

# FORMALIZANDO E EXPLORANDO CONHECIMENTO TÁCITO COM A TECNOLOGIA DE TEXT MINING PARA INTELIGÊNCIA

Stanley Loh<sup>1,2</sup>, Eliseo Reategui<sup>3</sup>, Leandro Krug Wives<sup>1</sup>,  
José Palazzo M. de Oliveira<sup>1</sup>, Maurício Almeida Gameiro<sup>4</sup>

sloh@zaz.com.br  
eliseo@godigital.com.br  
wives@inf.ufrgs.br  
palazzo@inf.ufrgs.br  
magameiro@uol.com.br

1- Programa de Pós-Graduação em Computação (PPGC) 2- Universidade Católica de Pelotas (UCPEL) e  
Instituto de Informática Universidade Luterana do Brasil (ULBRA)  
Universidade Federal do Rio Grande do Sul (UFRGS)

3- GoDigital  
Marketing de Precisão

4- Clínica Psiquiátrica Olivé Leite  
Pelotas, RS

## Abstract

This work presents an approach for exploring knowledge available with people, using Text Mining technology. The knowledge may come from internal collaborators or from customers. To make the knowledge concrete in electronic ways, the approach acquires information through textual documents. Text Mining tools are then used to extract concepts present in the texts. Concepts represent real world events, people, objects, etc., and they help to understand what themes or subjects are referenced by the texts. After the extraction step, data mining tools are used to discover new knowledge through the analysis of concept distributions and relations. The approach is useful to obtain intelligence about the organization, in order to improve products, services, internal processes and the relationship with customers.

Two applications of the approach are discussed in this paper: one for exploring knowledge from customers of a cable television company and other to capture the knowledge used by physicians of a psychiatric hospital. In the first case, information was captured as suggestions of the customers. Results from the Text Mining approach enabled the understanding of how customers saw the offered services and products. In the second application, the approach was applied over medical records about patients, written by physicians. The discovered knowledge was useful to analyze patients' characteristics and to understand how the diagnosis process is made.

Keywords: business intelligence, text mining, tacit knowledge, knowledge discovery

## 1 Introdução

Hoje em dia, os clientes costumam selecionar produtos e serviços analisando a competência das organizações e procurando por algum diferencial. Logo, para que as organizações ganhem e mantenham seus clientes elas precisam conhecer cada um de seus clientes, identificando seus desejos e anseios. Além disso, as organizações necessitam saber se elas podem oferecer esse diferencial e, caso negativo, como elas poderiam fazê-lo. Isso significa que as organizações precisam de conhecimento sobre seus clientes e sobre os meios e processos capazes de atender às necessidades do cliente.

Este conhecimento pode estar em diferentes formas e provir de diferentes fontes. Uma destas fontes são as próprias pessoas ligadas à empresa. Por ser um dos recursos mais importantes, o conhecimento proveniente das pessoas passou a ser chamado de capital intelectual (Stewart, 1998). Rodrigues e outros (2000, p.72) propõem um modelo onde as

pessoas são o foco das atenções da empresa. Neste caso, as pessoas são a fonte principal de conhecimento para que a organização atinja a competência e o diferencial desejados.

Este conhecimento pode estar disponível internamente com funcionários e colaboradores, mas também pode ser obtido dos clientes. Um modelo que começa a se popularizar é o de organizações centradas no cliente (Imhoff et al., 2001). Os clientes são fonte de conhecimento e inovação para a organização (Pereira e Angeloni, 2000). Muitas vezes, as organizações esquecem que existe mais de uma via na interação com clientes. As empresas fornecem produtos e serviços para satisfazer as necessidades dos clientes e esquecem que estes também têm algo para oferecer em troca, além do pagamento. Por exemplo, pode-se obter conhecimento sobre a área de atuação da empresa, sobre seus produtos ou sobre características dos concorrentes e do mercado..

Este conhecimento todo, seja de pessoas da própria organização ou de clientes, pode ser usado para entender e melhorar a organização, gerando o que se chama de Inteligência. O objetivo final da inteligência é criar diferencial e competência para a organização (inteligência do negócio - *Business Intelligence* – ou inteligência competitiva – *Competitive Intelligence*).

O conhecimento pode ser classificado em tácito ou explícito (Nonaka e Takeuchi, 1997). O primeiro é aquele conhecimento que não está formalizado. Em geral, este tipo de conhecimento encontra-se com as pessoas e não foi ou não pode ser transformado para representações rigorosas. Já o segundo tipo de conhecimento é justamente aquele que foi formalizado em documentos, bancos de dados, gráficos, desenhos, etc.

Em uma organização, os dois tipos de conhecimento devem coexistir harmonicamente e de alguma forma interagir para que todo o potencial de utilização possa ser aproveitado. Nonaka e Takeuchi (1997) identificaram 4 modos de conversão e interação entre os conhecimentos tácito e explícito. O processo de **externalização** é a transformação do conhecimento tácito em explícito. A **internalização** é o processo inverso. Já a **combinação** é o processo de interação entre conhecimentos explícitos para geração de novos conhecimentos. Por sua vez, a **socialização** é a interação entre conhecimentos tácitos.

