FERRAMENTA DE VISUALIZAÇÃO DE DADOS E PROCESSAMENTO DE TEXTO: ANÁLISE DE REVIEWS DE VIAJANTES NO TRIPADVISOR

Yussif Tadeu de Barcelos[1], Marlusa Gosling[1], Mariana de Freitas Coelho[1], Marcos Paulo Resende[1]
[1]Universidade Federal de Minas Gerais
barcelosyussif@gmail.com, mg.ufmg@gmail.com, marifcoelho@gmail.com, marcospaulodlresende@gmail.com

Resumo
A proposta deste estudo é demonstrar o desenvolvimento e a aplicação de softwares para a análise de dados, de modo a proporcionar novas diretrizes para o uso de grandes quantidades de texto não estruturado no contexto da gestão. Esse projeto teve como base um estudo sobre o engajamento de viajantes com hotéis de um destino turístico através de *reviews online* em *websites* de planejamento de viagens. Os dados foram coletados automaticamente, pré-processados, categorizados e disponibilizados em visualizações interativas e dinâmicas para análises qualitativas e quantitativas dos comentários. Os resultados demonstram como a categorização por temas e a visualização de dados em gráficos podem condensar as informações disponíveis em uma rede social especializada de turismo (TripAdvisor).

Palavras-chave: Text Mining. Visualização de dados. Avaliação de comentários de hotéis. Gerenciamento e Análise de dados da web. Gestão hoteleira.

FERRAMENTA DE VISUALIZAÇÃO DE DADOS E PROCESSAMENTO DE TEXTO: ANÁLISE DE REVIEWS DE VIAJANTES NO TRIPADVISOR

Abstract
The purpose of this study is to demonstrate software development and application for data analysis in order to provide new guidelines for its use in large amounts of unstructured text in management's context. This project was based on a study about traveler engagement with hotels at a tourist destination through online reviews on travel planning websites. Data were automatically collected, pre-processed, categorized and disposed into dynamic and interactive visualizations in order to provide qualitative and quantitative analysis of the reviews. The results demonstrate how the categorization by themes and graphics can condense the information available in a specialized tourism social network (TripAdvisor).

Key-words: *Text Mining. Data Visualization. Hotel Reviews' Assessment. Web Data Management. Hotel Management.*

1. Introdução

A visualização de informação (ou visualização de dados) pode ser entendida como o estudo e a prática de mapear conjuntos de dados e representá-los em formatos visuais para auxiliar os usuários a explorar e compreender esses conjuntos de dados (CARLIS; KONSTAN; 1998). Ao utilizarem interfaces visuais eficientes, as pessoas podem trabalhar com grandes volumes de dados de forma ágil e efetiva, descobrindo características, padrões e tendências que poderiam passar despercebidas à primeira vista (KIRNER *et al.*, 2004). Ainda, de acordo com Galvão e Marin (2009), durante a mineração de dados, definem-se as tarefas em atividades preditivas e descritivas com o objetivo de obter uma resposta para um problema.

Uma das áreas que pode utilizar os dados de maneira estratégica é o marketing, isto é, "a área do conhecimento que engloba todas as atividades concernentes às relações de troca, orientadas para a satisfação dos desejos e necessidades dos consumidores, visando alcançar determinados objetivos de empresas ou indivíduos e considerando sempre o meio ambiente de atuação e o impacto que essas relações causam no bem-estar da sociedade" (LAS CASAS, 1997, p. 27).

As atividades de um gestor de marketing abrangem um leque muito amplo, desde o estudo de mercado, a fixação de preço, a definição de estratégias, promoção, distribuição, vendas e a assistência pós-venda. Além disso, o crescente uso de ferramentas na internet para se comentar sobre marcas, produtos e experiências tem dificultado o processo de gestão de marcas e sua reputação. Especificamente no setor turístico, o impacto do uso da mídia social se relaciona à descentralização de uma mídia que era centrada na organização (mídia própria) para outra na qual a participação do consumidor influencia outros consumidores (mídia adquirida) (ELDEMAN, 2010). Consequentemente, empresas detêm menor controle sobre sua mídia adquirida, a citar canais criados por consumidores e comunidades de entusiastas de uma marca, serviço ou produto.

A necessidade de cruzar informações para executar uma gestão empresarial eficiente é uma realidade crescente. As empresas têm procurado ferramentas para análises estratégicas de seus dados, principalmente aquelas que se encontram em áreas altamente competitivas e que investem em gestão da informação e do conhecimento como diferencial competitivo.

Este trabalho explora o estudo abordado no artigo "Comentar Bem ou Mal na Internet? O Engajamento de Viajantes em Reviews de Hotéis" (COELHO; GOSLING, 2013) e apresenta uma solução computacional para auxiliar os usuários do TripAdvisor e gestores de hotéis na análise de satisfação dos consumidores, além de contribuir para a avaliação da reputação dos hotéis e dos termos e macrotemas (temas gerais) mais mencionados pelos consumidores.

Assim como no artigo citado, este estudo considera os mesmos hotéis em Ouro Preto, Minas Gerais, Brasil: Solar do Rosário, Pousada Arcanjo, Luxor, Pousada Clássica, Boroni Palaci Hotel e Grande Hotel Ouro Preto. Foram analisados mais de 250 *reviews,* publicados até 30 de junho de 2013, do TripAdvisor, enquanto o artigo analisou 606 comentários do TripAdvisor e Booking.com.

Entretanto, Coelho e Gosling (2013) coletaram e avaliaram os comentários de forma manual, com o auxílio de algumas ferramentas, como o Excel, para facilitar o processo. Já este artigo propõe o desenvolvimento e aplicação de softwares para coleta e análise dos dados não estruturados de forma automática e semi-automática, incorporando visualizações de dados para análises qualitativas e quantitativas no contexto da gestão. Em outras palavras, o objetivo do artigo é demonstrar como reduzir grandes quantidades de texto disponíveis na *web* para propiciar análises e visualização dos dados de maneira mais rápida e confiável, no contexto específico de comentários de hotéis do TripAdvisor. Ressalta-se também a característica interdisciplinar deste trabalho, que aborda temas correlatos das áreas de ciência da computação, administração, marketing e turismo.

A implementação da solução computacional proposta no trabalho envolveu as etapas de coleta e pré-processamento de *text mining* (mineração de texto); armazenamento de informações em um banco de dados; a categorização e a classificação de termos; o processamento dos textos dos comentários de avaliações de hotéis aplicando índice dos termos categorizados; o processamento e a contagem dos termos agrupando-os como avaliações e macrotemas e; finalmente, a disponibilização dos dados em visualizações web utilizando a biblioteca D3js (disponível em http://d3js.org).

Na próxima seção, apresenta-se a metodologia de mineração de texto, seguida pela literatura de mineração de dados e gestão (seção três). A quarta seção apresenta o desenvolvimento do trabalho; na quinta seção é descrita a coleta de dados; na sexta seção são apresentadas as visualizações dos dados e na última seção estão os resultados e considerações finais do trabalho.

