Revolutions R e a análise dos mesmos. Nessa etapa estão as tabelas com os resultados e a interpretação das mesmas, que foram criados a partir dos dados disponíveis no Twitter. Ainda nessa seção a análise se dividiu em duas etapas sendo que a primeira avalia os resultados obtidos a partir do Twitter e a segunda avalia esse método de mineração, se é válido, útil para as organizações, dentre outras avaliações.

Por fim essa pesquisa se encerra com as considerações finais e as referências utilizadas no referencial teórico para a realização desse trabalho.

# 2 LITERATURA PERTINENTE

A seguir será apresentada uma breve revisão de literatura pertinente para o auxílio no desenvolvimento da pesquisa.

## 2.1 A EVOLUÇÃO DAS GERAÇÕES

Para compreender o grande impacto das redes sociais nas organizações, e também no cotidiano das pessoas, é importante conhecer e entender a evolução das grandes gerações que se seguiram no decorrer da história da sociedade.

### 2.1.1 GERAÇÃO VETERANOS

A primeira geração reconhecida é a geração chamada de veteranos (site Foco em Gerações 2010). Ela é composta por pessoas que nasceram entre 1925 e 1945, no período das grandes crises econômicas e viveram a época da 2ª Guerra Mundial (site Foco em Gerações 2010). De acordo com Eline Kullock, especialista em gerações, os veteranos são pessoas mais rígidas e respeitam regras, por causa das dificuldades que viveram (site Foco em Gerações 2010).

Muitos dos homens que se tornaram empresários no pós-guerra vieram retomar suas posições quando a mesma acabava e voltavam dos campos completamente mudados (site Cine Gestão 2009). A geração dos veteranos é muito organizada, dedicada e prática, prefere que sejam apresentados resultados de mudança e não vem custos para isso (site Cine Gestão 2009).

Os indivíduos da geração dos Veteranos preferem a estabilidade, por isso pessoas dessa geração passam muito tempo na mesma empresa (site Cine Gestão 2009). Os funcionários dessa geração que não se aposentaram ainda, provavelmente estão em cargos de chefia ou de sociedade, tamanha a confiança devotada à empresa (site Cine Gestão 2009). Há ainda aqueles que nunca foram promovidos, ou então, foram promovidos a cargos significativamente baixos (site Cine Gestão 2009).

Essa geração é a geração dos sacrifícios (site Foco em Gerações 2010). Gostam de hierarquias rígidas e de padrões a serem seguidos (site Foco em Gerações 2010). Eles tiveram que aprender todo um sistema novo de automação vindo da II Revolução Industrial e das necessidades das guerras (site Cine Gestão 2009). Ainda tiveram que se adaptar às tecnologias futuras, tendo que aprender a lidar com computadores, celulares, automóveis modernos, dentre outras tecnologias (site Cine Gestão 2009).

Em uma cultura organizacional atual, o impacto dessa geração ainda é significativo (para o caso das empresas que ainda tem funcionários muito antigos) (site Foco em Gerações 2010). Eles acumularam sabedoria e experiência ao longo dos anos, sem abrir mão da moral e dos costumes que aprenderam na juventude, quando ainda estavam aprendendo sobre a vida (site Foco em Gerações 2010). A rigidez de alguns criou uma geração um tanto quanto libertária e otimista, os chamados baby boomers (site Cine Gestão 2009).

### 2.1.2 GERAÇÃO BABY BOOMERS

Nos livros de história, o fenômeno conhecido como “Baby Boom” aconteceu a partir de 1945, com o fim da Segunda Guerra (site Foco em Gerações 2010). Soldados sobreviventes voltaram vitoriosos aos Estados Unidos e cheios de esperança para formar ou continuar sua família (site Como Tudo Funciona, 2010). Os nascidos nessa geração mudaram a realidade, a história e a cultura organizacional das empresas (site Foco em Gerações 2010).

Em 1960 eles começaram uma revolução cultural diferente do que já havia sido visto antes (site Como Tudo Funciona, 2010). Liberdade de expressão, liberdade sexual, letras de música mais agressivas, abertura e anistia política. Muitas ações e inovações vieram dessa geração (site Como Tudo Funciona, 2010).

Mesmo fazendo parte de uma mesma geração, eles são constantemente divididos em duas categorias: aqueles que nasceram entre 1946 e 1954 (geralmente chamados de Primeiros Boomers) e os nascidos entre 1955 e 1964 (freqüentemente chamados de Boomers Posteriores ou Geração Jones) (site Como Tudo Funciona, 2010).

Os primeiros foram aqueles que presenciaram os assassinatos de líderes importantes como John Kennedy e Martin Luther King. Os outros viram a queda do presidente Nixon no escândalo do Watergate (site Como Tudo Funciona, 2010). Dois momentos culturais distintos para a mesma geração. Porém dois aspectos importantes unem os dois grupos de baby boomers (site Foco em Gerações 2010). O primeiro é a Guerra do Vietnã. Os Primeiros Boomers foram à guerra como combatentes e os Posteriores viram tudo de longe pela televisão. A televisão é o segundo aspecto que os une. Os baby boomers foram a primeira geração a crescer com um novo formato de difusão de informação (site Como Tudo Funciona, 2010).

Essa identidade foi de muitas formas, bastante cética (site Como Tudo Funciona, 2010). Aos vinte anos, os Boomers cunharam a famosa frase "Não confie em ninguém com mais de 30 anos" em plena Guerra do Vietnã (site Foco em Gerações 2010). Os Babyboomers passaram a confiar em si mesmos. Eles foram chamados de "Geração Eu" porque foi a primeira geração a fazer um intervalo entre a infância e a idade adulta, e a explorar o fato de ser jovem (site Foco em Gerações 2010). Eles se casaram e tiveram filhos mais tarde, e gastaram bastante com si mesmos (site Foco em Gerações 2010).

De modo controverso, eles também são uma das gerações mais ativas e menos egoístas de todos os tempos (site Foco em Gerações 2010). Sua luta contínua contra a injustiça criou o movimento das mulheres, o movimento pelos direitos civis, os protestos contra a Guerra do Vietnã dentre outros (site Foco em Gerações 2010).

Os babyboomers geralmente ocupam os cargos de chefia, depois do longo tempo que trabalharam na mesma empresa (site Como Tudo Funciona, 2010). São mais tranqüilos e racionais mas, por terem tido uma grande afirmação ideológica no passado, são pouco receptivos à mudanças (site Foco em Gerações 2010).

### 2.1.3 GERAÇÃO X

Pertencem à Geração X aqueles que nasceram entre 1965 e 1980 (site Foco em Gerações 2010). Nesse período, as condições materiais do planeta permitem pensar em qualidade de vida, liberdade no trabalho e nas relações (SERRANO, 2010). Com o desenvolvimento das tecnologias de comunicação, a Geração X já

pode tentar equilibrar vida pessoal e trabalho (SERRANO 2010). Mas, como enfrentaram crises violentas, como a do desemprego na década de 80, também se tornaram céticos e superprotetores (site Foco em Gerações 2010).

A Geração recebeu esse nome devido a um romance sobre o assunto. Seus membros são os filhos de mães que trabalham fora ou de pais divorciados (SERRANO 2010). Diferentemente dos integrantes da geração do baby boom, que tendiam a se especializar em ciências humanas, os membros desse grupo preferiram as áreas de administração e economia, trocando o idealismo por um realismo mais pragmático e cético (site Foco em Gerações 2010). Entre suas principais influências estão Ronald Reagan, a explosão do ônibus espacial Challenger e a Guerra do Golfo (site Foco em Gerações 2010).

Os indivíduos da Geração X, apesar lidar melhor com mudanças, são também os que mais se conformam com uma determinada situação (SERRANO 2010). Eles viveram em pleno período de Guerra Fria, onde uma nova ordem mundial estava prestes a ser implantada, mas não era implantada nunca (site Foco em Gerações 2010). A briga entre capitalismo e socialismo fez nascer uma sensação de insegurança e conformismo muito grande na Geração X (SERRANO 2010).

A maior parte daqueles que fazem parte dessa geração não se sente ameaçada pela vida corporativa (site Foco em Gerações 2010). Eles se motivam pelas mesmas razões que a geração dos boomers se motivou. Planejam deixar a vida corporativa em breve para iniciar algum empreendimento ou trabalhar em empresas pequenas, opções que se encaixam melhor, para eles, que os papéis corporativos que necessitarão assumir (site Foco em Gerações 2010).

Entretanto, esse perfil ainda é altamente necessário às grandes corporações (SERRANO 2010). São pessoas que entraram na empresa com a mente aberta, um pouco mais racional do que a geração anterior e que agregou valores importantes no seu desenvolvimento (SERRANO 2010).

Os nascidos nessa geração buscam um equilíbrio real entre trabalho e vida pessoal e são profundamente independentes (SERRANO 2010). A Geração X é a primeira geração que verdadeiramente domina os computadores (SERRANO 2010). É claro que os "Y" estão muito mais atualizados com as novidades tecnologicas, mas um "X" convive com informática há mais tempo (SERRANO 2010). E o mais importante: são os indivíduos da Geração X os maiores incentivadores do trabalho em equipe (SERRANO 2010).

### 2.1.4 GERAÇÃO Y

A geração Y é chamada também de geração Internet. O crescimento da geração Y foi marcada por momentos na história como o julgamento de O.J.Simpson, o massacre na escola de Columbine, o vazamento do petroleiro Exxon Valdez, a Guerra do Golfo, 11 de setembro, a Guerra da Iraque dentre outros (site Foco em Gerações 2010) Porém, a mudança mais significativa para essa geração foi a ascensão do computador, da Internet e de outras tecnologias digitais (TAPSCOTT, 2010). Dessa forma as crianças nascidas nessa época cresceram com essas tecnologias enquanto os adultos tiveram que se adaptar a ela, tornando a tecnologia indispensável para o cotidiano moderno (TAPSCOTT, 2010).

Nessa época é preciso buscar as informações em vez de somente observá-las. Isso obriga os jovens a desenvolver o raciocínio e habilidades investigativas (TAPSCOTT, 2010). Além de também ter de desenvolver um censo crítico, como por exemplo, que sites merecem credibilidade, o que é real e o que é fictício, dentre outros (TAPSCOTT, 2010).