A tecnologia da informação constitui-se num apoio importante para armazenar e explorar conhecimentos explícitos. Este trabalho apresenta uma abordagem para externalizar e explorar conhecimentos tácitos, disponíveis com clientes ou com pessoas internas à organização. O objetivo é gerar inteligência a partir da análise das informações capturadas e documentadas em textos livres. Para tanto, será utilizada a tecnologia de Text Mining (técnicas e ferramentas). A abordagem apoia os processos de:

- **formalização** de conhecimentos, transformando conhecimentos tácitos em explícitos (externalização); e
- **exploração** do conhecimento formalizado, analisando e integrando conhecimentos explícitos (combinação).

A etapa de **formalização** faz uma análise das informações contidas em textos livres. A tecnologia de Text Mining serve então para identificar os conceitos presentes nos textos. Conceitos representam “entes” do mundo real e permitem entender que temas estão presentes nos textos ou do que tratam os textos.

Em seguida, a **exploração** é feita através de um processo automático de mineração. Nesta etapa, são aplicadas ferramentas de data mining, não sobre dados estruturados de bancos de dados, mas sobre os conceitos extraídos dos textos livres, na etapa anterior. Esta mineração é feita analisando-se a distribuição dos conceitos em coleções (a frequência ou probabilidade com que aparecem) e a relação dos conceitos entre si, para descobrir associações e dependências.

Assim, o processo de formalização e exploração permite gerar novos conhecimentos (inteligência), com o objetivo de melhorar processos internos, serviços, produtos e relacionamento com clientes.

Dois aplicações da abordagem são apresentadas neste artigo. Uma para formalizar e explorar conhecimento de clientes adquiridos através de pesquisas. Neste caso, a pesquisa coletou sugestões e reclamações (textos livres) de clientes sobre produtos e serviços de uma empresa de TV por assinatura. Na segunda aplicação, a tecnologia foi utilizada para formalizar e explorar o conhecimento utilizado por médicos de uma clínica psiquiátrica, com o objetivo de analisar as características de pacientes e entender o processo de

diagnóstico. Neste segundo caso, foram usados prontuários (textos semi-estruturados), criados por médicos no momento da internação dos pacientes.

A seção 2 deste trabalho descreve como a tecnologia pode ser usada para gerar inteligência para a organização. Depois, na seção 3, é apresentada a tecnologia de Text Mining, de forma geral. Na seção seguinte (4), é detalhada a abordagem de Text Mining usada neste trabalho para gerar inteligência. A seção 5 apresenta as aplicações da abordagem, bem como discute as possibilidades de uso do conhecimento descoberto para melhorar a organização. Por fim, a seção 6 apresenta as conclusões do trabalho.

## 2 Inteligência e a Tecnologia de Informação

O ramo da ciência que estuda e aplica métodos e técnicas de análise de informações para geração de inteligência com o intuito de oferecer vantagem competitiva a uma empresa é chamado de Inteligência do Negócio (Wanderley, 1999). Alguns autores fazem uma pequena distinção entre os processos de inteligência: Inteligência do Negócio fica responsável pela análise de dados relativos à organização e Inteligência Competitiva é mais voltada para a análise dos dados do mercado e dos concorrentes.

Para gerar inteligência, é necessário armazenar, analisar e disseminar conhecimento dentro da empresa. Pereira e Angeloni (2000) comentam estratégias para os processos de transformação e interação entre conhecimentos tácitos e explícitos. Por exemplo, o processo de externalização pode ser feito através de metáforas, modelos, analogias, conceitos e hipóteses. Já o processo de combinação deve trabalhar com conjuntos diferentes de conhecimentos explícitos e pode utilizar classificação, acréscimo e combinação.

Entretanto, tais processos não são fáceis de serem feitos, ainda mais quando o volume de informações é muito grande. Para apoiar estas atividades, pode ser utilizada a tecnologia da informação. Por exemplo, para os processos de socialização podem ser usadas tecnologias que apoiem o trabalho cooperativo entre pessoas, tais como sistemas de *groupware*, listas de discussão, fóruns na Web, Intranets, etc.

Já a internalização pode ser apoiada por Intranets (manuais, por exemplo), sistemas de busca na Web e sistemas de recuperação de informação (para encontrar documentos) e tecnologias para ensino à distância (EAD) ou treinamento auxiliado por computador (*Computer-Based Training* e *e-learning*).

A tecnologia tem seu principal uso nos processos de combinação (explícito para explícito), já que é mais fácil trabalhar com conhecimento já formalizado. Neste caso, podem ser usados sistemas de data mining, sistemas de informações gerenciais (SIG), *executive information systems* (EIS), ferramentas OLAP e sistemas de informações geográficas (GIS).