2. Mineração de texto

De acordo com Liddy (2000), mineração de texto é o processo de análise natural de ocorrência de texto com a finalidade de descobrir e capturar informações semânticas para inserção e armazenamento em uma Estrutura de Organização do Conhecimento (*Knowledge Organization Structure - KOS*), com o objetivo de permitir a descoberta de conhecimento, através de acesso textual ou visual, para uso em uma ampla variedade de aplicações.

Hearst (1999, p. 6) define que "*Text Mining*, como análise de dados exploratória, é um método para apoiar pesquisadores e derivar novas e relevantes informações de uma grande coleção de textos." É um processo parcialmente automatizado onde o pesquisador ainda está envolvido, interagindo com o sistema.

Aranha (2007) complementa a definição ao dizer que a mineração de texto é interdisciplinar com influências de áreas como: Processamento de Linguagem Natural (PLN), Recuperação da Informação (RI), Inteligência Artificial (IA) e Ciência Cognitiva.

Liddy (2000, p. 14, *tradução nossa*) define que o processo de mineração de texto consiste em 3 etapas:

- **Text Preparation (*Preparação de Texto*)**: "seleção, limpeza e pré-processamento do texto. Nesta fase ocorre a seleção de sites ou fontes de texto com um pré-processamento inicial, geralmente sob a orientação de um especialista humano ou um algoritmo de software bem treinado, como a identificação de parágrafos/sentenças e classes de palavras".
- **Text Processing (*Processamento de Texto*)**: "uso do algoritmo de *data mining* (mineração de dados) para processar os dados preparados, comprimindo e transformando esses dados para identificar pepitas latentes de informação. Nesta etapa um sistema de PLN inteiramente caracterizado determina identidades canônicas e variante de entidades (pessoas, empresas, organizações, etc.), identificando relações conceituais entre entidades e até mesmo instanciando quadros de interesse".
- **Text Analysis (*Análise de Texto*)**: "avaliação da saída para ver se o conhecimento foi descoberto e determinar a sua importância. Após executado os algoritmos de processamento, os textos/dados extraídos são submetidos a várias técnicas que permitirão a utilização direta da informação extraída ou através da visualização em uma ferramenta que permitirá a análise humana para completar a análise iniciada pela ferramenta de mineração de texto".

Aranha (2007) apresenta em sua tese a expansão deste modelo para 5 etapas, descritas a seguir.

- **Coleta de Dados**: nesta etapa ocorre a busca dos dados para formar a base textual de trabalho que irá ser processada. Segundo Aranha (2007), coletar dados é uma atividade trabalhosa, pois os dados podem não estar disponíveis em um formato apropriado para serem utilizados no processo de mineração de textos.
- **Pré-Processamento**: "consiste em um conjunto de transformações realizadas sobre alguma coleção de textos com o objetivo de fazer com que esses passem a ser estruturados em uma representação atributo-valor" (ARANHA, 2007, p. 42). Em mineração de textos, pré-processamento normalmente significa dividir o texto em palavras, aplicar técnicas de *stemming*, remover as *stop-words* e classificá-las segundo a classe gramatical. Aranha (2007, p.45-46) relata que, segundo Spark- Jones e Willett (1997), esta etapa inclui:

> *A eliminação de palavras comuns: as palavras comuns (stop-words) são elementos de texto que não possuem uma semântica significativa; sua presença não agrega nenhuma indicação do conteúdo ou do assunto do texto correspondente. Normalmente as*

palavras comuns são constituídas de artigos, preposições, verbos auxiliares, etc, tais como 'que', 'de/do/das', 'o' ou 'a'.
A obtenção dos radicais (stems): em linguagem natural, diversas palavras que designam variações indicando plural, flexões verbais ou variantes são sintaticamente similares entre si. Por exemplo, as palavras 'real', 'realidade', 'realeza' e 'realizado' têm sua semântica relacionada. O objetivo é a obtenção de um elemento único que permita considerar como um único termo, portanto com uma semântica única, estes elementos de texto.

- **Indexação**: conforme explicado por Aranha (2007), indexação é o processo de catalogar documentos, segundo um critério, para que estes documentos sejam recuperados de forma mais rápida e precisa.
- **Mineração**: para Aranha (2007) esta fase envolve decidir e aplicar algoritmos aos dados para criar um modelo preditivo, podendo-se utilizar diversas áreas do conhecimento como Estatística, Aprendizado de Máquina e Redes Neurais.
- **Análise da Informação**: Aranha (2007) e Liddy (2000) apresentam definições muito semelhantes para esta fase, enfatizando a participação humana na análise da informação aplicabilidade e consistência dos resultados.

Em Salton (1983), a identificação das palavras nos documentos a serem indexados consiste na identificação de palavras analisando-se as sequências de caracteres no texto. Conforme apontado por Aranha (2007), Salton aconselha fazer um *Dictionary lookup*, ou seja, comparar as sequências de caracteres retiradas do texto com um dicionário a fim de validar se essas palavras realmente existem. Pode-se dizer que a abordagem escolhida para este estudo foi de mineração de padrões, já que buscou encontrar padrões existentes nos dados através de regras de associação, e também o processamento simplificado de linguagem natural, utilizando o texto para descobrir quem fez o quê, quando, como, onde e porquê (PRATES; BARBON JR., 2013).

Aranha (2007) observa que o dicionário pode também auxiliar a identificação de termos específicos, quando se deseja utilizar palavras pré-definidas no índice, evitando que palavras desconhecidas sejam identificadas (ou seja, evita a utilização de um vocabulário descontrolado).

3. Mineração de texto e Gestão

A aplicação de técnicas de mineração de dados textual para a descoberta de conhecimento útil e a classificação de dados textuais é uma área relativamente nova de investigação (UR-RAHMAN; HARDING, 2012). As minerações de dados e de texto permitem a extração de padrões, a partir de dados textuais, como por exemplo, a mensuração da produção e distribuição geográfica dos textos, o levantamento das palavras mais usadas e a avaliação do cumprimento do papel da empresa. Porém, para que a mineração de texto tenha um desempenho estratégico na gestão do conhecimento, é preciso uma interlocução entre a coleta de dados, a interpretação do significado dos dados coletados e, por fim, a geração de aprendizado e a tomada de ações após a análise da informação (SILVA; PRADO; FERNEDA, 2002).

Prates e Barbon (2013) afirmam que o crescimento da internet e das redes sociais tem oferecido informações que contribuem para a venda e aceitação de seus produtos por meio de análises dos perfis dos consumidores e aquisição de conteúdos significativos.

Conforme Benevenuto, Almeida e Silva (2011), cada pesquisa demanda um tipo específico de coleta, então, cada coleta de dados apresenta particularidades. Dentre os possíveis pontos de coleta de dados sobre redes sociais *online* destacados pelos autores estão: (1) metodologias de entrevistas com usuários de redes sociais, (2) proxies ou agregadores de tráfico (dados que passam por um provedor de serviços da internet ou de um agregador de redes sociais); (3) dados de servidores (obtidos diretamente de servidores de redes sociais *online*) ou coleta de dados públicos na web e; (4) dados de aplicações de terceiros (páginas de redes sociais com o uso de

uma ferramenta automática, denominada *crawler* ou robô, que coleta sistematicamente informações públicas de usuários e objetos) (BENEVENUTO, ALMEIDA E SILVA, 2011). Este trabalho coletou dados disponíveis publicamente na *web*, como citado anteriormente.