Os profissionais da geração Internet estão muito mais preocupados com o próprio crescimento do que com o crescimento da empresa (site Foco em Gerações 2010). Mas isso não quer dizer falta de comprometimento. Quando eles decidem pelo emprego certo, se eles observam a possibilidade de crescimento, se dedicam 100% ao local de trabalho e ficam felizes com o que fazem. Quando há um retorno positivo, cada vez mais eles desempenham melhor suas funções (TAPSCOTT, 2010).

O problema está quando nada disso compensa. Se o "Y" não se satisfaz no trabalho ele pode pensar de duas formas: 1-"Não posso sair agora, mas só vou aguentar o tempo necessário, até conseguir coisa melhor" ou 2-"Sou novo e tenho condições de conseguir algo melhor. Não vou ficar mais aqui" (TAPSCOTT, 2010). De uma forma ou de outra, para eles insatisfação é sinônimo de demissão. É o que especialistas no assunto tendem a chamar de infidelidade (TAPSCOTT, 2010).

Essa geração é aquela que não consegue ficar muito tempo no mesmo lugar; E às vezes não é nem questão de compromisso ou dinheiro. Eles simplesmente ficam entediados, se cansam da rotina e precisam de algo novo constantemente (TAPSCOTT, 2010). Os "Y" cresceram em um mundo com aceleradores de tempo,

como a televisão, o celular, a internet, os videogames, ente outros. Eles acompanharam a evolução tecnológica - muitos inclusive ajudaram a criá-las, vide Mark Zuckerman, o criador do Facebook (site Foco em Gerações 2010). Eles não gostam de ficar parados no tempo. Precisam ficar atentos às novidades tecnologicaes e em tudo o que está acontecendo e são os que mais acessam internet e dispositivos de mídia como Orkut, Facebook e Twitter (site Foco em Gerações 2010).

A geração Y gosta de ditar as regras, e não de segui-las. O que não significa insubordinação. Pesquisas nas empresas comprovam que a maioria das ideias inovadoras que surgem nas empresas vem de indivíduos da Geração Y (TAPSCOTT, 2010). A Geração Internet quer satisfação pessoal antes da satisfação financeira, mas faz o possível para aliar os dois (TAPSCOTT, 2010).

Para os “Y”, faz muito mais sentido agregar experiência em diversas áreas e empresas do que fazer carreira num lugar só. Uma exigência que o mercado começou a fazer, a do funcionário multiuso, agora está se voltando contra ele: cada vez mais profissionais querem ser multiuso, desde que sejam reconhecidos por isso (TAPSCOTT, 2010). Até mesmo porque, a Geração Y está muito mais voltada para o conteúdo de qualidade que tem dentro deles.

Essa geração irá comandar as empresas no futuro. Eles serão muito mais experientes em tecnologia, em trabalho e em informação (TAPSCOTT, 2010).

### 2.1.5 GERAÇÃO Z

Ao contrário do que possa parecer, a Geração Z não é formada pelos filhos da Geração Y (site Foco em Gerações 2010). A letra Z indica uma geração de indivíduos preocupados, cada vez mais com a conectividade com os demais indivíduos de forma permanente (site Foco em Gerações 2010).

Assim, se as gerações anteriores se conectavam com o seu mundo através de um computador de mesa, a nova geração passou a ficar constantemente disponível e conectada através de dispositivos móveis (SERRANO 2010).

A noção de grupo passa a ser virtual. Cada pessoa passa a ter o seu vídeo game, a sua TV, o seu celular e o seu equipamento de som. Isto muda a forma de

comportamento e relacionamento social sobremaneira, já que até então, essas formas de diversão, entretenimento ou comunicação eram coletivas(SERRANO, 2010).

Ao final do Século XX, a televisão ocupava um lugar central na sala, reunindo a família no que se chamava “horário nobre”. Da mesma forma no início do Século passado, o Rádio e equipamentos de som ocupavam esse lugar. A geração Z dispõe de todos esses dispositivos em equipamentos portáteis que não os prendem mais a lugar nenhum (site Foco em Gerações 2010).

Os indivíduos da geração Z, normalmente são datados como nascidos ao final do Século XX. Mas, os gerados no início do Século XXI, independente de outras denominações que possam ser dadas, mantêm as características da geração Z. (alguns estudiosos já estão chamando os nascidos a partir de 2010 de Geração Alfa) (SERRANO, 2010).

Assim, pessoas da geração Z acabam trazendo traços de comportamento das gerações anteriores, aliado a uma forte Responsabilidade Social e preocupação com o meio ambiente e a sustentabilidade do planeta (SERRANO, 2010). Já foi dito que a geração Z se parece mais com a geração Y do que os próprios indivíduos da Geração Y (SERRANO 2010).

## 2.2 WEB 2.0

A partir do surgimento da Internet muitas mudanças ocorreram na forma da sociedade obter informações e na comunicação das pessoas. A "Explosão Informacional” exigiu o desenvolvimento de tecnologias que permitissem o armazenamento e a recuperação de maneira satisfatória dos conteúdos, uma vez que os recursos disponíveis antigamente não seriam capazes de atender o volume de informação.

Como conseqüência dessa necessidade, a Internet evoluiu em seus serviços oferecidos, criando características virtuais aos modelos tradicionais de comércio, reuniões, testes, disciplinas, dentre outros. Como processo natural, serviços que jamais poderiam ser disponibilizados pelo meio tradicionais, podem agora serem oferecidos, pautados nas premissas de agilidade, praticidade e interação. E nestes três conceitos é que está pautada a Web 2.0.

O'Reily (2006), um dos precursores deste novo conceito de internet, afirma em seu artigo que o despertar para a Web 2.0 surgiu a partir de um brainstorming dele com a MediaLive International. No momento, foi despertada a necessidade de discutir novas soluções para a Rede, surgindo a primeira conferência denominada Web 2.0. Como um dos resultados desse evento, foi proposto o conceito de "Internet como uma plataforma".

## 2.3 REDES SOCIAIS

Falar em redes significa trabalhar com concepções variadas nas quais parecem misturar-se idéias baseadas no senso comum, na experiência cotidiana do mundo globalizado ou ainda em determinado referencial teórico-conceitual. Loiola & Moura (1997, p.54) ressaltam que: "A presença de um ponto central, de uma fonte geradora/ propulsora, não figura no significado popular de rede. A igualdade e a complementaridade entre as partes são seus aspectos básicos, reforçados pela regularidade entre as malhas".

Esse é apenas um dos muitos conceitos de redes. No contexto de compartilhamento de conteúdos as redes de conexões já são usadas a algum tempo, porém as organizações as estão visualizando como ferramenta organizacional a pouco tempo.

> O que é novo no trabalho em redes de conexões é sua promessa como uma forma global de organização com raízes na participação individual. Uma forma que reconhece a independência enquanto apóia a independência. O trabalho em redes de conexões pode conduzir a uma perspectiva global baseada na experiência pessoal. (LIPNACK e STAMPS, 1992, p. 19).

Apesar de essa afirmação já ser antiga, a sua ideia principal ainda é muito atual.

Mídias Sociais são definidas como tecnologias e práticas on-line usadas por pessoas ou empresas para disseminar conteúdo, provocando o compartilhamento de opiniões, ideias, experiências e perspectivas. Seus diversos formatos, atualmente podem englobar textos, imagens, áudio e vídeo, conceito de Wagner Fontura no seu site Boombust.

Um elemento importante no estudo de redes sociais é entender em que elas se diferenciam das mídias sociais. Mídias sociais foram definidas por Boyd e Ellison (2007) como sistemas que permitem I) a construção de uma *persona* por meio de um perfil ou página pessoal; II) a interação por meio de comentários; e III) a exposição pública da rede social de cada ator. Os sites dessas redes seriam uma categoria do grupo de softwares sociais com aplicação direta para a comunicação medida por computador.

Essa definição ganha significado ímpar quando utilizada nas empresas, que a veem como meio de comunicação com seus públicos, pois esperam que o conteúdo compartilhado ultrapasse a página eletrônica, gerando ação de compra de serviço ou produto ou o simples ato de expressar sua opinião sobre a organização. E isso, vai além de comercializar produtos antigos de maneiras novas (DIZARD, 2000, p.257), é necessário que haja uma mobilização e conscientização dos recursos organizacionais, isto é, objetivo e estratégia empresarial.

Uma característica importante para as redes sociais é a de ser democrática. Para Lemos e Palácios (2001, p.238) a questão do público e privado na Internet ganha significado único, pois o sentido de privado é aquele de natureza íntima, pessoal não sendo necessariamente restrito a determinado grupo de pessoas e o público ganha o sentido da publicidade, tendo conhecimento público por qualquer meio ou material.

Outro conceito que deve ser levado em consideração diz respeito à cultura empresarial, visto que a implantação de mídias sociais deve estar em harmonia com os valores e crenças da empresas porque o que é compartilhado na rede deve ser coerente com o que a organização acredita.

### 2.3.1 TWITTER

O modelo de comunicação todos para todos da web (Lévy, 1999) traz a possibilidade de que, ao menos em tese, qualquer pessoa possa produzir e publicar conteúdo na rede. Com o advento da Web 2.0 (O'Reilly 2005), passam a surgir novos espaços de participação que facilitam esse processo de produção e publicação de conteúdos, como blogs, sites de redes sociais e *wikis*.

O Twitter é uma rede social e servidor para microblogging, que permite aos usuários enviar e receber atualizações pessoais de outros contatos (em textos de até 140 caracteres, conhecidos como "*tweets*"), por meio do *website* do serviço, por SMS e por softwares específicos de gerenciamento.

O Twitter é estruturado com seguidores e pessoas a seguir, onde cada usuário pode escolher quem deseja seguir e ser seguido por outros. Há também a possibilidade de enviar mensagens em modo privado para outros usuários. A janela particular de cada usuário contém, assim, todas as mensagens públicas emitidas por aqueles indivíduos que o usuário segue. Mensagens direcionadas também são possíveis a partir do uso da "@'' antes do nome do destinatário. Cada página particular pode ser personalizada pelo twitter por meio de uma construção de um pequeno perfil.