Entretanto, uma das maiores dificuldades para o uso da tecnologia da informação é formalizar o conhecimento (externalização), ou seja, torná-lo disponível em algum meio eletrônico e em um formato que possa ser analisado. Este é o processo básico para que a tecnologia possa ser aplicada; sem ter como capturar o conhecimento, não se pode difundir-lo para outras pessoas (internalização) nem explorá-lo (combinação). Em geral, as organizações utilizam bancos de dados para formalizar o conhecimento, pois isto facilita o uso da tecnologia. Entretanto, as informações disponíveis em bancos de dados são codificadas de forma resumida e estruturada, após algum tipo de filtragem, o que certamente gera perdas. Além disto, 80% das informações de uma organização está disponível em forma textual, não estruturada (Tan, 1999).

Documentos textuais são mais fáceis de serem coletados e armazenados, pois permitem textos livres sem estruturas ou sem formatos limitadores. Isto gera uma riqueza de conhecimento maior que nos bancos de dados. Além disto, documentos textuais possuem conhecimento escondido, implícito nos textos ou em relações entre os documentos (Davies, 1989). Também com o crescente uso da Internet, o conhecimento está cada vez mais disponível em meios eletrônicos, e a forma mais utilizada são os textos. Garofalakis e outros (1999) estimam que a maior parte da informação humana estará disponível na Web em 10 anos.

Entretanto, na maioria dos casos, as pessoas e as organizações não sabem como analisar esta documentação textual para extrair informação nova e útil (combinação) e acabam desperdiçando importantes fontes de conhecimento. Para minimizar este tipo de problema, surgiu a tecnologia de Text Mining.

### 3 A Tecnologia de *Text Mining*

O meio mais simples de externalização é registrar, em textos livres, pensamentos, idéias, sentimentos e opiniões de pessoas. Nas organizações, há muito conhecimento deste tipo disponível na forma de:

- sugestões e reclamações de clientes em pesquisas, emails e serviços de atendimento;
- descrições de defeitos, causas e soluções aplicadas por funcionários;
- manuais, normas e procedimentos definidos como padrão;
- e-mails oriundos de listas de discussão;
- memorandos e comunicações formais, distribuídos através de meios eletrônicos; etc.

Entretanto, as organizações e as pessoas têm dificuldade para tratar adequadamente este tipo de informação por não estar estruturada. A área de Text Mining surgiu para minimizar este problema, ajudando a explorar conhecimento armazenado em meios textuais.

Tan (1999) define *Text Mining* (ou *KDT* – Descoberta de Conhecimento em Textos) como o processo de extrair padrões ou conhecimentos interessantes e não-triviais a partir de documentos textuais.

A tecnologia de Text Mining pode ser usada para formalizar e explorar conhecimento tácito. O conhecimento disponível com pessoas pode ser armazenado em textos, os quais serão analisados para se entender seu significado, ou seja, do que tratam os textos. Depois, pode-se explorar o conhecimento extraído dos textos para gerar novos conhecimentos. Também se pode combinar este conhecimento com o conhecimento explícito armazenado em bancos de dados estruturados.

Existem várias técnicas para Text Mining (Loh et al., 2000). Entretanto, por ser ainda uma área recente, as poucas ferramentas disponíveis são ainda ineficientes (Tan, 1999). Na maioria dos casos, as ferramentas apenas encontram textos que podem conter informações relevantes (ferramentas de recuperação de informação), deixando para os usuários a difícil tarefa de encontrar o conhecimento desejado. Ferramentas mais avançadas separam documentos em grupos por assunto ou afinidade (ferramentas de classificação e clusterização). Entretanto, não conseguem extrair conhecimento novo destes grupos. Também não existem ferramentas adequadas para combinar o conhecimento disponível em textos com conhecimentos formalizados de forma estruturada, por exemplo, em bancos de dados.

### 4 Abordagem de Text Mining para Inteligência

A abordagem apresentada neste artigo utiliza ferramentas de Text Mining para formalizar o conhecimento tácito, ou seja, para transformá-lo em conhecimento explícito (externalização), e para explorar este conhecimento depois de formalizado.

A etapa de **formalização** é feita através da análise de textos livres gerados por meios manuais. Nesta análise, ferramentas são utilizadas para identificar os conceitos presentes nos textos. Conceitos representam entes do mundo real e permitem entender que temas estão presentes nos textos ou do que tratam os textos.

Depois, a **exploração** é feita através de um processo automático de mineração. Nesta etapa, são aplicadas ferramentas de *data mining*, não sobre dados estruturados de bancos de dados, mas sobre os conceitos extraídos dos textos na etapa anterior. Esta mineração é feita analisando-se a distribuição dos conceitos em coleções (a frequência ou probabilidade com que aparecem) e a relação dos conceitos entre si, para descobrir associações e dependências.

A vantagem do uso de conceitos é que estes representam melhor que palavras os objetos, eventos, sentimentos e ações do mundo real. Abordagens baseadas em conceitos (*concept-based approaches*) já são usadas com sucesso para recuperação de informação. Lin e Chen (1996) comentam as vantagens deste tipo de abordagem em relação à busca por palavras-chave. Sua principal vantagem é minimizar o problema do vocabulário (uso de sinônimos, termos correlatos, palavras com vários significados). Uma área onde este tipo de abordagem está sendo usado de forma inovadora é a análise de discurso, para identificar idéias e ideologias presentes em textos. Por exemplo, Chen e outros (1994) usaram com sucesso a identificação de conceitos para organizar idéias discutidas num processo de *brainstorming* eletrônico.

