O caso de negócios do uso de análise de texto não estruturado em IBM Power Systems para tomada de decisão crítica

Stephen Markham, PhD
Michael Kowolenko, PhD
Universidade Poole de Gerenciamento
Universidade do Estado da Carolina do Norte

Essa é a questão; não a tecnologia
Michael Kowolenko

Os grandes volumes de dados prometem alterar dramaticamente o ambiente comercial, entretanto, a tecnologia é apenas um facilitador de decisão. Muitas firmas aplicam abordagens estruturadas aos grandes volumes de dados para tomar decisões operacionais de rotina. Tomar decisões estratégicas e críticas não programadas, entretanto, geralmente envolve dados não estruturados. Este documento descreve o valor comercial do uso de técnicas de grandes volumes de dados para coletar e analisar dados não estruturados. Esta abordagem usa o pensamento crítico inserido em um processo junto com servidores e softwares avançados para converter grandes volumes de dados em valor comercial. O processo e os resultados abrem novas oportunidades que requerem estruturas, cultura e conhecimento adaptáveis para cumprir a promessa dos grandes volumes de dados.

Este artigo tem três finalidades: Primeiro, explicar a diferença na forma em que os dados estruturados e não estruturados são usados na tomada de decisão. Segundo,

dar exemplos de como as empresas atingiram valor comercial usando dados não estruturados. Terceiro, dar orientação sobre como as empresas podem alcançar um valor comercial semelhante usando dados não estruturados, e por que escolher os servidores e os softwares certos podem fazer uma diferença.

1. A diferença na forma em que os dados estruturados e não estruturados são usados na tomada de decisão

Textos não estruturados compõem cerca de 80% dos dados disponíveis hoje em dia. Empresas que usam somente dados estruturados não se beneficiam da maioria das informações disponíveis. Dados não estruturados incluem todo o texto contido em relatórios governamentais da SEC, NIH, NSF, DOE, assim como todas as pesquisas acadêmicas, relatórios de análise financeira e comercial, resultados de pesquisa de consultoria e muitas outras fontes. Texto não estruturado também é encontrado em uma variedade de portais de mídias sociais, como Facebook, blogs, registros de reclamações de clientes, Twitter, transcrições de notícias, imprensa popular, revistas especializadas e muitas outras.

As necessidades dos clientes, as ações de concorrentes, as tendências emergentes e outras partes individuais de informações necessárias para tomar decisões comerciais críticas fazem parte dos dados estruturados. Abordagens estruturadas dos grandes volumes de dados coletam e agregam linhas e colunas de números para informar os tomadores de decisão. Grandes conjuntos de números são analisados por meio de técnicas estatísticas avançadas para revelar padrões valiosos nos dados. Essas técnicas permitem que os tomadores de decisão vejam o que aconteceu ou está acontecendo em tempo real. A análise de dados que podem ser adicionados, subtraídos, multiplicados e divididos conta com uma abordagem estruturada para agregar os dados e é essencial para a tomada de decisões operacionais relacionadas a preço, distribuição e inventário.

Abordagens não estruturadas por outro lado buscam isolar partes críticas de informações. Por exemplo, grandes volumes de dados não estruturados encontram anúncios de que um concorrente está construindo uma nova instalação ou de que um cliente está expandindo as operações. Isso dará tempo para os tomadores de decisão reagirem, antes que os dados estruturados revelem uma diminuição na receita de vendas. Para encontrar de forma confiável uma “agulha no palheiro” é necessário usar grandes volumes de dados para coletar vastas quantidades de texto não estruturado e usar programas especializados para procurar partes específicas de informações entre dezenas de milhões de documentos com o olhar inflexível de um computador.

O pensamento crítico direciona o uso de grandes volumes de dados.

A habilidade de tomar decisões direcionadas aos dados presume que as pessoas saibam quais perguntas fazer, mas a nossa experiência mostrou que nem sempre esse é o caso. Muitas organizações não têm processos para aplicar o pensamento crítico. Em cada um dos projetos patrocinados pelo setor conduzidos pelo CIMS (Center for

Innovation Management Studies)[1], empresas startups e Fortune 500 lutam para gerar consultas estratégicas.