As atualizações são exibidas no perfil de um usuário em tempo real e também enviadas a outros usuários seguidores que tenham assinado para recebê-las. As atualizações de um perfil ocorrem por meio do site do Twitter, por RSS, por SMS ou programa especializado para gerenciamento. O serviço é gratuito pela internet, entretanto, usando o recurso de SMS pode ocorrer a cobrança pela operadora telefônica.

O Twitter foi fundado por Jack Dorsey, Biz Stone, e Evan Willians ainda em 2006, como projeto da empresa Odeo. Uma das características mais importantes do sistema é permitir que sua API seja utilizada para a construção de ferramentas que utilizem o Twitter. Isso fez da ferramenta popular, sendo utilizada em inúmeras iniciativas, como o Summize, ferramenta de busca no sistema que posteriormente foi adquirida pelo Twitter e tornou-se sua busca "oficial".

## 2.4 ANÁLISE DE SENTIMENTO

A Análise de sentimento ou mineração de opinião é o estudo computacional de opiniões das pessoas, avaliações, e as emoções em direção a entidades, eventos e seus atributos.(LIU, 2010).

O estudo da análise de sentimentos apresenta uma justificativa de que opiniões são importantes porque não importa a decisão que se tome, é sempre importante ouvir a opinião de outras pessoas.(LIU,2010).

A pesquisa nessa área começou com a classificação de sentimento e subjetividade. Ela tratou o problema como sendo um problema de classificação de texto. A Classificação sentimentos é um documento opinativo (por exemplo, análises de produtos) ou sentença expressa uma opinião positiva ou negativa.(Lee e Pang 2008).

Lee e Pang também definiram a análise de sentimento de outra maneira.Eles interpretam esse tipo de análise como uma pesquisa que classifica se um documento opinativo (por exemplo, revisões de produto) ou sentença expressa uma opinião positiva ou negativa.

Esse tipo de análise também é interpretada como subjetiva. Wiebe, Wilson, Bruce, Bell e Martin definem essa classificação da subjetividade como determinante se uma frase é subjetiva ou objetiva.Porém Liu e Jindal chamam atenção para as aplicações na vida real que requerem uma análise mais detalhada, pois o usuário muitas vezes quer saber sobre o que as opiniões foram expressas.

## 2.5 MINERAÇÃO DE DADOS

A rápida evolução dos recursos computacionais ocorrida nos últimos anos permitiu que, simultaneamente, fossem gerados grandes volumes de dados. Estima-se que, em 1999, a quantidade de informação no mundo dobra a cada 20 meses e que o tamanho e a quantidade dos bancos de dados crescem com velocidade ainda maior (DILLY, 1999).

O conceito de Mineração de Dados (*Data Mining*) está se tornando cada vez mais popular como uma ferramenta de descoberta de informações, que podem

revelar estruturas de conhecimento, que possam guiar decisões em condições de certeza limitada.

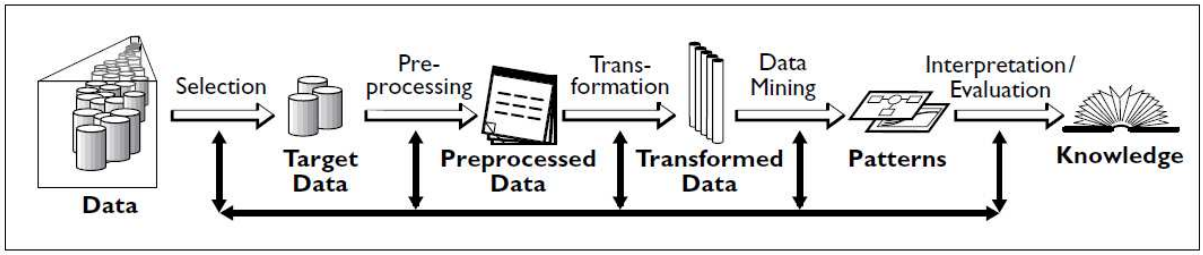
A Mineração de Dados é uma tecnologia que possui conceitos de três áreas: estatística clássica, inteligência artificial e aprendizado de máquina, sendo a primeira a mais antiga delas. Observa-se que a Mineração de Dados é parte de um processo maior conhecido como KDD (*Knowledge Discovery in Databases*) – em português, Descoberta de Conhecimento em Bases de Dados –, que, segundo Addrians & Zantinge (1996), permite a extração não trivial de conhecimento previamente desconhecido e potencialmente útil de um banco de dados.

Esse conceito é enfatizado por Fayyad *et al.* (1996b), ao afirmar que é "o processo não trivial de identificação de padrões válidos, desconhecidos, potencialmente úteis e, no final das contas, compreensíveis em dados".

O processo KDD é constituído de várias etapas, que são executadas de forma interativa e iterativa.

> As etapas são interativas porque envolvem a cooperação da pessoa responsável pela análise de dados, cujo conhecimento sobre o domínio orientará a execução do processo. Por sua vez, a iteração deve-se ao fato de que, com freqüência, esse processo não é executado de forma seqüencial, mas envolve repetidas seleções de parâmetros e conjunto de dados, aplicações das técnicas de *Data Mining* e posterior análise dos resultados obtidos, a fim de refinar os conhecimentos extraídos. (BRACHMAN & ANAND, 1996, p. ?).

**Figura 1 Visão geral das etapas que constituem o processo de KDD**




Fonte: **Advances in Knowledge Discovery in Data Mining.** FAYYAD *et al*. 1996

O processo KDD é descrito na Figura 1e foi feita por Fayyad em 1996. O processo de KDD é interativo e iterativo (com muitas decisões feitas pelo usuário), envolvendo algumas etapas, resumidas como:

1. Aprendizagem do domínio da aplicação: inclui um conhecimento prévio dos dados a serem analisados e os objetivos da aplicação (FAYYAD 1996).

2. Criar um conjunto de dados: inclui selecionar um conjunto de dados, a partir da base de dados iniciais, em que a descoberta de padrões deve realizada. Dentre as várias etapas do processo KDD, a principal, que forma o núcleo do processo e que, muitas vezes, confunde-se com ele, chama-se *Data Mining* (FAYYAD 1996).

3. Limpeza de dados e pré-processamento: inclui operações básicas, tais como remover os “ruídos” ou outliers (dados incorretos ou que diferem muito do grande grupo) se apropriados; Coletar informações para criar um modelo da amostra de dados ou até mesmo para montar uma amostra para os dados considerados “ruído”; Decidir sobre estratégias para lidar com falta de dados e a contabilidade para informações de tempo e seqüência de alterações conhecidas, decidir o tipo de dados, esquema e mapeamento de valores ausentes e desconhecidos (FAYYAD1996).

4. Redução de dados e projeção: inclui encontrar recursos úteis para representar os dados, dependendo sobre o objetivo da tarefa, e utilizando a redução de dimensionalidade ou métodos de transformação para reduzir o número efetivo de variáveis sob consideração ou para encontrar representações invariantes para os dados (FAYYAD 1996).

5. Escolher a função de mineração de dados: inclui decidir o propósito do modelo derivado pelo algoritmo de mineração de dados (FAYYAD 1996).

6. Escolher o algoritmo de mineração de dados: inclui método de seleção a ser utilizado para a busca de padrões nos dados, tais como decidir quais modelos e os parâmetros podem ser mais adequados (por exemplo, modelos para dados categóricos são diferentes dos modelos no vetores sobre reais) e combinando um método de mineração especial com os critérios gerais do processo KDD (por exemplo, o usuário pode estar mais interessados em compreender o modelo que na sua capacidade de previsão) (FAYYAD 1996).

7. Data mining: inclui a procura de padrões de interesse em uma forma particular de representação ou um conjunto de tais representações, incluindo a classificação de regras ou árvores, regressão, clusterização, modelagem seqüência, análise de dependência, linha de análise (FAYYAD 1996).

8. Interpretação: inclui interpretar a descoberto dos padrões e possivelmente voltar a qualquer um dos passos anteriores, bem como visualização possível do padrões extraídos, eliminando redundantes ou irrelevantes padrões, e traduzindo os

úteis em termos compreensíveis para os usuários (FAYYAD 1996).

9. Usar o conhecimento descoberto: inclui incorporação deste conhecimento para o desempenho do sistema, tomar ações baseadas no conhecimento, ou simplesmente documentá-lo e denunciá-lo aos interessados, bem como a verificação desses padrões, ou resolver potenciais conflitos com que se acreditava anteriormente (FAYYAD 1996).

O KDD tem início com o entendimento do domínio da aplicação e dos objetivos a serem atingidos.

Em seguida, é realizado um agrupamento organizado da massa de dados alvo da descoberta. Como em toda análise quantitativa, a qualidade dos dados é essencial para a obtenção de resultados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso do *Data Mining*, como afirmam Diniz & Louzada-Neto (2000). A limpeza dos dados (identificada na literatura como *Data Cleaning*) é realizada por meio de um pré-processamento, visando assegurar a qualidade dos dados selecionados. Destaca-se que, segundo Mannila (1996), essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido às dificuldades de integração de bases de dados heterogêneas.

Os dados pré-processados devem passar por outra transformação, que os armazena adequadamente, visando facilitar o uso das técnicas de *Data Mining*. Nessa fase, o uso de *Data Warehouses* expande-se consideravelmente, já que, nessas estruturas, as informações estão alocadas da maneira mais eficiente. Addrians & Zantinge (1996) definem *Data Warehouse* como um depósito central de dados, extraído de dados operacionais, em que a informação é orientada a assuntos, não volátil e de natureza histórica.

Devido a essas características, *Data Warehouses* tendem a se tornar grandes repositórios de dados extremamente organizados, facilitando a aplicação do *Data Mining* (FAYYAD 1996).

Prosseguindo no processo KDD, chega-se especificamente à fase de *Data Mining*. O objetivo principal desse passo é a aplicação de técnicas de mineração nos dados pré-processados, o que envolve ajuste de modelos e/ou determinação de características nos dados. Em outras palavras, exige o uso de métodos inteligentes para a extração de padrões ou conhecimentos dos dados (FAYYAD 1996).