A seguir, serão descritos os métodos e as ferramentas usadas em cada uma das etapas da abordagem (formalização e exploração).

#### **4.1 A extração de conceitos (formalização)**

A extração de conceitos é feita através de um processo semi-automático. As regras para identificação dos conceitos são definidas manualmente com auxílio de ferramentas automatizadas. Depois, um processo de categorização identifica automaticamente os conceitos presentes nos textos usando as regras previamente definidas.

Textos não referenciam explicitamente conceitos, mas sim utilizam palavras para fazer referência a entes do mundo real (Apté et al., 1994). Então é possível identificar os conceitos através da análise de palavras e construções gramaticais (Sowa, 2000).

Nesta abordagem, cada conceito deve ser definido através de uma ou mais regras para identificação. Cada regra será verificada contra todas as frases de um texto. As regras combinam termos positivos e negativos. Para um conceito estar presente em uma frase, todos os termos positivos devem estar presentes na frase e nenhum termo negativo pode aparecer. Se uma das regras for verdadeira para a frase sendo analisada, então o conceito está presente na frase e, conseqüentemente, no texto. Por exemplo, no domínio médico, o conceito “*álcoolismo*” pode ser definido pelas regras (o símbolo “-” indica um termo negativo):

(i) álcool –nega

(ii) hálito etílico

O termo negativo “*nega*” aparece para eliminar frases como “*o paciente nega uso de álcool*”.

Todas as frases são comparadas contra todos os conceitos (e todas as suas regras). Se um conceito está presente mais de uma vez no texto, este valor pode ser usado para indicar o quanto um conceito é referenciado num texto. Por exemplo, se um cliente reclama três vezes de um certo problema na mesma interação, isto é diferente de um cliente citando apenas uma vez o mesmo problema. Por enquanto, a abordagem não está utilizando estes valores, mas sim trabalhando com valores binários (conceito presente ou não).

A definição dos conceitos (quais conceitos serão analisados e as regras de identificação de cada um deles) pode ser feita de várias formas. No momento, a abordagem combina tarefas manuais/intelectuais com ferramentas automatizadas. As ferramentas ajudam as pessoas a entenderem como os termos estão sendo utilizados nos textos (que termos estão sendo usados e como, em que contexto). As pessoas podem ainda aumentar este vocabulário usando sinônimos e palavras correlatas extraídas de dicionários. As ferramentas também são utilizadas para analisar amostras de frases onde os conceitos aparecem, para verificar se as regras funcionam corretamente. Alarmes falsos podem ser analisados para identificar termos negativos.

#### **4.2 A mineração dos conceitos (exploração)**

O processo de mineração aplica ferramentas de *data mining* sobre os conceitos extraídos na etapa anterior. As técnicas utilizadas são as mesmas existentes na área de mineração de dados ou descoberta de conhecimento em bancos de dados (*Data Mining* ou *KDD – Knowledge Discovery in Databases*) (Fayyad et al., 1996). A diferença é que as

ferramentas devem ser aplicadas sobre os conceitos extraídos nos textos e não sobre itens de um banco de dados.

Duas ferramentas específicas estão sendo usadas: uma para análise de distribuições e outra para identificar associações. A primeira verifica a frequência com que ocorrem os conceitos num conjunto de textos (pode ser a coleção toda ou parte dela). O resultado é o que se chama de centróide (um vetor de conceitos e suas frequências). Isto permite analisar que temas são mais dominantes e quais aparecem menos. Também é possível comparar um centróide com outro (por exemplo, centróides de duas subcoleções diferentes). Assim, podem ser encontrados temas comuns em duas coleções ou temas exclusivos e também disparidades ou similaridades nas frequências dos conceitos.

Já a segunda ferramenta descobre relações ou associações entre conceitos, expressando os resultados na forma de regras  $X \rightarrow Y$  ( $X$  pode ser um ou mais conceitos e  $Y$  somente um conceito). A regra significa que “se  $X$  está presente em um texto, então  $Y$  também está presente com um certo grau de certeza”.

O grau de certeza é dado por valores de confiança e suporte. De acordo com a analogia proposta por Lin e outros (1998) e Garofalakis e outros (1999), os textos (ou documentos) são tratados como transações e os conceitos como os itens do banco de dados. Assim, a interpretação do grau de *confiança* (*confidence*) para uma regra associativa do tipo  $X \rightarrow Y$  é a proporção de textos que possuem  $X$  e  $Y$  em relação ao número de textos que possuem somente  $X$ . Da mesma forma, o *suporte* da mesma regra (*support*) é interpretado como o número de documentos onde  $X$  e  $Y$  estão presentes (ou a proporção em relação à coleção toda). O grau de confiança funciona como uma probabilidade condicional. Isto permite prever a presença de um conceito em função da presença de outro.

Nem todas as regras são importantes, novas ou úteis. Para filtrar regras interessantes, devem ser definidos limiares para os valores de confiança e suporte. Feldman e Dagan (1998) também sugerem fazer comparações entre subcoleções (regras comuns e exclusivas) ou comparar as regras das subcoleções com as regras da coleção toda. Também se pode separar a coleção por períodos de tempo e assim comparar as regras extraídas em cada período.