Escolheu-se a área do turismo, especificamente os dados de *reviews* de hotéis, dado à importância das informações disponibilizadas online para decisões de viagem. As novas gerações de turistas planejam suas viagens e buscam por destinos através de serviços *online* e redes sociais, tornando popular a exploração de tais serviços através de *travel social networks*, nas quais usuários compartilham suas opiniões sobre pontos turísticos famosos e sobre lugares desconhecidos (LEITE, BENEVENUTO E MORO, 2013). Ainda segundo os autores, ferramentas online de turismo oferecem informações recentes e atualizadas, com opiniões que são percebidas como mais confiáveis pelos turistas potenciais e há a possibilidade de identificar o que várias pessoas estão pensando e qual o sentimento delas sobre o lugar, ao invés de ter apenas a opinião de agentes de viagens e autores de livros ou guias turísticos.

No contexto hoteleiro, os sites de avaliações de hotéis na internet podem influenciar a escolha e reserva de hotéis pelos consumidores e serem úteis como ferramenta de apoio à decisão dos gestores (GOMES, CHAVES e PEDRON, 2011). O artigo de Leite, Benevenuto e Moro (2013) apresentou um estudo similar ao proposto por esta pesquisa, porém os autores tinham como objetivo desenvolver uma ferramenta para auxiliar turistas na busca de pontos turísticos. Assim, nenhum trabalho foi encontrado nas bases de pesquisa EBSCO que coletou informações do TripAdvisor ou sites similares de avaliação de hotéis e desenvolveu uma ferramenta de visualização e avaliação de dados destinados à gestão estratégica dos mesmos por gestores de hotéis. Dentre os resultados do estudo de Leite, Benevenuto e Moro (2013), tem-se que avaliações *online* têm potencial para auxiliar a gestão dos hotéis estudados, no aumento da satisfação dos seus clientes e na gestão dos recursos existentes de forma eficiente.

Portanto, tornar um conteúdo extenso em algo relevante para a decisão empresarial e para a geração de confiança pelos consumidores ainda é um desafio. Para Costa *et al.* (2012), a incorporação de dados não estruturados para um ambiente de tomada de decisão representa um desafio de negócios, já que as técnicas e tecnologias de *Business Intelligence* ainda representam um número muito pequeno da porcentagem de dados corporativos. Esta pesquisa aponta uma lacuna sobre o estudo do desenvolvimento de ferramentas para utilização empresarial, em especial no contexto turístico. Por este motivo, são demonstrados em detalhe os métodos utilizados para o processo de mineração de dados com base na literatura e o potencial existente da mesma para a gestão hoteleira, por meio dos relatórios com novas abordagens de visualização de dados.

4. Desenvolvimento do Trabalho

O trabalho foi desenvolvido com o intuito de propiciar uma experiência visual analítica amigável e interativa para os usuários, levando em consideração principalmente os conceitos e necessidades abordados por Coelho e Gosling (2013), além de facilitar as etapas de coleta e análise dos dados através de aplicativos com rotinas automáticas ou semiautomáticas com a intervenção do usuário.

As visualizações foram desenvolvidas utilizando a biblioteca D3js, html, css e javascript. O *site* oficial da biblioteca D3js explica que:

> *D3js (Data-Driven Document) é uma biblioteca JavaScript para manipular documentos com base em dados. O D3 ajuda a trazer os dados à vida usando HTML, SVG e CSS. A ênfase do D3 em padrões web dá-lhe todas as capacidades de navegadores modernos sem amarrar-se a uma estrutura de propriedade, combinando componentes de*

visualização potentes e uma abordagem orientada a dados para manipulação DOM[1] *(BOSTOCK, 2015).*

Neste trabalho não foi implementada a etapa de mineração e as *stop-words* (palavras comuns) foram mantidas no dicionário de termos para o pré-processamento e identificação de termos compostos.

4.1. Definições Gerais

Com base em Liddy (2000), foi necessário efetuar a preparação de texto, verificando-se os termos pertinentes para o processamento e a análise de texto no contexto da gestão. Algumas expressões e classificações foram utilizadas para detalhar o desenvolvimento do trabalho. Assim, é importante esclarecer as seguintes definições gerais:

- **Termo:** palavras, conjunto de palavras e caracteres presentes nos comentários
- **Termo Simples:** contém apenas uma palavra ou caractere, por exemplo *casa , "!", "-", comida*
- **Termo Composto:** contém duas ou mais palavras para formar um termo ou expressões significativas, por exemplo *café da manhã, café-da-manhã, comida mineira, piscina aquecida, o hotel estava cheio*
- **Dicionário de Termos:** contém o índice dos termos e classificação dos termos em categorias, macrotema, positivo e negativo. Este dicionário será a base para o pré-processamento, indexação e geração de dados para as visualizações
- **Categoria de Termos:** classificação do termo de acordo com sua função e tipos de interpretação nos textos de comentários
  Não identificado: termos ainda não classificados pelo usuário
  Auxiliares: são as *stop-words*, palavras que não possuem relevância na análise de termos simples, como exemplo *de para é sim não manhã talvez*
  Pontuação: sinais de pontuação de texto, como por exemplo *". ", "!", "?", "-", ":", ")"*
  Sentimentos: verbos e adjetivos de caráter positivo ou negativo que expressão reações e comportamentos sentimentais dos clientes, como exemplo *gostei detestei triste feliz*
  Avaliação: adjetivos e advérbios de caráter positivo ou negativo que expressam a análise ou satisfação do cliente em relação aos hotéis e serviços, como exemplo *quente frio bom ruim agradável amigável excelente péssimo*
  Característica: adjetivos e advérbios de caráter positivo ou negativo que expressam características dos hotéis ou serviços, como exemplo *antigo histórico decorado barulhento espaçoso*
  Tema: classificação temática dos termos relacionados aos hotéis sinalizados no artigo base deste trabalho, como *quarto atendimento gastronomia preço cultura*
- **Macrotemas:** classificação temática dos termos relacionados aos hotéis que serão utilizados para as análises
  Atendimento: funcionários, atendimento, recepção, colaboradores
  Cultura: igreja, histórico
  Decoração: bem decorado
  Gastronomia: café-da-manhã, almoço, restaurante
  Instalações: elevadores, salas, piscina
  Localização: próximo, longe, localização
  Preço: caro, barato

[1] *Document Object Model* (Modelo de Objeto de Documento) é uma interface multiplataforma e independente de linguagem que permite a programas e *scripts* acessarem e atualizarem o conteúdo, estrutura e estilo em documentos HTML, XHTML e XML.

Quarto: banheiro, cama, quarto, chuveiro

- **Rotina:** aplicativo, software, programa que executa automaticamente uma tarefa cuja intervenção do usuário está limitada a iniciar a execução da rotina através do clique de um botão ou menu, execução de um arquivo .exe ou .bat ou agendamento de *job*
- **Administrador**: usuário Administrador do sistema responsável por manter (cadastrar, editar e excluir) categorias, macrotemas e termos e iniciar o processamento dos dados
- **Sistema:** aplicativo disponibilizado para os Administradores. O aplicativo contém telas para pesquisa e manutenção de categoria, macrotemas, termos e para o processamento, todos acessíveis através do menu superior. O aplicativo foi nomeado VisMkt.