É importante destacar que cada técnica de *Data Mining* utilizada para conduzir as operações de Mineração de Dados adapta-se melhor a alguns

problemas do que a outros, o que impossibilita a existência de um método de *Data Mining* universalmente melhor. Para cada problema particular, tem-se uma técnica particular. Portanto, o sucesso de uma tarefa de *Data Mining* está diretamente ligado à experiência e à intuição do analista (FAYYAD 1996).

A etapa final do processo de mineração consiste no pós-processamento, que engloba a interpretação dos padrões descobertos e a possibilidade de retorno a qualquer um dos passos anteriores. Assim, a informação extraída é analisada (ou interpretada) em relação ao objetivo proposto, sendo identificadas e apresentadas as melhores informações (FAYYAD 1996).

Dessa forma, o propósito do resultado não consiste somente em visualizar, gráfica ou logicamente, o rendimento da pesquisa realizada por meio do *Data Mining*, mas, também, em filtrar a informação que será apresentada, eliminando possíveis ruídos (ou seja, padrões redundantes ou irrelevantes) que podem surgir no processo (BRACHNAD e ANAND 1996).

O processo de mineração de dados em si, pode ser entendido como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomada de decisões. É uma metodologia aplicada em diversas áreas que usam o conhecimento, como empresas, indústrias e instituições de pesquisa. *Data Mining* define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para descrever características do passado como para predizer tendências para o futuro (BRACHNAD e ANAND 1996).

Para encontrar respostas ou extrair conhecimento interessante, existem diversos métodos de *Data Mining* disponíveis na literatura. Mas, para que a descoberta de conhecimentos seja relevante, é importante estabelecer metas bem definidas (FAYYAD 1996).

Essas metas são alcançadas por meio dos seguintes métodos de *Data Mining*: Classificação, Modelos de Relacionamento entre Variáveis, Análise de Agrupamento, Sumarização, Modelo de Dependência, Regras de Associação e Análise de Séries Temporais, conforme citação e definição feita por Fayyad *et al.* (1996a).

É importante ressaltar que a maioria desses métodos é baseada em técnicas das áreas de aprendizado de máquina, reconhecimento de padrões e estatística. Essas técnicas vão desde as tradicionais da estatística multivariada, como análise

de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos (FAYYAD 1996).

Os métodos tradicionais de *Data Mining,* definidos por Faayad em 1996, são:

a) *Classificação*: associa ou classifica um item a uma ou várias classes categóricas pré-definidas. Uma técnica estatística apropriada para classificação é a análise discriminante. Os objetivos dessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de várias populações, além da classificação das observações em uma ou mais classes predeterminadas. A idéia é derivar uma regra que possa ser usada para classificar, de forma otimizada, uma nova observação a uma classe já rotulada. Segundo Mattar (1998), a análise discriminante permite que dois ou mais grupos possam ser comparados, com o objetivo de determinar se diferem uns dos outros e, também, a natureza da diferença, de forma que, com base em um conjunto de variáveis independentes, seja possível classificar indivíduos ou objetos em duas ou mais categorias mutuamente exclusivas.
b) *Modelos de Relacionamento entre Variáveis*: associa um item a uma ou mais variáveis de predição de valores reais, consideradas variáveis independentes ou exploratórias. Técnicas estatísticas como regressão linear simples, múltipla e modelos lineares por transformação são utilizadas para verificar o relacionamento funcional que, eventualmente, possa existir entre duas variáveis quantitativas, ou seja, constatar se há uma relação funcional entre X e Y. Observa-se, conforme Gujarati (2000), que o método dos mínimos quadrados ordinários, atribuído a Carl Friedrich Gauss, tem propriedades estatísticas relevantes e apropriadas, que tornaram tal procedimento um dos mais poderosos e populares métodos de análise de regressão.
c) *Análise de Agrupamento* (*Cluster*): associa um item a uma ou várias classes categóricas (ou *cluster*s), em que as classes são determinadas pelos dados, diversamente da classificação em que as classes são pré-definidas.Os *cluster*s são definidos por meio do agrupamento de dados baseados em medidas de similaridade ou modelos probabilísticos. A análise de *cluster* (ou agrupamento) é uma técnica que visa detectar a

existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso de sua existência, determinar quais são eles. Nesse tipo de análise, segundo Pereira (1999), o procedimento inicia com o cálculo das distâncias entre os objetos estudados dentro do espaço multiplano constituído por eixos de todas as medidas realizadas (variáveis), sendo, a seguir, os objetos agrupados conforme a proximidade entre eles. Na seqüência, efetuam-se os agrupamentos por proximidade geométrica, o que permite o reconhecimento dos passos de agrupamento para a correta identificação de grupos dentro do universo dos objetos estudados.

d) *Sumarização*: determina uma descrição compacta para um dado subconjunto. As medidas de posição e variabilidade são exemplos simples de sumarização. Funções mais sofisticadas envolvem técnicas de visualização e a determinação de relações funcionais entre variáveis. As funções de sumarização são frequentemente usadas na análise exploratória de dados com geração automatizada de relatórios, sendo responsáveis pela descrição compacta de um conjunto de dados. A sumarização é utilizada, principalmente, no pré-processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas – como mínimo, máximo, média, moda, mediana e desvio padrão amostral –, no caso de variáveis quantitativas, e, no caso de variáveis categóricas, por meio da distribuição de freqüência dos valores. Técnicas de sumarização mais sofisticadas são chamadas de visualização, que são de extrema importância e imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados. Exemplos de técnicas de visualização de dados incluem diagramas baseados em proporções, diagramas de dispersão, histogramas e *box plots*, entre outros. Autores como Levine *et al.* (2000) e Martins (2001), entre outros, abordam com grande detalhamento esses procedimentos metodológicos.

e) *Modelo de Dependência*: descreve dependências significativas entre variáveis. Modelos de dependência existem em dois níveis: estruturado e quantitativo. O nível estruturado especifica, geralmente em forma de gráfico, quais variáveis são localmente dependentes. O nível quantitativo especifica o grau de dependência, usando alguma escala numérica.

Segundo Padovani (2000), análises de dependência são aquelas que têm por objetivo o estudo da dependência de uma ou mais variáveis em relação a outras, sendo procedimentos metodológicos para tanto a análise discriminante, a de medidas repetidas, a de correlação canônica, a de regressão multivariada e a de variância multivariada.

f) *Regras de Associação*: determinam relações entre campos de um banco de dados. A idéia é a derivação de correlações multivariadas que permitam subsidiar as tomadas de decisão. A busca de associação entre variáveis é, freqüentemente, um dos propósitos das pesquisas empíricas. A possível existência de relação entre variáveis orienta análises, conclusões e evidenciação de achados da investigação. Uma regra de associação é definida como *se X então Y*, ou $X \Rightarrow Y$, onde X e Y são conjuntos de itens e $X \cap Y = \emptyset$. Diz se que X é o antecedente da regra, enquanto Y é o seu conseqüente. Medidas estatísticas como correlação e testes de hipóteses apropriados revelam a freqüência de uma regra no universo dos dados minerados. Vários métodos para medir associação são discutidos por Mattar (1998), de natureza paramétrica e não paramétrica, considerando a escala de mensuração das variáveis.

g) *Análise de Séries Temporais*: determina características seqüenciais, como dados com dependência no tempo. Seu objetivo é modelar o estado do processo extraindo e registrando desvios e tendências no tempo. Correlações entre dois instantes de tempo, ou seja, as observações de interesse, são obtidas em instantes sucessivos de tempo – por exemplo, a cada hora, durante 24 horas – ou são registradas por algum equipamento de forma contínua, como um traçado eletrocardiográfico. As séries são compostas por quatro padrões: tendência, variações cíclicas, variações sazonais e variações irregulares. Há vários modelos estatísticos que podem ser aplicados a essas situações, desde os de regressão linear (simples e múltiplos), os lineares por transformação e regressões assintóticas, além de modelos com defasagem, como os autoregressivos (AR) e outros deles derivados. Uma interessante noção introdutória ao estudo de séries temporais é desenvolvida por Morettin & Toloi (1987).

A partir dessas informações pode-se perceber que a *Data Mining* preocupa-se também com a análise secundária dos dados, num sentido mais amplo e mais

indutivo do que uma abordagem hipotético-dedutiva, freqüentemente considerada como o paradigma para o progresso da ciência moderna. Assim, *Data Mining* pode ser visto como o descendente direto da estatística, e algumas de suas técnicas utilizam métodos estatísticos puros, com as redes bayesianas (VEIGA e SILVA, 2002).

## 2.6 LINGUAGEM "R"

R é uma linguagem e ambiente para computação estatística e gráficos. É um projeto General Public License (GNU) que é similar à linguagem e ambiente S, que foi desenvolvida no laboratório Bell (anteriormente AT & T, agora Lucent Technologies) por John Chambers e outros colegas. A linguagem R pode ser considerada como uma implementação diferente de S. Há algumas diferenças importantes, mas muitos códigos escritos para a linguagem S também funciona em R.

A linguagem R fornece uma ampla variedade de estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, clustering, dentre outros) e técnicas gráficas, e ela também é extensível. A linguagem S é, muitas vezes, o veículo de escolha para a pesquisa em metodologia estatística, porém a linguagem R fornece uma rota de código aberto para a participação nessa atividade de maneira interativa. Um dos pontos fortes da linguagem R, é a facilidade com que, se bem projetado, publicações de qualidade-plots podem ser produzidos, incluindo símbolos e fórmulas matemáticas, quando necessário. Um dos objetivos no desenvolvimento dessa linguagem foi cuidar para que os padrões de designfossem feitos em gráficos não muito elaborados, mas mesmo com essa opção o usuário ainda mantém o controle total do programa.