O pensamento deve direcionar o uso dos grandes volumes de dados. Os grandes volumes de dados não podem conduzir o pensamento. O pensamento crítico é a base para encontrar as fontes de dados e as ferramentas para reconhecer o significado subjacente no texto não estruturado. Por exemplo, a afirmação "Nossa empresa precisa examinar as redes sociais para analisar os sentimentos" pode ser dividida em uma série de subperguntas, tais como:

- Sentimento no que diz respeito ao nosso produto
- Sentimento relacionado à concorrência
- O cliente gosta ou não gosta da classe de produtos em questão
- Como os novos produtos da nossa empresa lidam com a insatisfação do cliente
- O que a concorrência está desenvolvendo para lidar com as insatisfações do cliente

Com base nos princípios do pensamento crítico, foi desenvolvido um processo para obter valor comercial do texto não estruturado. Na prática, a aplicação do pensamento crítico ao desenvolvimento do processo de tomada de decisão comercial tem sido bem-sucedida em várias empresas.

O processo de uso do pensamento crítico para analisar dados não estruturados

O processo (Figura 1) requer a participação de uma equipe interdisciplinar de participantes afetados desde o início e tem sido comprovadamente bem sucedido em vários setores. Esse processo usa técnicas de pensamento crítico para: 1) definir o problema e preparar perguntas específicas para fazer consultas, 2) identificar fontes de informações, 3) identificar os termos de pesquisa e definir os relacionamentos entre os termos (chamados de regras), 4) aplicar a tecnologia de grandes volumes de dados aos termos e regras para coletar, armazenar e analisar quantidades massivas de dados coletados de fontes externas e internas, 5) analisar a suficiência, aplicabilidade e a veracidade dos dados, e finalmente 6) quando os filtros tiverem sido aplicados nos dados, a equipe avalia a evidência que apoia ou refuta as presunções ou condições necessárias para tomar a decisão. É importante enfatizar que esse é um processo iterativo. O ato de coletar e filtrar os dados frequentemente resulta em novas ideias que requerem mais investigação. Isso requer que a equipe supere uma tendência individual e de grupo ao confrontar uma nova evidência.

[1] O CIMS é um IUCRC graduado. Formado em 1984, ele é o único centro de pesquisa patrocinado pela NSF dedicado a
investigar os efeitos gerenciais e organizacionais na inovação.

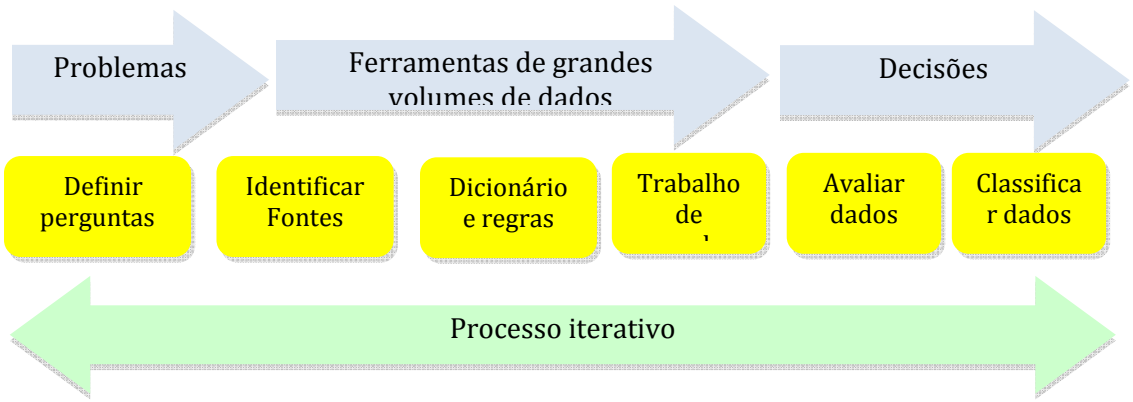

Figura 1. Processo para tomada de decisão possibilitada por dados não estruturados

2. Exemplos de como as empresas atingiram valor comercial usando dados não estruturados

Apenas cinco das dezenas de exemplos de questões comerciais tratadas com êxito por este processo são destacadas na tabela abaixo. Cada exemplo representa um tipo diferente das questões mais frequentes. Outra ideia extraída desses exemplos é que as decisões não são baseadas em análises quantitativas de dados estruturados, mas na interpretação dos fatos derivados da análise de dados não estruturados. A seleção das questões, o uso de termos e a criação de dicionários e regras permitem a configuração do software para usá-lo para responder a uma ampla variedade de decisões críticas.