4.2. Coleta dos dados

Como etapa ressaltada por Liddy (2000) e Aranha (2007), a coleta de dados é realizada automaticamente através de uma rotina que acessa a url (endereço do site) no TripAdvisor do hotel para cada paginação dos comentários, percorre o texto do arquivo HTML extraído da url em busca de padrões de textos que identificam os dados do usuário e do comentário. Na Figura 1 pode-se observar um trecho do HTML do comentário registrado para o hotel Solar do Rosário.

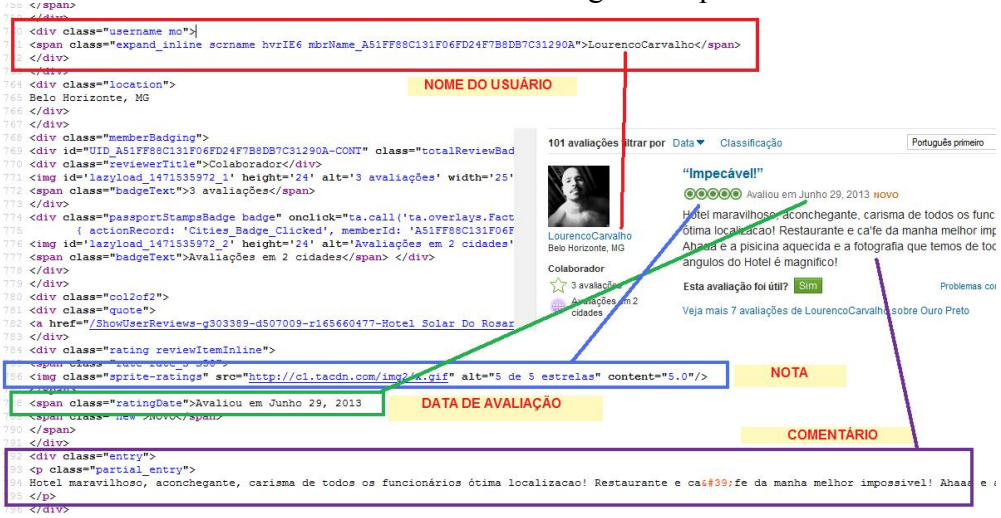


Figura 1 – HTML do Hotel Solar do Rosário

Os seguintes dados foram coletados do html extraído:

- Usuário: Nome, Localização (cidade, estado, país)
- Comentário: Data, Nota (avaliação), Conteúdo (texto do comentário).

Para executar esta rotina, o Administrador informa os campos do código do hotel, nome do arquivo de dados para conferência, número de páginas de *reviews* do hotel e url da página do hotel no TripAdvisor.

Ao executar a rotina "Carregar Review On-Line", os dados dos *reviews* são carregados no banco de dados e uma cópia dos dados fica disponível para o Administrador conferir a coerência dos dados.

4.3. Limitações

O *site* do TripAdvisor limita o número de palavras que são exibidas na visão geral dos conteúdos dos comentários. Para os usuários verem o restante do texto do comentário, o usuário deve clicar na opção "MAIS", localizada sob o conteúdo do comentário. Ao clicar neste botão é apresentado ao usuário o restante do texto do comentário e as avaliações dos macrotemas definidos pelo TripAdvisor (Custo-Benefício, Localização, Qualidade do Solo, Quartos, Limpeza, Serviço). Essa necessidade de clicar no "MAIS" elimina a captura automática do restante do conteúdo do comentário e a avaliação dos macrotemas pela rotina de carga. Também conforme Coelho e

Gosling (2013) tal limitação também pode interferir na escolha do hotel pelos potenciais viajantes, uma vez que o volume de informação é tão grande que pode limitar a leitura e entendimento dos consumidores.

Outra limitação refere-se aos comentários que não estão em Português: estes comentários utilizam um serviço do Google Tradutor para traduzir o conteúdo do comentário. Essa limitação elimina a captura automática do conteúdo do comentário pela rotina de carga.

4.4. Cadastro de Termos

Ao realizar o cadastro ou edição de um termo simples ou composto (Figura 2), o Administrador classificará o termo em uma categoria. Caso opte por uma categoria que seja do tipo "Tema" as opções "MacroTema" e "Tema" são habilitadas. Caso opte por uma categoria do tipo "Avaliação" a opção de "Avaliação: Positivo/Negativo" é habilitada.

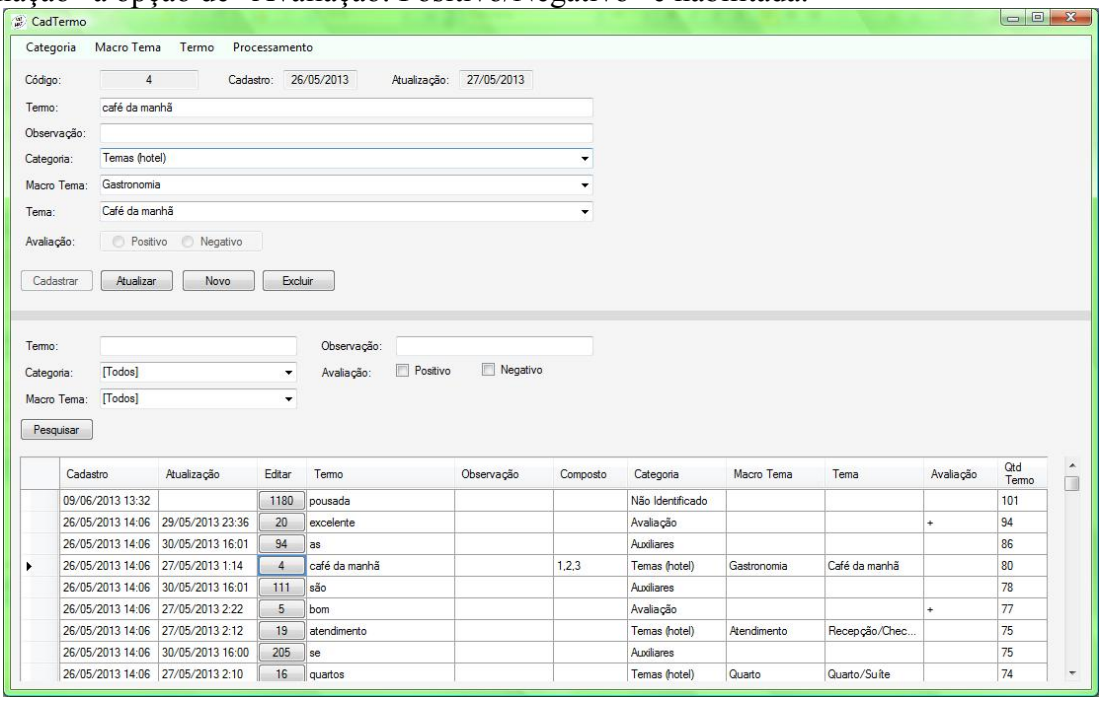


Figura 2 – Cadastro de Termos
Fonte: tela de cadastro do software VisMkt

Ao cadastrar/editar um termo composto o seguinte fluxo é realizado:

1. sistema identifica os termos individuais;
2. sistema verifica se estes termos não estão cadastrados;
3. sistema cadastra novos termos individuais;
4. sistema recupera os códigos dos termos individuais na ordem do termo composto;
5. sistema cadastra novo termo composto e armazena na tabela de termos (dicionário de termos) a sequência de códigos dos termos individuais separados por vírgula.