O ambiente R está disponível como Software Livre sob os termos da licença da Fundação Software Livre General Public License em forma de código fonte. Ele compila e roda em uma grande variedade de plataformas UNIX e sistemas similares (incluindo FreeBSD e Linux), Windows e MacOS. O ambiente R é um conjunto integrado de instalações de software para manipulação de dados, cálculo e apresentação gráfica. A linguagem inclui:

- uma eficaz forma de manuseio e instalação de armazenamento de dados;
- um conjunto de operadores para cálculos sobre arrays, matrizes em particular;
- uma coleção grande, coerente e integrada de ferramentas intermediárias para análise de dados;
- instalações gráficas para análise de dados e exibição ou não no ecrã ou na via impressa;
- uma desenvolvida linguagem de programação que inclui condicionais, loops, funções definidas pelo usuário e recursiva de entrada e saída de instalações.

O termo "ambiente" destina-se a caracterizá-la como um sistema planejado e coerente, em vez de um acréscimo incremental de ferramentas muito específicas e inflexíveis, como é frequentemente o caso com outro software de análise de dados.

A linguagem R, como a linguagem S, é projetado em torno de uma linguagem de computador verdadeira, e que permite aos usuários adicionar funcionalidade adicional definindo novas funções. Grande parte do sistema em si é escrito na linguagem R parecida com o dialeto S, o que torna fácil para os usuários seguir as escolhas algorítmicas feitas. Para tarefas de computação intensiva, C, C + + e código Fortran podem ser ligados e chamados em tempo de execução. Os usuários avançados podem escrever código C para manipular objetos R diretamente.

Muitos usuários pensam que a linguagem R é um sistema de estatísticas. Porém define-se ela como um ambiente no qual as técnicas estatísticas são implementadas. A R pode ser estendida (facilmente) através de pacotes. Há cerca de oito pacotes fornecidos com a distribuição R e muitos mais estão disponíveis através da família CRAN de sítios da Internet que cobrem uma vasta gama de estatísticas modernas.

R tem o seu formato de documentação própria LaTeX-like, que é usado para fornecer documentação completa, tanto on-line em vários formatos e em via impressa.

# 3 METODOLOGIA

Para que os objetivos sejam cumpridos, procura-se caracterizar e analisar o ambiente de pesquisa, a definição da amostra dos dados, a forma de coleta, sistematização e análise dos dados obtidos.

Quanto aos objetivos, Gil (2002) define que uma pesquisa pode ser:

a) exploratória, que tem por objetivo a exploração de um assunto por meio da aproximação de fatos, o que se realiza através da identificação, do acesso e uso de materiais informacionais – fontes;

b) descritiva, cujo objetivo é observar, registrar e analisar os fenômenos sem, entretanto, entrar no mérito de seu conteúdo;

c) explicativa, que, além de registrar, analisar e interpretar os fenômenos estudados tem como preocupação primordial identificar os fatores que determinam ou que contribuem para a ocorrência dos fenômenos, isto é, suas causas.

No que diz respeito à natureza da pesquisa, Severino (2002) esclarece que ela pode ser quantitativa ou qualitativa. Uma pesquisa quantitativa traduz em números os resultados obtidos, valendo-se de técnicas estatísticas. E na pesquisa qualitativa, os resultados não são quantificáveis, sendo a interpretação da realidade um fator primordial.

Conclui-se, a partir das definições, que pesquisas descritivas tem por objetivo, descrever um ambiente, processo de um determinado fato ou fenômeno. Portanto esta pesquisa pode ser classificada como descritiva.

Ainda a partir das definições dos autores citados pode-se concluir que a pesquisa terá a abordagem qualitativa, pois se trata de um estudo caracterizado pela identificação e análise de dados.

Para cumprir os objetivos desse estudo, inicialmente será realizada uma revisão da literatura pertinente. A seguir serão coletadas as informações livres na Internet sobre os usuários da rede social Twitter, em especial os “seguidores” da empresa escolhida. Dando sequência ao estudo será aplicado um dos métodos de

mineração de dados para identificar um possível padrão entre esses usuários e por fim esses dados tratados serão analisados e avaliados pela autora dessa pesquisa.

O ambiente da pesquisa será a rede social Twitter, mais especificamente o perfil da organização Little Brown Book Group. A organização é uma editora de livros impressos.

## 3.1 AMBIENTE DE PESQUISA: LITTLE, BROWN BOOK GROUP

A organização escolhida para esse estudo é a Little, Brown Book Group. A escolha dessa organização se deu por ela atuar em um ramo de négocio que possui muitos concorrentes, como por exemplo os E-books, livros virtuais, Audio books dentre outros. Uma vantagem competitiva pode vir a se tornar decisiva para atrair novos clientes. Outro motivo também é que o software utilizado para realizar os testes utiliza a linguagem inglesa o que ocasionaria traduzir os dados do português para o inglês e, pelo tempo não muito longo da pesquisa, não seria viável essa tradução.

Essa editora venceu três vezes o prêmio “Publisher of the Year Award (editor do ano)”, venceu também, em 2010, o “Bookseller Industry Awards (Prêmio da Indústria Livreira)”. A Little, Brown Book Group tem um histórico de entrega de *best sellers* em todas as esferas de publicação da editora.

A Little, Brown Book Group é uma editora criada por Charles Coffin Little e seu parceiro, James Brown. Desde 2006 tem sido uma unidade constituinte da Hachette Book Group no Estados Unidos.

A organização é focada na publicação de livros de ficção ou não-ficção, buscando publicar um número limitado de títulos porém com variedade temas. A empresa apresenta uma missão de “Comprar o melhor material em qualquer categoria e publicá-lo com talento e entusiasmo”.

A Little, Brown Book Group define sua visão em “Concentrar-se em obter o máximo de cada livro através da atenção detalhada a cada fase do processo, desde editorial de design, as campanhas de marketing e publicidade e da estratégia de varejo com clientes-chave.” (Site Little Brown Book Group 2011)

A empresa iniciou suas atividades especializada em tratados jurídicos e títulos importados. Por muitos anos ela foi a maior editora das leis dos Estados Unidos, e

também a maior importadora da lei Inglesa padrão além de importar obras diversas. Ela introduziu também, aos compradores americanos, a Enciclopédia Britânica, os dicionários de William Smith, e muitas outras obras-padrão. Ainda assim, nos primeiros anos Little e Brown publicou as obras de Daniel Webster, a História dos Estados Unidos de George Bancroft, Ferdinand William H. Prescott 's e Isabella Jones, dentre outros.

A partir desse histórico a organização foi adquirindo novos títulos e direitos autorais com diversos autores e ganhando novos clientes com os mais diversos tipos de livros.

## 3.2 DESCRIÇÃO DO MÉTODO DE RECUPERAÇÃO DE TWITTES COM A LINGUAGEM R

O desenvolvimento dessa pesquisa deu-se com a ajuda do aluno de mestrado Frank Coelho de Alcantara, o qual desenvolveu um material de apoio para auxiliar o melhor entendimento da nova linguagem e a aplicação da mesma.

Primeiramente, a partir das pesquisas realizadas nas referências bibliográficas disponíveis, optou-se por utilizar o método de mineração de dados de sumarização, este por ser mais adequado a base de dados utilizada e ao tipo de resultado esperado dessa pesquisa.

Para realizar esse procedimento, foi baixada a ferramenta “Revolutions R” na versão para estudante acadêmico. Esta ferramenta que viabiliza a leitura da linguagem R. A partir da instalação desse programa deve-se instalar também alguns pacotes de funções mais específicas que trazem um resultado mais relevante para uma análise posterior. Os pacotes que devem ser adicionados para pesquisas com o Twitter são o TwitterR, o plyr e o stingr pois, esses são os pacotes que utilizam a linguagem R e que são adaptadas para pesquisa no twitter.

A partir desses pacotes já é possível iniciar o código. Esses pacotes devem ser carregados no código para uso de suas funções. Depois desses pacotes carregados deve-se iniciar a construção da base de dados.

Deve-se iniciar a primeira linha definindo qual é o termo a ser recuperado do Twitter que terá relevância para a análise. Essa etapa deve ser descrita dessa forma:

**Comando 2: Comando de busca dos termos usados na pesquisa**

```
recuperados = searchTwitter('termo buscado', n=500, lang='en')
```

Fonte: <http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> 2011

Nesse exemplo a parte destacada em itálico deve ser trocada pelo termo da pesquisa, o n representa o número de palavras que serão recuperados e que contém esse termo e o **lang** representa a língua em que serão recuperados os dados. Nesse caso os dados trabalhados serão em inglês. A escolha da linguagem se deu porque o dicionário de palavras escolhido como critério de decisão está na língua inglesa. Porém a pesquisa pode ser feita em qualquer linguagem desde que o dicionário das palavras seja da mesma língua dos termos buscados.

A próxima linha irá apresentar uma função da linguagem R que cria um método de texto que lê os “twittes” recuperados. Essa função é descrita dessa forma:

**Comando 3: Comando de leitura dos termos recuperados**

```
recuperados.text=laply(recuperados,function(t) t$getText() )
```

Fonte: <http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> 2011

Após essa linha deve-se complementar a função com a seguinte linha, para finalizar a leitura dos termos:

**Comando 4: Comando de leitura dos termos recuperados (continuação)**

```
head(recuperados.text, 10)
```

Fonte: <http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> 2011

Executando esses comandos deve-se visualizar uma base de dados parecida com o exemplo abaixo:

**Comando 5: Exemplos de dados recuperados sem tratamento**

```
[1] "RT @onemorepage: Read my interview with @AliMcNamara here: http://t.co/HlRsEKZ6
#breakfastatdarcys @LittleBrownUK"
[2] "Read my interview with @AliMcNamara here: http://t.co/HlRsEKZ6 #breakfastatdarcys
@LittleBrownUK"
[3] "@LittleBrownUK http://t.co/PGY3lUaL"
[4] "LiteraryFrenzy #FF @LittleBrownUK @sarahduncan1 @ViragoBooks @HodderBooks
@simonschuster @elizabethbuchan @KatieFforde Thanks for the follow"
[5] "@Clairywoowoo @davidhepworth @nilerodgers @LittleBrownUK That's a wonderful snippet
Clairy"
[6] "@Clairywoowoo @nilerodgers @LittleBrownUK When you're recording a podcast there's nobody
more welcome than a trouper. And he's a trouper."
```

```
[6] "@Clairywoowoo @nilerodgers @LittleBrownUK When you're recording a podcast there's nobody
more welcome than a trouper. And he's a trouper."
[7] "@davidhepworth @nilerodgers I sang to him at a @LittleBrownUK lunch. It got very emotional at
the end. I love that man. I bet you do too ;-)"
[8] "RT @LittleBrownUK: RT @rockmother: Stage invaaaasiioon @nilerodgers and Chic Org London
style http://t.co/LOvyvVbV"
[9] "RT @Shotsblog: Info on Shotsblog about @1pcornwell event @sharoncanavar
@TheakstonsCrime @LittleBrownUK\r http://t.co/ZVYmabqz"
[10] "@LittleBrownUK http://t.co/c4RvWM3t"
```

Fonte: A autora 2011

Os números que aparecem antes de cada objeto representam a ordem em que foram recuperados pela função descrita nas três linhas anteriores.