## 5 Aplicações (Estudo de Casos)

A aplicação da abordagem de Text Mining tem por objetivo gerar novos conhecimentos sobre a organização, para melhorar processos internos, serviços, produtos e relacionamento com clientes. Para tanto, o conhecimento tácito (de colaboradores ou de clientes) deve ser armazenado de forma livre em textos não estruturados.

Neste artigo, duas aplicações são apresentadas e discutidas. Uma para formalizar e explorar o conhecimento de clientes, adquiridos através de pesquisas e contendo reclamações ou sugestões sobre o negócio de uma empresa de TV por assinatura.

Na outra aplicação, a abordagem foi utilizada para formalizar e explorar o conhecimento utilizado por médicos de uma clínica psiquiátrica, com o objetivo de analisar as características de pacientes e entender o processo de diagnóstico. Neste segundo caso, o conhecimento tácito foi capturado em textos escritos por médicos (prontuários) no momento da internação dos pacientes e contendo descrições de sinais, sintomas e comportamento social do paciente.

Dois métodos diferentes de definição dos conceitos foram usados (seleção de conceitos e definição das regras para identificação nos textos). Na primeira aplicação, os conceitos foram definidos por leigos, analisando os termos presentes nos textos da coleção e o seu contexto (como eram usados). Na segunda, especialistas da área ajudaram na definição dos conceitos e das regras.

### 5.1 Primeira aplicação: conhecimento de clientes

Neste caso, o conhecimento tácito foi coletado através de uma pesquisa com clientes de uma empresa de TV por assinatura. Sugestões e reclamações dos clientes sobre produtos e serviços da empresa (num total de 225) foram registradas em formato de texto

livre (um registro para cada cliente). Depois de coletados os textos, o processo de formalização seguiu com a identificação dos conceitos presentes.

Nesta pesquisa, havia também informações estruturadas, como tipo de plano ou pacote do cliente e seu canal preferido. Estes últimos dados foram usados no processo de exploração (mineração ou combinação) para separar a coleção de textos por classes.

Na tabela 1, são apresentados alguns exemplos de padrões interessantes descobertos nesta primeira aplicação. Na primeira coluna, aparecem padrões referentes à distribuição dos conceitos na coleção toda, e na segunda coluna, as regras associativas derivadas da coleção toda com o seu grau de confiança.

Algumas conclusões podem ser obtidas dos resultados apresentados na tabela 1. Metade dos clientes tem alguma sugestão ou reclamação sobre filmes. Em geral, uma sugestão vem de uma insatisfação e também pode ser considerada uma reclamação, só que não explícita. Destas reclamações (sobre filmes), 39,5% falam também de repetição, como pode ser notado nas regras associativas. Segundo o senso comum, infere-se que esta é uma insatisfação dos clientes. Este então é um ponto fraco da empresa e seu negócio pode ser melhorado diminuindo-se a repetição de filmes. Ainda pode-se notar que alguns poucos clientes citaram a concorrência (5,3%), mas este pode ser um valor alto para a empresa (deve-se analisar a proporção de perdas de clientes). Destes, segundo as regras associativas, 33,3% citaram o custo. Infere-se que estes clientes estão dizendo que a concorrência tem custo menor.

Tabela 1: padrões descobertos na coleção toda

<b>Distribuições</b>	<b>Regras Associativas</b>
filmes – 50,7%	imagem → qualidade (80,00%)
custo – 20,4%	pacote A → custo (66,67%)
programação – 19,6%	concorrência → filmes (58,3%)
pacote – 15,6%	filmes → repetição (39,5%)
revista – 10,7%	atendimento → demora (37,5%)
pay per view – 6,2%	concorrência → custo (33,3%)
esportes – 5,3%	filmes → qualidade (18,4%)
concorrente – 5,3%	filmes → concorrência (6,1%)
imagem – 4,4%	filmes → pay per view (6,1%)
som – 4,4%	filmes → lançamento (4,4%)
documentários – 3,1%	
seriados – 3,1%	
futebol – 2,7%	

Como discutido anteriormente, o conhecimento tácito formalizado em textos pode ser combinado com conhecimento explícito presente em bancos de dados. Nesta primeira aplicação, o tipo de plano ou pacote que o cliente assina bem como seu canal preferido foram explicitamente registrados. Assim, foi possível separar a coleção de textos em subcoleções, com o intuito de explorar as sugestões e reclamações referentes a cada tipo ou perfil de cliente.

Na tabela 2, são apresentados os conceitos mais frequentes nas reclamações dos clientes do pacote A (mais caro) e dos clientes do pacote D (mais barato).

Tabela 2: padrões por tipo de pacote

<b>Pacote A</b>	<b>Pacote D</b>
filmes – 36,4%	custo – 38,5%
custo – 24,2%	atendimento – 23,1%
repetição – 22,7%	filmes – 15,4%
programação – 22,7%	

Pode-se observar na tabela 2 que os clientes do pacote A reclamam menos do custo que os clientes do pacote D e que os primeiros estão mais insatisfeitos com os filmes da programação geral do que os segundos. Este conhecimento permite entender melhor os interesses dos clientes de cada pacote, podendo-se gerar um perfil de clientes por tipo de pacote.