O Administrador pesquisa os termos que deseja editar ou excluir na listagem localizada na parte inferior da tela de cadastro e pode ordenar pelo atributo (coluna) desejado. A listagem apresenta as informações (atributos) do termo e uma contagem do número de ocorrências do termo em questão. Assim, o Administrador poderá realizar um filtro pelas características do termo e por termos mais frequentes. Essa contagem de termos é interessante para o usuário identificar os termos mais frequentes e assim classificar e analisar os mesmos.

4.5. Pré-processamento, indexação simples e análise de termos de comentários

Nesta seção são detalhados os impactos de termos compostos, os processos de identificação de termos, indexação dos termos nos comentários e análise dos termos.

Após executar a coleta de dados, o Administrador deve executar o próximo passo com a rotina de processar *reviews* on-line. Esta rotina percorre os comentários cujos termos ainda não foram indexados. O seguinte fluxo é executado:

1. sistema consulta os comentários para indexação e percorre cada comentário retornado;
2. sistema busca pontuações no comentário e adiciona espaço antes e depois da pontuação (dessa forma a pontuação não ficará no termo antecessor à ela);
3. sistema realiza um *split* (quebra) por espaços em branco no texto do comentário;
4. sistema percorre cada termo retornado da quebra removendo os espaços extras:
    a. sistema envia o termo para o banco de dados;
    b. banco de dados pesquisa o termo e cadastra o termo caso ele não exista;
    c. banco de dados retorna o código do termo (seja ele novo ou já existente);
    d. sistema concatena o código à lista de códigos (índices) de termos.
5. sistema cadastra a análise do comentário, incluindo a lista de códigos (índices) dos termos, separando os códigos por vírgula.

Após executar o passo anterior, o Administrador deve executar a etapa de processar termos compostos para análise dos comentários que não possuem termos compostos analisados. O seguinte fluxo é executado:

1. sistema consulta os comentários para análises e percorre cada comentário retornado;
2. sistema busca sequências de termos compostos nos índices do comentário;
3. sistema percorre cada termo composto retornado:
    a. sistema procura o termo mais relevante;
    b. sistema substitui na indexação do comentário a sequência do termo composto pelo código do termo composto (C#), identificando assim que o termo é composto.
4. sistema encerra o processamento ao término do último comentário;

COD_COMENTARIO	DTA_ANALISE	LST_TERMO_COMENTARIO	LST_TERMO_ANALISE
70	30/06/2013 15:...	43,250,7,701,7,2263,32,715,106,39,337,14,10,146,41,...	43,250,7,701,7,2263,32,715,106,39,337,14,10,1...
71	30/06/2013 15:...	20,43,7,39,98,719,7,84,546,584,32,728,2272,145,2273...	20,43,C142,98,719,7,84,546,584,32,728,2272,1...
72	30/06/2013 15:...	2279,21,87,88,89,1247,2280,12,945,41,208,2281,2282...	2279,21,87,88,89,1247,2280,12,945,41,208,228...
73	30/06/2013 15:...	250,43,7,14,2294,7,19,2295,7,16,149,7,95,73,20,109,...	250,43,7,14,2294,7,19,2295,7,16,149,7,95,73,2...
74	30/06/2013 15:...	751,299,498,2304,2305,80,87,88,89,32,441,442,7,51,7...	751,299,498,2304,2305,80,87,88,89,32,441,442,...
75	30/06/2013 15:...	64,51,97,2316,113,216,81,447,339,12,51,19,2317,7,41...	64,51,97,2316,113,216,81,447,339,12,51,19,231...
76	30/06/2013 15:...	1116,106,16,84,111,1252,7,29,21,43,2329,9,41,90,118...	1116,106,16,84,111,1252,C180,21,43,2329,9,41...
77	30/06/2013 15:...	67,33,68,69,70,71,72,73,74,75,76,77,7,78,79,80,81,41...	67,33,68,69,70,71,72,73,C1155,76,77,7,78,79,8...
78	30/06/2013 15:...	43,90,141,7,39,143,2,144,145,146,10,106,16,111,1147...	43,C1121,C142,143,2,144,145,146,10,106,16,11...
79	30/06/2013 15:...	193,32,194,90,1147,195,7,21,43,41,196,197,10,10,63	193,32,194,90,C1146,7,21,43,41,196,197,10,10

Figura 3 – Exemplos de comentários indexados e analisados
Fonte: registros em banco de dados do software VisMkt

4.6. Interpretação de termos compostos

O objetivo dos termos compostos é unir termos que isoladamente não são interessantes para a análise e interpretação, mas que são significativos quando estão juntos. Caso seja encontrado no processamento de análise de termos uma lista de códigos de termos que seja uma sublista de mais de um termo composto, será considerado como mais relevante o termo composto que possui mais termos individuais. Por exemplo, pode-se analisar a expressão "hotel não é ruim".

No sistema, estão cadastrados os seguintes termos compostos:

- hotel não é ruim;
- não é ruim;
- é ruim.

Após percorrer a indexação de termos do comentário o sistema identifica que existem 3 termos compostos para esse intervalo de 4 termos individuais.

1. sistema ordena a listagem de termos compostos pelo número de termos individuais;
2. sistema substitui a sequência de caracteres da indexação pelo código do maior termo composto acrescido da letra C (ex. C10), identificando que o código refere-se a um termo composto;
3. sistema procura próxima lista de termos compostos e repete o processo até que não exista mais sequências de termos na indexação que são sublistas de termos compostos.

Adotou-se essa heurística de aplicar o maior termo composto por considerar que quanto mais termos individuais o termo composto possuir, mais precisa é a interpretação do termo.