A partir desses comandos a base de dados a ser analisada está pronta. Após esses comandos, é preciso acrescentar um dicionário de palavras positivas e palavras negativas para que seja possível atribuir valores aos “twittes” que mais tarde servirão como base para uma análise e posteriormente a transformação desses dados em informação.

Um dicionário de dados utilizado para atribuir valores para esses dados é o dicionário Opinin Lexicon. Ele é uma lista de palavras consideradas positivas e negativas criada na Universidade de Illinois em Chicago, Estados Unidos, no Departamento de Ciência da Computação. Esta lista foi compilada durante anos a partir do primeiro trabalho realizado sobre o tema de mineração de dados web (Hu e Liu, KDD-2004). A lista do Opinin Lexicon pode ser baixada por qualquer usuário já que é um documento livre.Para utilizar esse dicionário, deve-se carregar os arquivos no diretório do console do projeto no programa. Pode-se realizar essa tarefa pelo comando no menu File e escolher a opção R Working Directory. Então deve-se escolher a pasta que foi salvo o dicionário.

A partir disso deve-se escrever o seguinte comando:

**Comando 6: Comando de leitura do dicionário de palavras positivas**

```
positivos = scan('positive-words.txt', what='character', comment.char=';')
```

Fonte: <http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> 2011

Após esse comando deve-se digitar um “enter”. O programa irá mostrar que foram lidos 2006 itens para a base de dados em análise. A seguir deve-se digitar o comando:

**Comando 7: Comando de leitura do dicionário de palavras negativas**

```
negativos = scan('negative-words.txt', what='character', comment.char=';')
```

Fonte: <http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> 2011

Esse comando irá retornar a leitura de 4783 itens. A partir dessas listas, o próximo passo é criar uma função que fará o levantamento do “humor” de cada “twitte”. Essa função é chamada de score.sentiment. O objetivo dessa função é criar uma tabela com duas colunas: o score e o “twittet”. O código a seguir vai criar a função que é necessária para contar as palavras positivas e negativas em um “tweet” e atribuir um valor, positivo se o número de palavras positivas for maior que o número de palavras negativas e vice-versa.

**Comando 8: Código de programação na linguagem R para a recuperação e classificação dos tweets**

```
score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
{
require(plyr)
require(stringr)
scores = laply(sentences, function(sentence, pos.words, neg.words) {
sentence = gsub('[[:punct:]]', '', sentence)
sentence = gsub('[[:cntrl:]]', '', sentence)
sentence = gsub('\\d+', '', sentence)
sentence = tolower(sentence)
word.list = str_split(sentence, '\\s+')
words = unlist(word.list)
pos.matches = match(words, pos.words)
neg.matches = match(words, neg.words)
pos.matches = !is.na(pos.matches)
```

```
neg.matches = !is.na(neg.matches)
score = sum(pos.matches) - sum(neg.matches)
                return(score)
                }, pos.words, neg.words, .progress=.progress )
scores.df = data.frame(score=scores, text=sentences)
return(scores.df)
}
```

Fonte: <http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> 2011

Uma vez que esta função tenha sido digitada, pode-se utilizar a função para fazer o score dos tweets que foram recuperados. Deve-se digitar o seguinte comando:

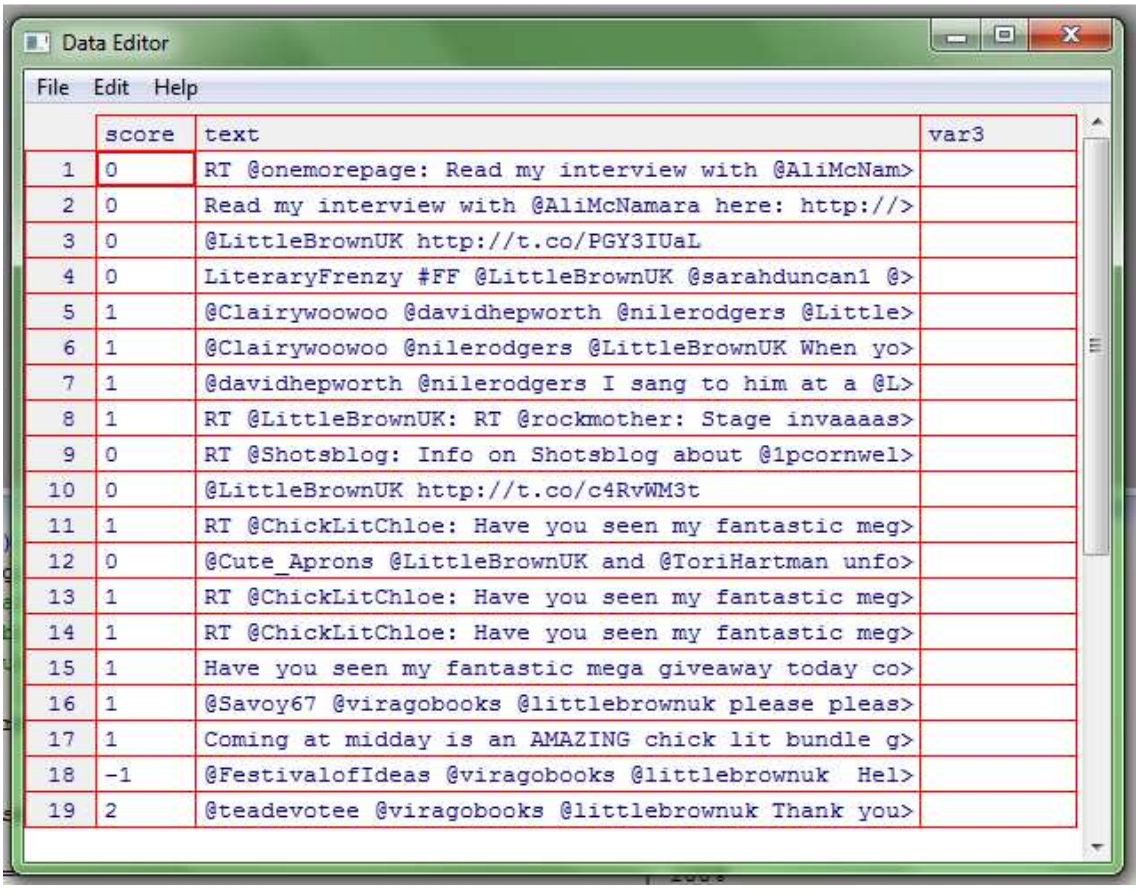
**Comando 9: Comando para agregar valor aos dados recuperados**

```
recuperados.score = score.sentiment(recuperados.text,positivos,negativos,
.progress='text')
```

Fonte: <http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> 2011

Este comando irá fazer a classificação e contagem de cada linha do objeto recuperados.text. Uma opção é ver o resultado em forma de tabela. Clicando-se com o botão direito do mouse sobre o objeto “recuperados.score” na janela “Object Browser” do programa Revolutions e escolher a opção “Edit Objeto”. Deve-se ver algo como na figura a seguir:

**Figura 10: Classificação dos dados recuperados (exemplo)**




| | score | text | var3 |
|---|---|---|---|
| 1 | 0 | RT @onemorepage: Read my interview with @AliMcNam> | |
| 2 | 0 | Read my interview with @AliMcNamara here: http://> | |
| 3 | 0 | @LittleBrownUK http://t.co/PGY3IUaL | |
| 4 | 0 | LiteraryFrenzy #FF @LittleBrownUK @sarahduncan1 @> | |
| 5 | 1 | @Clairywoowoo @davidhepworth @nilerodgers @Little> | |
| 6 | 1 | @Clairywoowoo @nilerodgers @LittleBrownUK When yo> | |
| 7 | 1 | @davidhepworth @nilerodgers I sang to him at a @L> | |
| 8 | 1 | RT @LittleBrownUK: RT @rockmother: Stage invaaaas> | |
| 9 | 0 | RT @Shotsblog: Info on Shotsblog about @1pcornwel> | |
| 10 | 0 | @LittleBrownUK http://t.co/c4RvWM3t | |
| 11 | 1 | RT @ChickLitChloe: Have you seen my fantastic meg> | |
| 12 | 0 | @Cute_Aprons @LittleBrownUK and @ToriHartman unfo> | |
| 13 | 1 | RT @ChickLitChloe: Have you seen my fantastic meg> | |
| 14 | 1 | RT @ChickLitChloe: Have you seen my fantastic meg> | |
| 15 | 1 | Have you seen my fantastic mega giveaway today co> | |
| 16 | 1 | @Savoy67 @viragobooks @littlebrownuk please pleas> | |
| 17 | 1 | Coming at midday is an AMAZING chick lit bundle g> | |
| 18 | -1 | @FestivalofIdeas @viragobooks @littlebrownuk Hel> | |
| 19 | 2 | @teadevotee @viragobooks @littlebrownuk Thank you> | |

Fonte: A autora 2011

O resultado do código R será apresentado nessa tabela. A partir dela devem ser tiradas as conclusões para uma análise aprofundada.