Como havia o registro explícito do canal preferido de cada cliente, a coleção de textos foi dividida em 3 partes referentes ao tipo de canal preferido (esportes, filmes e notícias). Na tabela 3, são apresentados os conceitos mais frequentes por tipo de canal preferido.

Tabela 3: padrões por canal preferido

<i>Esportes</i>	<i>Filmes</i>	<i>Notícias</i>
filmes – 39,4%	filmes – 60,9%	filmes – 65,4%
custo – 30,3%	custo – 17,2%	custo – 19,2%
pay per view – 15,2%	pay per view – 4,7%	pay per view – 7,7%
concorrência – 15,2%	concorrente – 3,1%	concorrente – 0
atendimento – 6,1%	atendimento – 7,8%	atendimento – 11,5%
clube – 0	clube – 3,1%	clube – 7,7%
ponto extra – 0	ponto extra – 3,1%	ponto extra – 0

O quadro comparativo da tabela 3 permite traçar um perfil do cliente por interesse. Uma das constatações é 15,2% dos clientes que preferem canais de esportes citaram também o conceito “*pay per view*”, talvez estando mais suscetíveis a fazer aquisições deste tipo do que os demais (4,7% em filmes e 7,7% em notícias). Pode-se também notar que os clientes que citaram os canais de filmes como preferidos também citaram o conceito “*ponto extra*” (os outros não). Disto pode-se inferir que estes clientes estão mais interessados em ter um ponto extra. Dos clientes que escolheram um canal de esporte como favorito, 15,2% citaram a concorrência (bem mais que os demais clientes). Isto levanta a hipótese de que a concorrência possa estar oferecendo algo melhor em termos de esportes. Analisando-se as regras associativas desta primeira subcoleção (não apresentadas neste trabalho), não foi detectado nenhum padrão associativo entre “*pay per view*” e “*concorrente*”. Assim, pode-se inferir que estes 15,2% referentes aos dois conceitos não são os mesmos clientes, ou seja, quem cita um destes dois conceitos provavelmente não cita o outro (nesta subcoleção).

## 5.2 Segunda aplicação: conhecimento interno à organização

Na segunda aplicação, a abordagem foi utilizada para formalizar e explorar o conhecimento utilizado por médicos de uma clínica psiquiátrica. O processo de formalização (externalização) iniciou com o registro em textos livres do resultado da entrevista do médico com o paciente e seus familiares, feita na internação. Estes textos formam parte do prontuário do paciente na clínica e contém informações sobre o comportamento social e familiar do paciente, história pregressa, remédios que toma ou que foram prescritos, além de sinais e sintomas identificados pelo médico durante a entrevista.

Durante 4 meses, foram coletados 400 textos. Para cada texto havia associado um diagnóstico, decidido por um médico da clínica para representar a doença mental do paciente. Entretanto, a indicação do diagnóstico não estava explicitamente expressa no texto. Os textos podem ser considerados semi-estruturados, pois, apesar de serem escritos em linguagem livre, continham informações previamente planejadas. Isto é, o médico anotava somente informações relevantes para o diagnóstico.

A formalização foi completada com a identificação dos conceitos presentes nestes prontuários através das ferramentas de Text Mining. Os conceitos definidos para esta aplicação representavam sinais, sintomas, pessoas, objetos, eventos e referências ao comportamento do paciente, por exemplo: insônia, agressividade, familiares, mãe, irmãos, arma de fogo, choro, uso de álcool.

Para o processo de exploração (mineração), a coleção toda de textos foi analisada de forma conjunta e também de forma separada por diagnóstico, combinando-se o conhecimento tácito formalizado (conceitos) com o explícito previamente existente (doença do paciente). Também foi possível separar a coleção por remédio utilizado, já que os diferentes remédios foram definidos como conceitos e identificados nos textos, ou seja, não era necessário ter esta informação de maneira prévia e estruturada (os remédios foram inferidos dos textos livres).

Na tabela 4, são apresentados alguns padrões interessantes encontrados na coleção toda. Na coluna da esquerda aparecem os conceitos mais frequentes e na coluna da direita as regras associativas com maior grau de confiança.

Tabela 4: padrões descobertos na coleção toda

Distribuições	Regras Associativas
familiares – 84,5%	alcoolismo → inapetência (84%)
agressividade – 77,0%	autismo → alteração de pensamento (95,3%)
inapetência – 76,0%	agressividade → familiares (92,8%)
remédios – 74,5%	depressão → insônia (85,1%)
insônia – 71,0%	religião → remédios (85,1%)
alteração de pensamento – 70,5%	
nervosismo – 68,5%	
alteração de atenção – 54,5%	

Este conhecimento descoberto permite traçar um perfil do paciente típico que é atendido na clínica. Isto quer dizer que, pela tabela 4: 84,5% dos pacientes têm familiares ou fazem algum tipo de referência a estes; 77% apresentam sinais de agressividade; 76% citaram sofrer de algum tipo de inapetência (falta de apetite), 74,5% já fazem uso de algum remédio, etc.