No exemplo, percebe-se que o usuário do TripAdvisor fez uma avaliação não negativa do hotel. Observe:

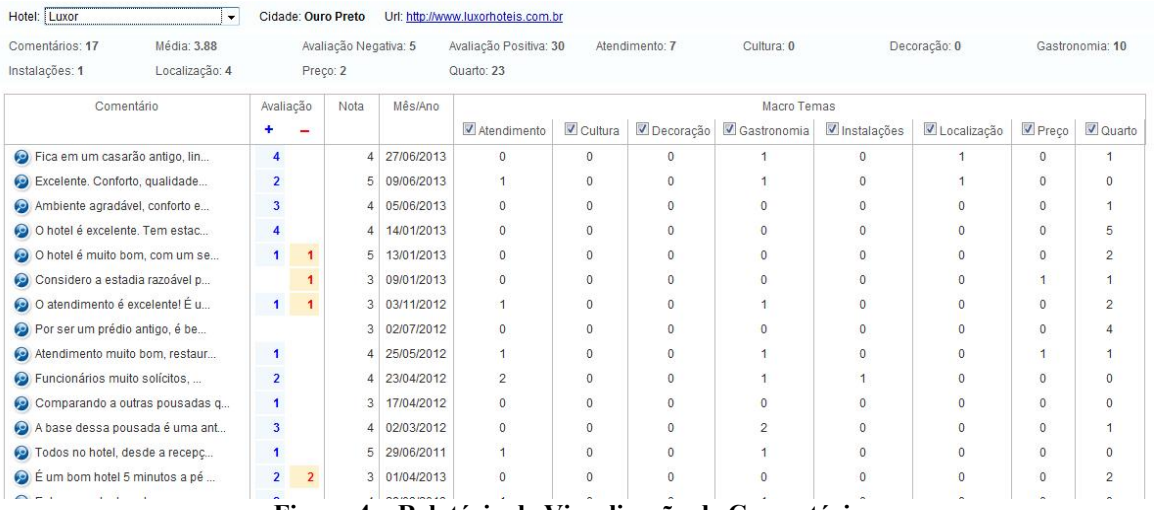
- os termos isolados (hotel, não, é, ruim) não possuem uma significância forte;
- o termo composto é ruim retorna uma avaliação negativa do usuário;
- o termo composto não é ruim retorna uma avaliação mais positiva que negativa;
- a composição hotel não é ruim é mais relevante para as análises de usuários do TripAdvisor e para os gestores dos hotéis.

5. Exibição de informações em visualizações

Para a visualização dos dados proposta nesse trabalho foram elaborados 4 relatórios, especificados a seguir, que visam contribuir para última etapa da mineração de texto, a análise de texto (LIDDY, 2000; ARANHA, 2007) .

5.1. Visualização de comentários

Após selecionar o hotel desejado, a visualização exibe uma tabela com a contagem de termos negativos e positivos (coluna avaliações), a contagem de termos de macrotemas (respectiva coluna), a nota atribuída, a data do comentário e os 100 primeiros caracteres do comentário. No cabeçalho da visualização são exibidos os totalizadores de comentários, das avaliações, dos macrotemas e a nota média dos comentários.


Hotel: Luxor Cidade: Ouro Preto Url: http://www.luxorhoteis.com.br

Comentários: 17 Média: 3.88 Avaliação Negativa: 5 Avaliação Positiva: 30 Atendimento: 7 Cultura: 0 Decoração: 0 Gastronomia: 10
Instalações: 1 Localização: 4 Preço: 2 Quarto: 23

Comentário	Avaliação +	Avaliação −	Nota	Mês/Ano	Macro Temas: Atendimento	Cultura	Decoração	Gastronomia	Instalações	Localização	Preço	Quarto
Fica em um casarão antigo, lin...	4		4	27/06/2013	0	0	0	1	0	1	0	1
Excelente. Conforto, qualidade...	2		5	09/06/2013	1	0	0	1	0	1	0	0
Ambiente agradável, conforto e...	3		4	05/06/2013	0	0	0	0	0	0	0	1
O hotel é excelente. Tem estac...	4		4	14/01/2013	0	0	0	0	0	0	0	5
O hotel é muito bom, com um se...	1	1	5	13/01/2013	0	0	0	0	0	0	0	2
Considero a estadia razoável p...		1	3	09/01/2013	0	0	0	0	0	0	1	1
O atendimento é excelente! É u...	1	1	3	03/11/2012	1	0	0	1	0	0	0	2
Por ser um prédio antigo, é be...			3	02/07/2012	0	0	0	0	0	0	0	4
Atendimento muito bom, restaur...	1		4	25/05/2012	1	0	0	1	0	0	1	1
Funcionários muito solícitos, ...	2		4	23/04/2012	2	0	0	1	1	0	0	0
Comparando a outras pousadas q...	1		3	17/04/2012	0	0	0	0	0	0	0	0
A base dessa pousada é uma ant...	3		4	02/03/2012	0	0	0	2	0	0	0	1
Todos no hotel, desde a recepç...	1		5	29/06/2011	1	0	0	1	0	0	0	0
É um bom hotel 5 minutos a pé ...	2	2	3	01/04/2013	0	0	0	0	0	0	0	2

Figura 4 – Relatório de Visualização de Comentários
Fonte: tela de visualizações do software VisMkt

A coluna de comentário possui uma imagem de uma lupa (detalhes) que, ao posicionar o mouse sobre a mesma, exibe uma caixa com o texto identificando por cores os macrotemas e termos positivos e negativos (Figura 5). Dessa forma o usuário consegue ter uma visão do texto destacando o que foi considerado como importante. Também é exibido nos detalhes informações do usuário (nome e localização) e do comentário (data e nota).

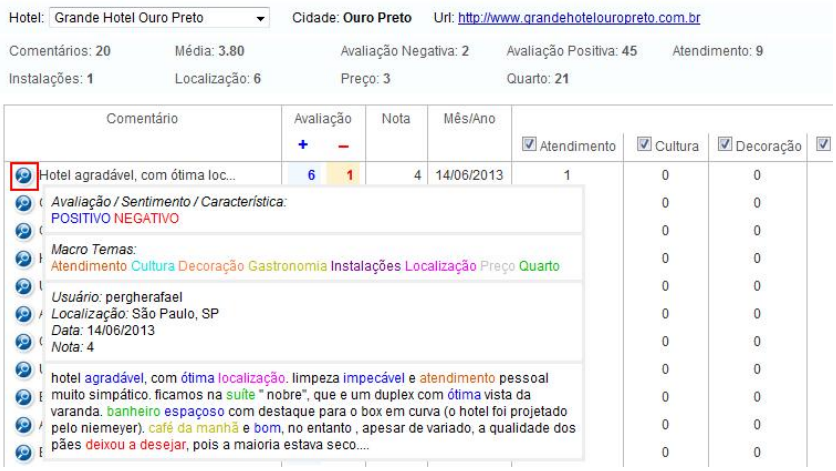


Figura 5 – Detalhes de Comentários em Relatório de Visualização de Comentários
Fonte: tela de visualizações do software VisMkt

Nas colunas da tabela de comentários existe um "*checkbox*" ao lado do nome do macrotema para que o usuário possa marcar se vai destacar esse macrotema ou não no detalhe do comentário. Isso é interessante para que o comentário não fique exageradamente colorido e o usuário possa definir melhor o que deseja destacar. Vale lembrar que a coluna "avaliação" pode contribuir para a identificação dos pontos fracos e fortes do hotel, contribuindo para ações futuras dos gestores, como planejamento estratégico, gestão do conhecimento, recuperação de falhas e aumento na satisfação dos clientes.

5.2. Visualização de macrotemas em bolhas

Os macrotemas são representados pela área da bolha e sua identificação está no texto no centro da bolha e no texto ao posicionar o mouse sobre a bolha. Os hotéis são categorizados por cores e essas cores são percebidas na visualização.