# 4 RESULTADOS DOS EXPERIMENTOS NA FERRAMENTA REVOLUTIONS R COM A LINGUAGEM R

Esse experimento foi realizado no dia vinte de novembro de dois mil e onze com o objetivo de verificar a opinião dos usuários do Twitter sobre alguns termos relacionados a organização escolhida. O primeiro teste foi realizado com o termo "@LittleBrownUK " que é o perfil da empresa no Twitter. Foram obtidos os seguintes resultados da tabela 1:

**Tabela 11: Tabela dos resultados do teste 1**

| | |
|---|---|
| Objetos recuperados considerados positivos | 5 |
| Objetos recuperados considerados neutros | 8 |
| Objetos recuperados considerados negativos | 6 |
| Total de objetos Recuperados | 19 |

Fonte: A autora 2011

A tabela acima é interpretada da seguinte forma: os objetos recuperados são os tweets lidos por meio do código descrito na seção anterior, que totalizaram 19 menções sobre o termo buscado. Como resultado foi obtido cinco objetos positivos, oito objetos considerados neutros e seis objetos considerados negativos. Todas essas classificações foram realizadas com base no dicionário de palavras descrito também na seção anterior.

Essa mesma tabela mostrou uma tendência maior para as palavras neutras. Isso pode representar que a imagem da organização está com uma imagem neutra para os usuários do Twitter.

Essa afirmação representa que a organização cumpre o seu propósito de publicar bons livros para o consumidor final. Porém ela não está tendo um produto, ou vários, com um destaque positivo que reflita na opinião dos usuários nessa rede social.

Como foi realizado um teste único sobre um termo específico, para se ter certeza da opinião dos usuários deveria-se repetir esse teste mais de uma vez para

concluir melhor se isso é uma opinião consolidada ou apenas um fato isolado desencadeado por algo que pudesse ter provocado esse comportamento.

Por meio dessa tabela pode-se perceber que, no momento da coleta de dados, a maioria deles foi classificado de forma neutra. A seguir os objetos considerados negativos tiveram a segunda maior frequência e por uma pequena margem os dados que tiveram menor frequência foram os objetos considerados positivos.

O próximo teste foi realizado no mesmo dia do teste anterior porém com o termo “StevesJobsBook”. Esse termo representa um livro publicado pela organização, classificado como o livro do mês pela própria organização. O resultado está representado na tabela:

**Tabela 12: Tabela dos resultados do teste 2**

| | |
|---|---|
| Objetos recuperados considerados positivos | 6 |
| Objetos recuperados considerados neutros | 10 |
| Objetos recuperados considerados negativos | 3 |
| Total de objetos Recuperados | 19 |

Fonte: A autora 2011

Esta tabela também é interpretada da mesma forma que a primeira: como resultado foi obtido seis objetos positivos, dez objetos considerados neutros e três objetos considerados negativos. Todas essas classificações também foram realizadas com base no dicionário de palavras descrito também na seção anterior.

Interpretando a tabela ve-se que a maioria dos objetos estão classificados na categoria de objetos neutros. A seguir a maior frequência é a dos objetos de palavras positivas e por último estão os objetos classificados como negativos.

Avaliando a segunda tabela apresentada, sobre o termo “SteveJobsBook”, nota-se, primeiramente,que a categoria das palavras neutras ocuparam praticamente a metade dos dados analisados.

Esta classificação mostra que esse livro também recebe a classificação mediana. O livro é uma biografia do executivo Steve Jobs que morreu nesse ano de

2011. Jobs foi o criador da marca Apple e uma figura importante no mundo dos negócios. Por causa de sua morte, ele tem tido sua história repetida em vários meios de comunicação. Um deles é o livro publicado pela organização estudada. Por meio do teste, pode-se perceber que o livro tem uma boa aceitação, já que as palavras negativas apresentaram uma baixa frequência, o que significa que o livro não recebeu muitas críticas no momento do teste ser realizado.

O próximo teste foi realizado com o termo "@AtomBooks" que é uma editora na Inglaterra que faz parte do grupo Little, Brown Books. O resultado foi o seguinte:

**Figura 13: Tabela dos resultados do teste 3**

| | |
|---|---|
| Objetos recuperados considerados positivos | 9 |
| Objetos recuperados considerados neutros | 7 |
| Objetos recuperados considerados negativos | 4 |
| Total de objetos recuperados | 20 |

Fonte: A autora 2011

A tabela apresentada é interpretada do mesmo jeito que as anteriores: os resultados foram obtidos como nove objetos positivos, sete objetos considerados neutros e quatro objetos considerados negativos. Todas essas classificações também foram realizadas com base no dicionário de palavras descrito também na seção anterior.

Le-se essa tabela a partir de uma grande quantidade de objetos na categoria de objetos considerados positivos, sendo nove deles classificados nessa categoria. A segunda maior frequência vem da categoria de objetos considerados neutros, sendo que esses tem um número de sete objetos. Por fim a categoria com objetos negativos apresentam uma frequência menor de quatro objetos.

Essa tabela apresentou a melhor avaliação dos três testes realizados. Isso pode ter ocorrido devido ao fato de o público alvo é diferente, já que o mesmo está localizado na Inglaterra. Esse teste foi o único a ter o maior resultado sendo de palavras positivas. Os dados considerados negativos para a organização foram classificados bem abaixo com relação aos outros portanto faz-se uma conclusão que a filial do reino Unido tem uma imagem positiva frente aos usuários no Twitter.

Analisando os três resultados juntos pode-se realizar uma avaliação positiva da organização. Alguns resultados isolados apresentaram uma opinião diferente da maioria porém, o Twitter possui muitos usuários em todo o mundo, mais de 100 milhões, segundo o site globo.com, o que significa que é muito difícil atingir uma aceitação alta da maioria dos usuários.

Nenhum dos testes apresentou um resultado muito diferente uns dos outros o que também indica que, como os termos estão relacionados entre si, mostra a homogeneidade da opinião dos usuários/consumidores nessa rede social específica.

Esse teste de mineração de dados web é válido por seguir todo o processo do KDD explicado na literatura pertinente. O primeiro passo de seleção da base de dados foi escolhido no momento de decisão da escolha do tema e do interesse pela pesquisa da autora. A seleção da base de dados está explicada na seção onde explica a metodologia de recuperação dos Twittes. A etapa de pré-processamento, transformação de dados e mineração de dados também está explicado nessa mesma seção. A interpretação dos dados foi explicada na seção anterior e a construção do conhecimento foi detalhada nessa própria etapa.

A partir de todas essas etapas percebe-se que a mineração de dados web é válido para buscar informações e criar conhecimento sobre o mercado consumidor em uma base de dados livre onde pode-se buscar vários tipos de dados e criar o conhecimento necessário para gerar vantagens competitivas e atrair a geração Y e Z para não só como consumidores mas também como futuros funcionários.

# 5 CONSIDERAÇÕES FINAIS

Dada a importância das informações no mundo atual para a competitividade das organizações, as novas fontes de dados não devem ser descartadas por falta de conhecimento de como utilizá-las. As redes sociais são um canal de comunicação direto com o consumidor final, onde não é preciso pedir que os clientes dêem suas opiniões sobre as empresas e seus produtos, eles mesmos mostram sem a censura de um pesquisador ou a pressão de responder um questionário.

Ao se expor a uma ferramenta na Web 2.0, onde o compartilhamento de informações é o foco principal, com consumidores cada vez mais exigentes e buscando produtos que representem cada vez mais a sua própria identidade, a organização pode atrair cada vez mais esse público e construir uma relação de confiança e fidelidade, os principais objetivos de todas as empresas.

Porém dados sem tratamento não possuem um valor adequado para uma avaliação ou teste de aceitação de algum novo produto por exemplo. É nesse quesito que a mineração de dados web, nesse caso específico as redes sociais, pode ajudar as organizações a compreenderem melhor a opinião de seus clientes sobre ela e até mesmo sobre os concorrentes já que, as redes sociais são uma “base de dados” livre na Internet, basta saber o que buscar e como buscar.

Acompanhar todas essas mudanças num mundo globalizado não é uma tarefa fácil. Se a organização é uma Multinacional, ela precisa conhecer não só a cultura de origem da organização, mas também precisa estudar como os funcionários das filiais agem e constroem suas opiniões.Essa é outra vantagem das redes sociais, pois ela é livre de barreiras como língua, costumes, fronteiras ou classe social. Qualquer um que tenha um computador com acesso a Internet pode criar um perfil no Twitter por exemplo. Não é necessário nem que o seu nome no perfil seja verdadeiro, desse modo os usuários se sentem livres para mostrar suas opiniões sem medos ou censura de conhecidos.

Portanto as redes sociais apresentam uma riqueza e quantidade de dados grandes disponíveis o que torna o estudo da mineração desses dados uma vantagem competitiva, além de tornar a imagem da organização mais adequada e mais identificável com um público que permanece conectado as redes sociais praticamente 24 (vinte e quatro) horas por dia.

A partir das pesquisas realizadas para este estudo, percebeu-se que o tema

de busca de dados em redes sociais ainda é um tema pouco estudado no Brasil. No momento da busca de materiais referentes a esse tema, verificou-se que a maioria da literatura disponível foi realizada por pesquisadores dos Estados Unidos e da Europa. Isso pode ser justificado pela criação da maioria das redes sociais terem sido criadas nesses lugares.

## 5.1 OBJETIVOS ALCANÇADOS

Este estudo teve como objetivo analisar os usuários do Twitter, mais especificamente os que citaram alguma opinião relevante sobre a organização Little, Brown Book Group, para verificar se essa avaliação apresenta algum padrão que possa ser utilizado pela própria organização. A partir dos resultados apresentados nos três testes realizados, pode-se perceber que há um padrão na opinião dos usuários relacionados com a organização estudada.

Nos três testes a opinião dos usuários foi positiva em relação ao perfil principal da organização, a um produto específico e a ao perfil de uma das editoras que faz parte do grupo Little, Brown Group. Isso representa que os usuários do Twitter estão satisfeitos com a organização. Mas isso também representa que a organização não é um assunto muito discutido entre os usuários.

Derivado desse objetivo principal tem-se os objetivos específicos. Porém, para que os resultados fossem alcançados de maneira satisfatória, foi feito um levantamento bibliográfico dos seguintes temas: Evolução das Gerações, Web 2.0, Redes Sociais, Twitter, Mineração de dados, Análise de sentimento e Linguagem R

Para se conseguir uma boa fundamentação teórica para início desse estudo, buscou-se compreender a evolução das gerações da sociedade mundial, para compreender o porque de o Twitter ser acessado por mais de 100 milhões (site Globo.com) de pessoas em todo o mundo.