Setor: Empresa	Pergunta	Fontes	Resultado
Força de trabalho temporária: Kelly Services	Desenvolver novas ofertas de serviço para a equipe de assistência médica	SEC, URLs, Publicações comerciais, Publicações profissionais, Provedores de seguro	A decisão de seguir adiante em um domínio de assistência médica inesperado
Gases industriais: Air Products	Encontrar novos clientes e oportunidades de mercado	SEC, feeds de notícias, Publicações do setor, permissões de construção	A identificação de um novo cliente que esteja planejando construir novas instalações
Universidade: NC State	Identificar padrões comerciais de novas tecnologias	SEC, URLs, Publicações do setor	Parceiros potenciais identificados para colaborações
Organização de pesquisa clínica: PRA International	Oferecer inteligência comercial para novos testes	Clintrials, PubMed	A identificação de novos médicos/hospitais com conhecimento

	clínicos		em áreas de testes clínicos.
Organização não governamental: Clinton Health Care Access Initiative	Encontrar o ajuste entre as novas tecnologias e as oportunidades de mercado para o diagnóstico de doenças	Clintrials, PubMed Firmas VC	A identificação de laboratórios de pesquisa ativos em pesquisa de diagnóstico de ponta

Tabela 2. Exemplos do setor

Kelly Services. Desenvolver novas ofertas de serviço para a equipe de assistência médica

A Kelly Services é uma líder em equipe temporária e expandiu suas ofertas para uma empresa de soluções de equipe de serviço integral. Eles queriam explorar se a oferta de serviços de equipe temporária para o setor de assistência média seria viável. Trabalhar com especialistas externos e internos, dados não estruturados sobre a necessidade desse novo serviço, concorrência e regulamentos, não somente revelou inúmeras oportunidades, mas também ofereceu orientação sobre como proceder.

Descobriu-se que havia uma grande necessidade de as enfermeiras trabalharem remotamente para oferecer avaliações médicas e aconselhamento para uma variedade de clientes, incluindo hospitais, instalações de cuidados de enfermagem, companhias de seguro e organizações de auto-seguro. Também se chegou à conclusão de que havia uma escassez de enfermeiras e que a demanda não estava sendo atendida local ou regionalmente. Combinar os recursos de equipe nacional da Kelly com as necessidades regionais permite que a Kelly atenda exclusivamente às demandas desconhecidas por seus concorrentes.

Também se descobriu que a habilidade desses provedores de serviços de assistência médica de serem pagos por esses tipos de serviços de telemedicina variava muito por estado. A análise de texto não estruturado foi usada para descobrir os regulamentos exatos em cada estado. Isso forneceu à Kelly o conhecimento para entrar nos estados certos com regulamentos favoráveis com os serviços certos. Esse novo serviço agora é uma das duas principais iniciativas inovadoras na Kelly.

Air Products and Chemicals. Encontrar novos clientes e oportunidades de mercado

A Air Products oferece gases industriais e produtos químicos especializados em uma dúzia de empresas verticais. Eles queriam identificar novos clientes potenciais e fornecer informações específicas para sua força de vendas sobre esses clientes. No processamento de metal, é importante identificar clientes antecipadamente, mesmo antes de a produção começar. As fábricas de processamento de metal são projetadas com alimentações de gás específicas, sendo assim, é importante que os fornecedores de gás saibam o quanto antes para poder atender às necessidades do novo cliente.

O processo de coleta de dados começou com a “leitura” de todos os dados SEC de empresas com códigos SIC de processamento de metal para identificar as empresas que estavam planejando gastos de capital em novas fábricas e equipamentos. Além da leitura de todos os jornais locais em uma tentativa de encontrar empresas mencionadas em relação à contratação de pessoal. Além disso, a leitura de jornais estrangeiros (em 22 idiomas) em países que produzem equipamentos de processamento de metal de anúncios sobre novas vendas de equipamentos. Finalmente, a leitura de todas as permissões de construção nos EUA para identificar empresas de processamento de metal que estão construindo novas fábricas.

A pesquisa descobriu uma empresa de processamento de metal que está construindo uma nova fábrica no Alabama, que pediu um tipo específico de equipamento e anunciou um número específico de novos empregos. Não somente a Air Products identificou um novo cliente potencial, mas eles também descobriram qual tipo de equipamento eles iam usar (definindo, portanto, o tipo de gás necessário), quanto metal a empresa planejava processar (indicando o volume de gás necessário). Tudo 18 meses antes de a construção começar. Isso deu à Air Products uma vantagem competitiva distinta na abordagem desse novo cliente. Uma análise adicional também revelou inúmeros clientes existentes que estavam expandindo suas operações, informação que não era conhecida anteriormente.