Analisando-se as regras associativas da tabela 4, é possível prever características pela presença de outras ou inferir relações de dependência entre as características. Por exemplo, pode-se notar que: em 84% dos casos, pacientes com sintomas de alcoolismo também apresentam inapetência (falta de apetite); quase sempre, sintomas de autismo são acompanhados de alteração de pensamento; 85,1% dos pacientes com sintomas de depressão apresentam também insônia; e 85,1% dos pacientes que citaram algo relacionado a religiões tomam remédios. Pode-se também levantar a hipótese de que a agressividade esteja relacionada à família, já que 92,8% dos pacientes com sintomas de agressividade citaram familiares.

Na tabela 5, são apresentados padrões (conceitos mais frequentes e regras associativas) referentes aos pacientes com o diagnóstico de esquizofrenia. Analisando tais padrões, é possível verificar que os pacientes esquizofrênicos apresentam maior incidência de “*alteração de pensamento*” (83,5%) do que a média geral (70,5% na tabela 4). Pelas regras de associação, é possível inferir que pacientes que “*ouvem vozes*” têm predisposição para agressividade, e pacientes homicidas (com alguma história envolvendo ameaças de morte) também apresentam sintomas de agressividade.

Tabela 5: padrões para o diagnóstico de esquizofrenia

Distribuições	Regras Associativas
familiares – 84,5%	agressividade → familiares (92,94%)
alteração de pensamento – 83,5%	alteração de atenção → agressividade (88,89%)
agressividade – 82,5%	homicida → agressividade (97,67%)
nervosismo – 75,7%	homicida → familiares (95,35%)
insônia – 75,7%	ouvir vozes → agressividade (90,91%)
remédios – 73,8%	perseguição → insônia (84,85%)
inapetência – 68,9%	insônia → remédios (80,77%)
perseguição – 64,1%	
ouvir vozes – 64,1%	
alteração de atenção – 52,4%	

Os padrões de esquizofrenia podem ser comparados com os de outras doenças. Na tabela 6, por exemplo, é apresentado o conhecimento descoberto sobre os pacientes com distúrbios afetivos. Pode-se notar que a incidência de insônia é bem maior nos pacientes com distúrbios afetivos que nos esquizofrênicos (85,2% contra 75,7%). Nesta segunda classe, aparecem novos sintomas entre os mais frequentes: “*suicida*” (81,5%), “*depressão*” (74,1%) e “*choro*” (63%). As regras associativas também apresentam diferenças.

Os padrões descobertos em cada classe de paciente permitem traçar o perfil de doenças para estudos epidemiológicos. Também podem ser usados para treinamento de pessoas, para validação de decisões médicas ou para a construção de sistemas automáticos

de classificação. Nestes casos, a abordagem de Text Mining funciona como um mecanismo de aprendizagem supervisionada, onde textos são selecionados como exemplos de uma determinada classe, e os padrões descobertos servem para descrever as características desta classe. Em um experimento anterior, um sistema automatizado criado com este conhecimento chegou a mais de 60% de acertos no diagnóstico de novos casos.

Tabela 6: padrões para o diagnóstico de distúrbios afetivos

Distribuições	Regras Associativas
insônia – 85,2%	agressividade → familiares (87,50%)
familiares – 85,2%	homicida → suicida (92,31%)
suicida – 81,5%	remédios → suicida (85,71%)
inapetência – 81,5%	morte → inapetência (90,91%)
remédios – 77,8%	inapetência → insônia (86,36%)
depressão – 74,1%	inapetência → suicida (81,82%)
pensamento – 70,4%	
alteração de atenção – 66,7%	
choro – 63,0%	
nervosismo – 63,0%	
agressividade – 59,3%	

Nesta segunda aplicação, foi possível também descobrir padrões referentes ao uso de remédios. A seguir, são apresentados os conceitos mais freqüentes identificados na subcoleção de pacientes que utilizaram o remédio Dienpax: “*inapetência*” (91.8%), “*agressividade*” (83.7%), “*alteração de pensamento*” (78.3%), “*nervosismo*” (75.6%), “*insônia*” (64.8%), “*alcoolismo*” (62.1%), “*ouvir vozes*” (59.4%). Este conhecimento permite avaliar o uso da medicação ou pode ser utilizado no treinamento de estudantes e médicos assistentes.

## 6 Conclusão

Este artigo mostrou que a tecnologia de Text Mining pode ser usada para formalizar e explorar conhecimento tácito, se este for capturado em textos.

Pelo que se pôde ver nas aplicações discutidas, a abordagem permitiu gerar novos conhecimentos sobre a organização.

As duas aplicações apresentadas demonstram que é possível analisar tanto o conhecimento interno das organizações quanto o conhecimento disponível com os clientes.

Na primeira aplicação, a abordagem foi utilizada para gerar inteligência do negócio a partir do conhecimento dos clientes de uma empresa de TV por assinatura. Foram apontados caminhos para melhorar processos, serviços, produtos e relacionamento com clientes.