Também foi estudada a Web 2.0. Ela é a base das redes sociais, já que seu objetivo principal é o compartilhamento. Com essa etapa da pesquisa, pode-se compreender a temática das redes sociais e o futuro das mesmas.

Inserido no contexto da sociedade atual, foi escolhido para a próxima etapa de estudo as próprias redes sociais. Nessa etapa foi estudado o início das dessas

redes, o porque de serem tão populares e como elas funcionam. Foi dado um aprofundamento maior ao Twitter, pois ele seria o objeto desse estudo.

Estudadas as redes sociais foi a vez da análise de sentimentos. Esse tema foi escolhido para compreender melhor os resultados finais da sumarização e o que eles representam.

A seguir foi estudada a mineração de dados em si. Foram buscadas informações sobre início dessa área de trabalho, métodos mais utilizados e conhecidos e também a temática principal da mineração de dados. Assim como nas redes sociais foi dada uma ênfase à linguagem R, pois ela foi utilizada nos experimentos desse estudo.

Feito o referencial teórico deu-se início aos objetivos específicos. O primeiro objetivo específico realizado foi a descrição da organização estudada, a Little, Brown Book Group. Esse objetivo foi cumprido com uma leve dificuldade devido a origem da organização ser os Estados Unidos, onde a língua oficial é a Inglesa.

O segundo objetivo específico era aplicar o método de sumarização no Twitter por meio da ferramenta Revolutions R. Esse objetivo atingiu as expectativas com a ajuda do aluno de mestrado Frank Coelho de Alcantara, o qual ajudou a aluna desenvolvedora dessa pesquisa a entender como essa linguagem funcionava e onde poderia ser buscadas mais informações sobre esse assunto.

Por fim o terceiro objetivo específico era analisar esses dados recuperados no Twitter para compreender se esse método de pesquisa representa um valor significativo para as organizações investirem seu tempo e dinheiro ou não.

Como conclusão desses objetivos pode-se perceber que a mineração de dados é uma ferramenta útil para as organizações pois, com os testes realizados viu-se que os dados se transformam em informações por meio de gráficos e da interpretação do analista. Esses testes podem trazer informações como: aceitação de um novo produto ou concorrente, conhecimento de alguma ação que a organização realizou e prejudicou sua imagem (uma ação contra o meio ambiente, por exemplo) e até o conhecimento dos produtos e serviços que a organização deve investir ou desistir.

## 5.2 PROJETOS FUTUROS

O tema de mineração de dados em redes sociais possui muito campo para pesquisas e aprofundamento do tema. Uma opção para a continuação desse estudo seria utilizar outros métodos e outras ferramentas para comparação com os resultados obtidos por meio da linguagem R.

Finalmente pode ser realizado um experimento com diversas empresas do mesmo ramo para analisar a diferença de opiniões dos usuários dentre as empresas para compreender as oportunidades, ameaças e novas opções de atuação.

Por fim pode-se escolher criar um dicionário das palavras negativas e positivas para a língua portuguesa. Com esse dicionário seria possível realizar estudos nos perfis das organizações brasileiras com mais facilidade e de maneira mais freqüente.

## REFERÊNCIAS


ADDRIANS, P. & ZANTINGE, D. **Data Mining.** Inglaterra: Addison-Wesley, 1996

BOOMBUST Disponível em: <http://www.boombust.com.br/a-hora-e-a-vez-das-midias-sociais/> conceito MS. Acesso em 23 maio 2011.

BOYD, D. Friendster and Publicly Articulated Social Network. **Conference on Human Factors and Computing Systeams** (CHI 2004). Vienna:ACM, Abril 24-29, 2004.

BRACHNAD, R.J. & ANAND, T. The process of knowledge discovery in databases. *In*: FAYYAD, U.M. *et al*. **Advances in Knowledge Discovery in Data Mining.** Menlo Park: AAAI Press, 1996.

CINE GESTÃO <http://cinegestao.blogspot.com/2009/12/conflito-de-geracoes-post-1-os.html>

DILLY, R. **Data Mining**: an introduction. Belfast: Parallel Computer Centre, Queens University, 1999.

DINIZ, C.A. &LOUZADA-NETO, F. **Data Mining**: uma introdução. São Carlos: Associação Brasileira de Estatística, 2000.

FAYYAD, U.M. *et al*. The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: **Advances in Knowledge Discovery in Data Mining.** Menlo Park: AAAI Press, 1996a.

FAYYAD, U.M. *et al*. **Advances in Knowledge Discovery and Data Mining.** California: AAAI Press, 1996b.

FOCO EM GERAÇÕES<http://www.focoemgeracoes.com.br>

GIL, A. C. **Como elaborar projetos de pesquisa**. 4 ed. São Paulo. Atlas, 2002.

GOUVEIA, L. M. B. **Notas de contribuição para uma definição operacional** Disponível em <http://www2.ufp.pt/~lmbg/reserva/lbg_socinformacao04.pdf> Acesso em 14 jun. 2011

HSMGLOBAL<http://br.hsmglobal.com/adjuntos/14/documentos/000/060/0000060367.pdf> Acesso em 17 abr. 2011

JEFFREY BREEN Disponível em: <http://jeffreybreen.wordpress.com/2011/07/04/twitter-text-mining-r-slides/> Acesso em: 07 nov. 2011

Jindal N, Liu B, **Opinion Spam and Analysis** Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2008.

LEMOS, A; PALÁCIOS, M. **As janelas do Ciberespaço**. Porto Alegre: Sulina. 2001.

Leslie D'Monte **Swine flu's tweet tweet causes online flutter.** Business Standard.

LEVINE, D.M. *et al*. **Estatística: teoria e aplicações.** Trad. Teresa C.P. de Souza. Rio de Janeiro: LTC Editora, 2000.

LÉVY, P. **Cibercultura**. São Paulo: Ed. 34, 1999.

LIPNACK, J, STAMP, J. **Network,redes de conexão:pessoas conectando-se com pessoas.** São Paulo: Aquarela,1992.

LITTLE, BROWN BOOK GROUP < http://www.littlebrown.co.uk/home> Acesso em 28 out. 2011

Liu B, **Sentiment Analysis and Subjectivity**. Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

LOIOLA, E.; MOURA, S. Análise de redes: uma contribuição aos estudos organizacionais. In: FISHER, T. (Org.). **Gestão Contemporânea, cidades estratégias e organizações locais.** Rio de Janeiro: Fundação Getúlio Vargas, 1997. p.53-68.

MANNILA, H. Data mining: machine learning, statistics and databases. **International Conference on Statistics and Scientific Database Management** Estocolmo, 8, 1996.

MARTINS, G.A. **Estatística Geral e Aplicada.** São Paulo: Atlas, 2001.

MATTAR, F.N. **Pesquisa de Marketing.** São Paulo: Atlas, 1998.

MORETTIN, P.A. & TOLOI, C.M. **Séries Temporais**. 2.ª ed. São Paulo: Atual, 1987.

MUNDO DO MARKETING <http://www.mundodomarketing.com.br/16,11349,maioria-das-empresas-usa-redes-sociais-para-promocoes-e-vendas.htm> Acesso em 08 maio 2011

O´REILLY, T. **What Is Web 2.0**: Design Patterns and Business Models for the Next Generation of Software. Disponível em: <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>. Acesso em: 17 jun. 2011.

PADOVANI, C.R. **Estatística na Metodologia da Investigação Científica.** Botucatu: UNESP, 1995.

Pang B, Lee L, **Opinion Mining and Sentiment Analysis.** Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008.

PEREIRA, J.C.R. **Análise de Dados Qualitativos.** São Paulo: Edusp/Fapesp, 1999.

PESSOAS<http://pessoas.hsw.uol.com.br/baby-boomers.htm> Acesso em 15 abr. 2011

REVISTA PEGN<http://revistapegn.globo.com/Revista/Common/0,,EMI124097-17171,00AS+VANTAGENS+DO+USO+DE+REDES+SOCIAIS+NAS+EMPRESAS.html> Acesso em 08 maio 2011

R-PROJECT < http://www.r-project.org/> Acesso em 23 out. 2011

SERRANO, D. P. **Geração X** 17/04/2011 Disponível em: <http://www.portaldomarketing.com.br/Artigos3/Geracao_X.htm> Acesso em 17 abr. 2011

SERRANO, D. P. **Geração Z** 17/04/2011 Disponível em: <http://www.portaldomarketing.com.br/Artigos3/Geracao_Z.htm> Acesso em 17 abr. 2011

SEVERINO, A. J. **Metodologia do trabalho científico**. 22 ed. São Paulo: Cortez, 2002.

SOUSA, L. M. M. de, AZEVEDO, L. E. "O Uso de Mídias Sociais nas Empresas: Adequação para Cultura, Identidade e Públicos" - **IX Congresso de Ciências da Comunicação na Região Norte** – Rio Branco – AC – 27 a 29 de maio 2010. Disponível em: <http://www.intercom.org.br/papers/regionais/norte2010/resumos/R22-0015-1.pdf> Acessado em 15 maio 2011

TAPSCOTT, D. **A hora da geração digital**: como os jovens que cresceram usando a Internet estão mudando tudo, das empresas aos governos tradução Marcello Lino – Rio de Janeiro: Agir negócios 2010

TEIXEIRA, J. **O sucesso do editor pop** Disponível em: <http://veja.abril.com.br/260510/sucesso-editor-pop-p-168.shtml> Acessado em 15 out. 2011

VEIGA, S.C.A e SILVA, W,T. **Redes Bayesianas**: uma visão geral. Brasília, 2002. Disponível em: <http://samuelveiga.pro.br/arq/Redes%20Bayesianas%20-%20Uma%20visao%20geral.pdf>. Acesso em: 07 dez. 2011.

VENTICINQUE, Danilo **Todo o poder aos fãs** Disponível em: <http://revistaepoca.globo.com/Revista/Epoca/0,,EMI219478-15220,00.html> Acessado em 15 out. 2011

Wiebe J, Wilson T, Bruce R, Bell M, Martin M, **Learning Subjective Language** Computational Linguistics, vol. 30, pp. 277–308, September 2004