Universidade do Estado da Carolina do Norte. Identificar padrões comerciais potenciais de novas tecnologias

Como muitas universidades e empresas, a NC State possui muitas tecnologias patenteadas que não estão sendo comercializadas. A universidade quis apontar quais empresas podem precisar dessas tecnologias. Usando um subconjunto de tecnologias, a análise de texto não estruturado foi usada para identificar as características das empresas de cada uma das tecnologias.

Os resultados identificaram alvos de grande potencial que foram abordados por funcionários da universidade com os recursos de tecnologia correspondentes aos usos da empresa-alvo. Isso resultou em negociações e licenças adicionais de tecnologias não usadas anteriormente.

PRA. Oferecer inteligência comercial para novos testes clínicos

A PRA Internal é uma organização de pesquisa clínica que conduz testes de drogas para companhias farmacêuticas. Eles queriam ter uma ideia melhor sobre o que outras organizações de pesquisa clínica estavam fazendo, além de prever com mais precisão o que os clientes vão querer no futuro próximo para poder atendê-los melhor.

A análise de texto não estruturado foi usada para a leitura de quatro milhões de arquivos de dados contendo mais de 75 milhões de páginas da Web. Essas fontes de informações não estruturadas foram depois vinculadas a informações contidas em www.clinicaltrails.gov e a arquivos de dados específicos da empresa para oferecer uma visão muito mais detalhada do que estava acontecendo no setor.

Em um projeto, a PRA quis saber o que estava acontecendo na pesquisa de mieloma. Eles queriam saber uma variedade de coisas que não estavam comumente disponíveis, tal era o motivo de falha no teste de mieloma anterior, e quais empresas podem estar trabalhando em mieloma sem que isso apareça em relatórios do governo e do setor. Essa pesquisa fez várias descobertas relevantes. A maioria das falhas de teste foram causadas por falta de inscrição, mas outros motivos importantes surgiram. Interessantemente, foram encontradas sete empresas que usam os mesmos alvos de genes para uma variedade de outras indicações como asma, câncer de mama e de pulmão, mal de Parkinson, Alzheimer, anemia falciforme e cegueira. Também foram encontrados os nomes dos gerentes desses projetos, o que permitiu que a PRA estabelecesse relacionamentos para avançar mais em sua própria pesquisa.

Clinton Healthcare Access Initiative (CHAI). Encontrar um ajuste entre as novas tecnologias e as oportunidades de mercado para o diagnóstico de doenças

A CHAI quis avaliar se seus investimentos de assistência médica em diagnóstico eram eficazes. Eles estavam apoiando um grande número de iniciativas e queriam maximizar o efeito que elas estavam tendo.

Os grandes volumes de dados foram usados para avaliar a eficácia dos novos diagnósticos, mas também qual impacto as novas ferramentas de diagnóstico podem ter. Analisando dados não estruturados de milhares de artigos e relatórios, descobriu-se que o impacto do desenvolvimento de novos diagnósticos de novas doenças seria muito menor do que fazer diagnósticos existentes para doenças comuns. Isso causou uma reavaliação de sua estratégia e o realinhamento de seus recursos para impactar favoravelmente mais pessoas.

Conclusão. Esses exemplos demonstram o valor comercial do uso da análise de texto não estruturado para encontrar e avaliar novas oportunidades de negócios, localizar e qualificar novos clientes, oferecer inteligência comercial extremamente aprimorada e mais detalhada e tomar decisões estratégicas de alocação de recursos. Dados não estruturados podem ser usados em muitas outras situações onde as decisões requerem o conhecimento de informações específicas que não podem ser adicionadas, subtraídas, multiplicadas ou divididas.

3. Como as empresas podem obter valor comercial usando dados não estruturados

A vantagem de usar grandes volumes de dados não estruturados é que as ferramentas e as técnicas agora foram refinadas para permitir que as pessoas com habilidades de pensamento crítico respondam a questões comerciais importantes sem ser programadores de software ou estatísticos. O processo pode parecer complicado a princípio e o software especializado, mas não é mais de competência exclusiva de cientistas de dados.