Na segunda aplicação, o conhecimento interno disponível com colaboradores de uma clínica psiquiátrica pôde ser formalizado e explorado. Este conhecimento servirá para entender melhor o perfil dos pacientes da clínica e o processo de diagnóstico feito pelos médicos. Também poder-se-á utilizar o conhecimento descoberto para treinar novos colaboradores, para validar decisões e para a construção de sistemas automatizados (sistemas especialistas ou de suporte à decisão).

Alguns cuidados se fazem necessários no uso desta abordagem. Primeiro, o processo de mineração fica condicionado à qualidade da identificação dos conceitos nos textos. A avaliação do processo de extração conduzida no segundo experimento apontou um resultado de 90% de acertos na identificação dos conceitos. Também podem surgir erros quando o conhecimento tácito é registrado em textos livres. Erros ortográficos ou informações incorretas, imprecisas, ambíguas e incompletas podem distorcer os resultados finais. Por fim, a interpretação dos resultados também está condicionada à interpretação dos conceitos. Por exemplo, na segunda aplicação, o conceito “*alcoolismo*” deve ser interpretado como uma referência a tal nos textos e não como a certeza de que o paciente tem este sintoma (a referência pode ser de que um familiar usa álcool).

## 7 Agradecimentos

Este trabalho tem o apoio parcial de: CNPq, CAPES, FIDEPS (Fundo de Incentivo ao Desenvolvimento do Ensino e da Pesquisa em Saúde).

## 8 Referências Bibliográficas

- APTÉ, C. et al. (1994). "Automated learning of decision rules for text categorization". *ACM Transactions on Information Systems*, v.12, n.3, pp.233-251.
- CHEN, H. et al. (1994). "Automatic concept classification of text from electronic meetings". *Communications of the ACM*, v.37, n.10, pp.56-73. Online at <http://ai.bpa.arizona.edu/papers/ebs92/ebs92.html>
- DAVIES, Roy. (1989). "The creation of new knowledge by information retrieval and classification". *Journal of Documentation*, v.45, n.4, pp.273-301.
- FAYYAD, Usama M. et al. (eds) (1996). *Advances in Knowledge Discovery and Data Mining*. Menlo Park: The MIT Press.
- FELDMAN, R. & DAGAN, I. (1998). "Mining text using keyword distributions". *Journal of Intelligent Information Systems*, v.10, n.3, pp. 281-300.
- GAROFALAKIS, Minos N. et al. (1999). "Data mining and the web: past, present and future". In: *ACM Workshop on Information and Data Management*, Kansas City.
- IMHOFF, C.; LOFTIS, L.; GEIGER, J. (2001). *Building the customer centric enterprise*. John Wiley & Sons.
- LIN, C.H. & CHEN, H. (1996). "An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese-English) documents". *IEEE Transactions on Systems, Man and Cybernetics*, v. 26, n.1, pp. 1-14. Disponível por WWW em <http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html>
- LIN, S.H. et al. (1998). "Extracting classification knowledge of Internet documents with mining term associations: a semantic approach". In *Proc. 21<sup>st</sup> International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, Melbourne, August 1998, pp.241-249.
- LOH, Stanley; WIVES, Leandro K.; OLIVEIRA, José Palazzo M. "Descoberta proativa de conhecimento em textos: aplicações em inteligência competitiva". In: LETHÉLIER, E. et al. (eds). *Proceedings, International Symposium on Knowledge Management / Document Management*, Novembro de 2000. Curitiba: Editora Universitária Champagnat, p.125-147.
- NONAKA, I. & TAKEUCHI, H. (1997). *Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação*. Rio de Janeiro: Campus.
- PEREIRA, Rita C. F. & ANGELONI, M. T. (2000). "O relacionamento com os clientes para transformação do conhecimento na organização". In: LETHÉLIER, E. et al. (eds). *Proceedings, International Symposium on Knowledge Management / Document Management*, Novembro de 2000. Curitiba: Editora Universitária Champagnat, p.89-104.
- RODRIGUES, Hugo T. et al. (2000). "Arquitetura da gestão pelo conhecimento focada na inovação". In: LETHÉLIER, E. et al. (eds). *Proceedings, International Symposium on*

Knowledge Management / Document Management, Novembro de 2000. Curitiba: Editora Universitária Champagnat, p.59-76.

SOWA, J.F. (2000). Knowledge representation: logical, philosophical, and computational foundations, Pacific Grove: Brooks/Cole Publishing Co.

STEWART, Thomas A. (1998). Capital intelectual: a nova vantagem competitiva das empresas. 2<sup>a</sup> ed. Rio de Janeiro: Campus.

TAN, Ah-Hwee. (1999). "Text mining: the state of the art and the challenges". In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases – PAKDD'99, p. 65-70, Beijing, April 1999. Disponível por WWW em <http://textmining.krdl.org.sg/publications.html>.

WANDERLEY, A.V.M. (1999). "Um instrumento de macropolítica de informação: concepção de um sistema de inteligência de negócios para gestão de investimentos de engenharia". Revista Ciência da Informação, v.28, n.2.