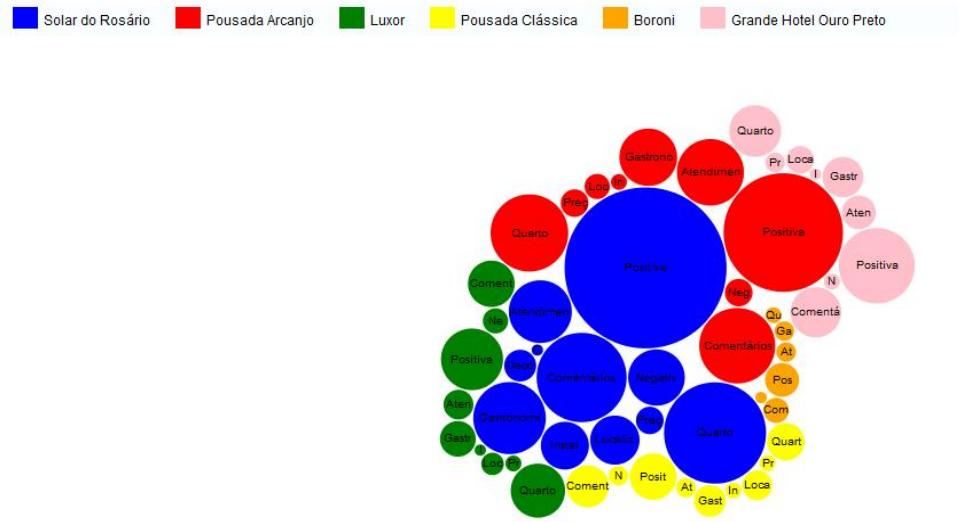


Figura 6 – Visualização em Bolhas de Macro Temas por Hotel
Fonte: tela de visualizações do software VisMkt

A visualização distribui os macrotemas por hotéis com maior volume de termos em macrotemas na região central e na periferia/borda estão os hotéis com menos termos. Nesta visualização é possível ter uma percepção espacial e da representatividade dos comentários por

macrotema e por empresa, o que pode ser de grande valia para a análise da concorrência da competitividade hoteleira.

5.3. Visualização de macrotemas em barras

Diferente da visualização em bolhas, a visualização em barras permite ao usuário realizar uma avaliação quantitativa mais clara do número de temos por macro tema e empresa e, assim, permite a comparação entre os macrotemas e as empresas analisadas. As cores dos macrotemas nesta visualização são as cores cadastradas no sistema e aplicadas na visualização de comentários.

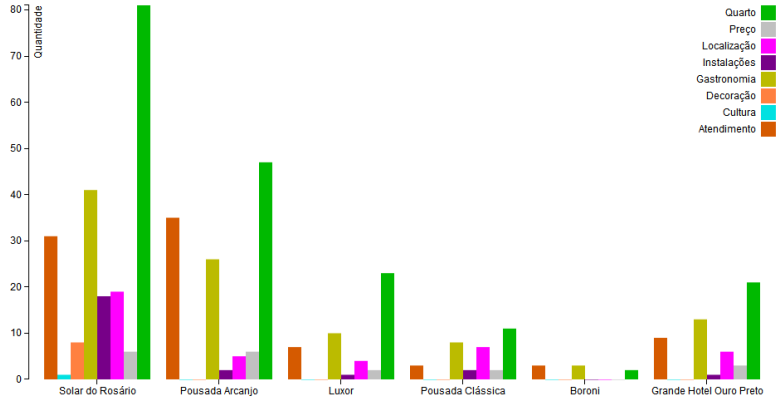


Figura 7 – Visualização em Barras de Macrotemas por Hotel
Fonte: tela de visualizações do software VisMkt

O resultado é similar ao de Coelho e Gosling (2013) que encontrou que as avaliações escritas sobre os hotéis destacaram características do quarto, localização, qualidade da gastronomia (café da manhã e/ou restaurante), qualidade do atendimento/ funcionários, preço, limpeza, decoração, além do barulho percebido dentro das instalações do hotel. Porém, o método utilizado neste artigo permite a visualização por hotel, contribuindo ainda mais para avaliações no contexto mercadológico.

5.4. Visualização mapa *mundi*

Por meio da "Visualização Mapa *Mundi*" é possível ter um panorama global da origem dos clientes dos hotéis. Após realizar a filtragem do hotel, são destacados os países dos clientes que comentaram no TripAdvisor. O nome do país é exibido ao posicionar o mouse sobre o mesmo.

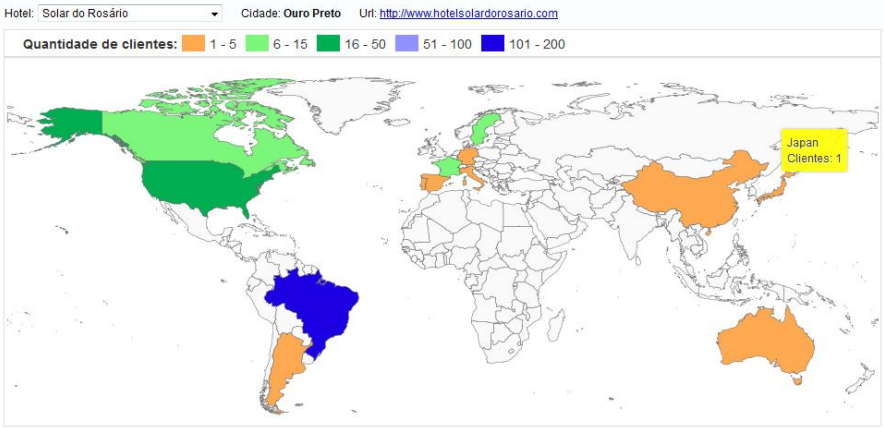


Figura 8 – Visualização Mapa Mundi
Fonte: tela de visualizações do software VisMkt

Dessa forma, a visualização pode contribuir para se realizar análises sobre os hóspedes e traçar estratégias de planejamento, como identificar necessidade de ações voltadas ao treinamento da equipe em idiomas estrangeiros e a elaboração de planos para adequação do hotel

aos padrões internacionais. Ressalta-se que esta análise não foi efetuada por Coelho e Gosling (2013), de modo que a gestão dessa informação também se mostra valiosa para a gestão hoteleira. A Figura 8, por exemplo, demonstra que a maior parte dos visitantes do Hotel Solar do Rosário que avaliam o hotel na internet é proveniente do Brasil, mas visitantes dos Estados Unidos, e, em menor número, Canadá, França e Suécia também avaliaram o hotel, podendo recomendá-lo a outros turistas internacionais.

Considerações finais

O processo de mineração de dados para a extração de informações relevantes para predizer e correlacionar conhecimento a partir de grandes volumes de dados (GALVÃO; MARIN, 2009) também pode trazer contribuições para a gestão de organizações que são avaliadas na internet, a citar hotéis de destinos turísticos.