Um projeto de grandes volumes de dados não estruturados requer a participação do departamento de TI, para dar suporte ao software e ajudar a coletar e armazenar os dados, e dos analistas estatísticos, para conduzir o processo de grandes volumes de

dados. Mas de longe o componente mais importante para o sucesso da análise de dados não estruturados é a habilidade de pensamento crítico dos especialistas da empresa (SMEs). Esta é uma capacidade do processo de tomada de decisão interdisciplinar. Não se trata de um evento único. Portanto, a empresa deve se comprometer com os recursos de SME necessários para conduzir o processo e tomar as decisões, e o gerenciamento sênior com a autoridade de agir a respeito delas.

Essa democratização dos grandes volumes de dados para uso por empresários não é apenas um benefício nem mesmo um objetivo dessas ferramentas, mas uma necessidade. O conteúdo comercial deve direcionar as questões, termos, fontes e regras necessários para obter valor comercial.

É fundamental escolher o software correto para conduzir a análise de texto não estruturado. Existem muitos programas comerciais e de código-fonte aberto que podem ser usados. Se você não escolher um programa que tenha uma interface de usuário gráfica e a habilidade de interagir continuamente com grandes conjuntos de dados, você precisará de cientistas de dados apenas para executar a parte técnica do software.

O software IBM Content Analytics Studio (ICA) atende exclusivamente a todos os critérios para permitir que a equipe de conteúdo comercial se comprometa com a análise de grandes volumes de dados não estruturados por si. Com alguns dias de treinamento, muitos empresários aprendem a usar o ICA bem o suficiente para aplicá-lo continuamente em sua área. Portanto, os grandes volumes de dados podem ser usados rotineiramente por uma ampla variedade de pessoas em vez de um projeto especial.

Também é importante usar a plataforma de servidor certa para grandes volumes de dados. Embora busquemos isolar informações altamente específicas para tomar decisões em vez de agregar grandes conjuntos de dados, só podemos identificar essa informação crítica se pudermos coletá-la e processá-la de forma precisa. Consequentemente, escolher a plataforma de servidor correta pode ser um aspecto crucial da implantação de um projeto de grandes volumes de dados não estruturados bem-sucedido.

Algumas empresas iniciantes decidiram usar o x86 porque já têm os servidores instalados e o software de grandes volumes de dados, incluindo o ICA, sendo executado no x86. Existem limitações profundas, entretanto, para essa abordagem. Enquanto os servidores x86 podem ser suficientes para pequenas demonstrações, a baixa confiabilidade da plataforma prejudica a adoção dos grandes volumes de dados. Talvez você não consiga demonstrar realmente o valor total dos grandes volumes de dados em servidores x86. Usamos servidores x86 e IBM Power Systems para o processamento de grandes quantidades de dados. Não há dúvidas de que os servidores Power são muito superiores. Os servidores x86 travam tão frequentemente que optamos por não usá-los com nossos clientes.

O servidor Power System, com a tecnologia baseada no processador POWER8, foi projetado para grandes volumes de dados para executar mais consultas simultâneas em paralelo rapidamente, entre muitos cores com mais threads por core. Eles também têm maior largura de banda de memória e ES mais rápida para ingerir, mover e acessar dados. Isso permite que as empresas executem essas consultas de análise de dados mais rapidamente.

Obtendo uma vantagem competitiva com grandes volumes de dados

A análise de dados não estruturados pode ser a fonte de grande valor comercial se os processos, as ferramentas, as habilidades e a estrutura apropriados forem implementados em conjunto. Frustação e desapontamento o esperam se você não implementar esses elementos em conjunto. O custo da falha é ficar atrás dos concorrentes que estão implementando com êxito os grandes volumes de dados.

A chave para uma implantação bem-sucedida é estabelecer um processo simples para os especialistas em conteúdo comercial e os tomadores de decisão usarem regularmente. Isso significa que as entradas e as saídas do processo de tomada de decisão devem ser apropriadamente dotadas de recursos, e as funções e as responsabilidades estabelecidas. Você deve estabelecer um relacionamento de subordinação com os tomadores de decisão para garantir a responsabilidade. As expectativas de desempenho e de resultado devem ser estabelecidas em relação ao que a sua empresa quer dos grandes volumes de dados. A infraestrutura certa deve ser usada para garantir que os níveis necessários de desempenho e confiabilidade sejam atendidos. Uma implementação apropriada promete valor comercial e vantagem competitiva, dotando os tomadores de decisão de informações mais direcionadas.