As visualizações implementadas (visualização de comentários, visualização de temas em bolhas ou em gráficos, e visualização do mapa mundi) e o processamento dos termos cumpriram com a proposta do trabalho e apresentaram novas formas de se avaliar os dados de *reviews* do site TripAdvisor. Como exemplo, o gráfico de mapas gerado a partir da análise dos dados possui um grande potencial para o entendimento da origem de turistas estrangeiros. Dessa forma, a mineração de textos e dados pode contribuir para se traçar estratégias de planejamento e gestão, e ainda auxiliar na identificação de problemas. Por exemplo, se o hotel identificar reclamações sobre funcionários que não falam idiomas estrangeiros, ações voltadas ao treinamento da equipe e a elaboração de planos para adequação do hotel aos padrões internacionais podem ser tomadas. Além disso, destaca-se a necessidade de desenvolvimento de ferramentas que simplifiquem a informação de *sites* com muito texto e volume de informações, de maneira simples e confiável.

A possibilidade de utilização de ferramentas em áreas e empresas multidisciplinares ressalta a importância de desenvolvimento do tema em estudos futuros, uma vez que o uso das técnicas de mineração de dados e mineração de texto em diversas áreas do conhecimento pode contribuir para monitorar as percepções e consumo de clientes, prevenir fraudes e diminuir riscos, dentre outras (MARCANO AULAR; TALAVERA PEREIRA, 2007). Portanto, o uso da ferramenta proposta no texto atuaria como forma de simplificar e sintetizar um conjunto de dados extensos e que, de outra forma, demandaria mais tempo e trabalho para a avaliação dos dados de maneira segura.

No entanto, a técnica não é livre de limitações como citado no desenvolvimento do texto. Segundo Araújo Júnior e Tarapanoff (2006), apesar da mineração de textos contribuir para o processo de busca e recuperação da informação, a precisão da informação gerada em comparação com a indexação manual sempre merece atenção. Assim, a mineração de textos pode ser empregada como ferramenta complementar no processo de indexação, visando ao aumento do índice de precisão na recuperação da informação, devendo sempre se conhecer as necessidades de informação dos usuários, ou seja, os gestores precisam apontar quais informações são importantes para eles.

As empresas podem, também, colher informações estratégicas da concorrência, do mercado como um todo ou de alguma empresa individual. Outra possibilidade é que os gestores do hotel tenham conhecimento de seus pontos fracos e fortes com aporte dos comentários, para corrigir erros e tomar decisões estratégicas. Deste modo, pesquisas futuras devem avançar no condensamento das informações na *web* e nas diversas possibilidades de síntese com o objetivo de gerar informações relevantes e facilmente analisáveis.

Referências Bibliográficas

ARANHA, C. N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português**. 2007. Tese (Doutorado)- Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, 2007.

ARAUJO JUNIOR, R. H.; TARAPANOFF, K. Precisão no processo de busca e recuperação da informação: uso da mineração de textos. **Ci. Inf.**, Brasília , v. 35, n. 3, Dec. 2006.

BENEVENUTO, F; ALMEIDA, J.; SILVA, A. Explorando Redes Sociais Online: Da Coleta e Análise de Grandes Bases de Dados às Aplicações. **Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**. *(SBRC).* Campo Grande, Brazil. Maio 2011.

BOSTOCK, M. Data-Driven Documents. 2015. Disponível em: <http://d3js.org>. Acesso em: 20 fev. 2015.

CARLIS, J. V.; KONSTAN, J. A. Interactive visualization of serial periodic data. In: **Proceedings of the 11th annual ACM symposium on User interface software and technology**. ACM, 1998. p. 29-38.

COELHO, M. F.; GOSLING, M. Comentar bem ou mal na Internet? O Engajamento de Viajantes em Reviews de Hotéis. In: **EnANPAD** – Encontro da Associação Nacional de Programas de Pós-graduação em Adminsitração. Rio de Janeiro: Anais do Enanpad, 2013.

COSTA, P.R.; SOUZA, F. F.; TIMES, V.; BEVENUTO, F. Towards integrating Online Social Networks and Business Intelligence. In: **Proceedings of the IADIS International Conference on Web Based Communities and Social Media** *(WBC'12).* Lisbon, Portugal, July 2012.

ELDMAN, D. C. Branding in The Digital Age. You're Spending Your Money In All the Wrong Places. **Harvard Business Review**, v.88, n.12, p. 62-69, Dez. 2010.

FRIENDLY, M.; DENIS, D. J. Milestones in the history of thematic cartography, statistical graphics, and data visualization. 2008. Disponível em: <http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/Downloads/Friendly_milestone.pdf> Acesso em: 28 fev. 2015.

GALVÃO, N. D.; MARIN, H. F. Técnica de mineração de dados: uma revisão da literatura. **Acta paul. enferm. [online]**. v. 22, n.5, 2009, p. 686-690.

PEDRON, C. Impacto da Web 2.0 e das avaliações online no apoio à gestão de pequenos e médios hotéis em Portugal: um estudo exploratório. In: A**nais da 11ª Conferência da Associação Portuguesa de Sistemas de Informação** (CAPSI 2011), 2011. Disponível em: <http://repositorio-cientifico.uatlantica.pt/jspui/bitstream/10884/306/1/gomes-capsi-2011.pdf>. Acesso em: 26 fev. 2015.

HEARST, M. A. **Untangling Text Data Mining**. Proceedings of the ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics. College Park: University of Maryland, 1999.

KIRNER, C.; KIRNER, T.; CALONEGO Jr., N.; BUK, C. Uso de realidade aumentada em ambientes virtuais de visualização de dados. In: **VII Symposium on Virtual Reality**, SP. 2004.

LAS CASAS, A. L. **Marketing conceitos, exercícios, casos**. 4° ed. São Paulo: Atlas, 1997.

LEITE, A.; BENEVENUTO, F; MORO, M. TripTag: Ferramenta de planejamento de viagens baseada em experiências de usuários de redes sociais. In: **28o Simpósio Brasileiro de Banco de Dados**, Recife, p. 37, set. 2013.

LIDDY, E. D. **Text Mining**. Bulletin of the the American society for Information Science, October/November, 2000. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/bult.184/full> Acesso em: 15 abr. 2013.

MARCANO AULAR, Y.J.; TALAVERA PEREIRA, R. Minería de datos como soporte a la toma de decisiones empresariales. **Opcion**, v.23, n.52, 2007, p.104-18.

PRATES, N. A.; BARBON JR., S. Text Mining em Redes Sociais para Análise de Marketing. Universidade Federal de Londrina, Departamento de Computação, 2013. Disponível em: <http://www.uel.br/cce/dc/wp-content/uploads/ProjetoTCC-NeanderPrates1.pdf>. Acesso em: 26 fev. 2015.

SALTON, G. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill, 1983.

SILVA, E. M.; PRADO, H. A.; FERNEDA, E. Suporte à criação de inteligência organizacional em uma empresa pública de jornalismo com o uso de mineração de textos. In: **Workshop brasileiro de inteligência competitiva e gestão do conhecimento**, 3., 2002, São Paulo. Anais. Congresso anual da sociedade brasileira de gestão do conhecimento, 1., 2002, São Paulo. Anais.

SPARK-JONES, K. e WILLET, P. **Readings in Information Retrieval**. San Francisco, CA: Morgan Kaufmann, 1997.

UR-RAHMAN, N.; HARDING, J.A. Textual data mining for industrial knowledge management and text classification: A business oriented approach. **Expert Systems with Applications**, v. 39, n. 5, p. 4729–4739, abr. 2012.