Artefactum
Revista de estudos em Linguagens e Tecnologia

MINERAÇÃO DE TEXTOS: CONFLUÊNCIA DE SABERES NA BUSCA DA PRODUÇÃO DE CONHECIMENTOS

Daniel Costa Vianna Mucciolo
danielmucciolo@unc.br
http://lattes.cnpq.br/9827685950079560

RESUMO

Este trabalho tem o intuito de apresentar as potencialidades do campo da mineração de textos, uma área que se utiliza de ferramentas para processar bancos de dados textuais com o intuito de produzir conhecimentos. Esta disciplina vem avançado bastante nos últimos anos graças aos avanços tecnológicos e dos mecanismos de análise e processamento dos computadores, aliado a isso, estamos vivendo uma era em que a produção de dados digitais é extremamente volumosa. E por último, o desenvolvimento desse campo é marcado pelo esforço congruente de pesquisadores de várias áreas de saber, como por exemplo, a tecnologia da informação, estatística, linguística, cognição e outras, que se empenham para formação de um campo que se dedique a produção de conhecimento através de análises de fontes textuais.

Palavras-chave: Mineração de texto; Análise de dados; Transdisciplinaridade

INTRUDOÇÃO

O presente texto tem o intuito de explorar as potencialidades da mineração de textos na produção de conhecimento, torna-se necessário para tal empreitada, apresentar as transformações tecnológicas das últimas décadas, ressaltar a característica de nossa época de massiva produção de dados digitais e demonstrar que a emergência de ferramentas de mineração de textos só puderam emergir a partir da contribuição e diálogo de diferentes disciplinas. Para fundamentar a discussão foi realizada uma pesquisa bibliográfica sobre autores que trabalham a mineração de dados, mineração de textos e se dedicam a temática da transdisciplinaridade.

A evolução das tecnologias de informação e comunicação causaram uma revolução em diversas áreas da sociedade como, por exemplo, os meios de produção, as áreas de entretenimento, comunicação, educação e tantas outras. No meio dessa transformação tão intensa, se torna difícil identificar alguma área de conhecimento da

humanidade da qual não tenha sido modificada com a utilização dos novos dispositivos tecnológicos.

O campo da pesquisa foi um dos que se beneficiou imensamente das novas tecnologias de informação e comunicação. Dentre as mudanças ocorridas podemos citar: os computadores que tiveram sua potencialidade muito elevada permitindo a realização de cálculos em minutos que levariam anos em computadores mais antigos, surgimento de diversas ferramentas que permitem coletar e manipular áudio, foto e vídeo com maior qualidade e também podemos ver mudanças até em tarefas mais simples como softwares que gerenciam as referências bibliográficas para pesquisadores.

A ERA DOS DADOS DIGITAIS

Dentre as maiores transformações desse período recente da história da humanidade, e que vem impactando enormemente a produção de conhecimento, estão o nível gigantesco de dados produzidos anualmente e a maior facilidade de acesso à informação. A quantidade de dados digitais que vem sendo produzido nos últimos anos atingiu índices muito altos, “os dados digitais se expandem rapidamente – dobram em pouco mais de três anos” (MAYER-SCHONBERGER e CUKIER, 2013, p.4). E como a capacidade de armazenamento e processamento também aumentou junto, resta o dilema de como aproveitar melhor toda essa gama de informações disponíveis. Para a área da mineração de dados, e consequentemente, para a mineração de textos quanto mais dados significa maior probabilidade de conseguir conhecimento sobre esses dados, pois isto possibilita modelos de análise mais complexos.

O banco mundial através de seu site World Bank Open Data é um exemplo de instituição que permite acesso em seu site a dados de forma gratuita e livre pois acredita que podem ser de grande valia para pesquisadores de todo o mundo. O banco de dados consta com informações de diversos países, sobre diversos aspectos e indicadores ligados ao desenvolvimento dos mesmos, dados como: o crescimento populacional, PIB per capita, expectativa de vida e muitas outras informações do mundo inteiro estão disponíveis com apenas alguns cliques. A premissa de que através da disponibilização de informações podem propiciar conhecimento que sejam úteis para todos tem sido tão validada, que até mesmo empresas privadas como Uber e Waze estão liberando o acesso

aos dados que possuem sobre o deslocamento de veículos nas cidades do mundo, de forma anônima resguardando as informações dos usuários, mas que quando analisados e combinados, estas informações podem ser utilizadas para que pesquisadores possam produzir conhecimentos e soluções para o trânsito das cidades em que operam.

Voltado especificamente a área de textos, cabe ressaltar a existência do Gutenberg Project, este projeto, que possui financiamento através de doações, disponibiliza em seu site mais de 50.000 livros digitalizados por milhares de voluntários ao redor do mundo, os livros em sua maioria já se encontram em domínio público por isso podem ser distribuídos livremente, sendo assim uma excelente base de dados textuais que podem ser exploradas para fim de pesquisa.

A área comercial tem investido bastante e já tem se beneficiado da mineração de dados textuais. Suas aplicações são diversas podendo, por exemplo, avaliar o impacto de suas campanhas de marketing na internet analisando os comentários de seus clientes nas redes sociais, monitoramento da concorrência ou direcionar sua propaganda para algum tópico que esteja sendo mais debatido na rede.

MINERAÇÃO DE TEXTOS

A mineração de textos se insere dentro do campo da mineração de dados, que por sua vez, tem sua origem no campo do KDD - Knowledge Discovery in Databases (Descoberta do Conhecimento em Bases de Dados) que pode ser definido como: “o processo geral de conversão de dados brutos em informações úteis” (KUMAR; STEINBACH; TAN, 2009, p.4). O conceito de mineração de dados tem se tornado muito popular nos últimos anos, aqui temos uma definição bastante clara sobre o assunto:

> Citação Data Mining, ou mineração de dados, é o processo de descoberta de padrões e tendências existentes em repositórios de dados. Este processo visa basicamente à análise de grandes quantidades de dados com o objetivo principal de descoberta de conhecimento. (PINHEIRO, 2008, p.97)

Já a mineração de textos se limita na exploração do conhecimento em repositórios textuais. Hoje a maior parte dos dados produzidos no mundo se refere a dados não estruturados, especialmente textos. Para a computação o conceito de dados estruturados se refere à concepção de informação que seja encontrada de forma

organizada e classificada na maioria das vezes em formas de tabela. Já os dados não-estruturados, não se encontram tabulados, são as formas mais presentes e produzidas pela humanidade, que podem ser, por exemplo, áudio, imagens que precisam ser classificadas ou um texto sobre qualquer assunto. O processo de produzir conhecimento através de uma base de dados textuais demanda uma série de etapas como podemos observar nesse fragmento:

> O processo de mineração de textos consiste em um mecanismo de coleta de dados, uma etapa de pré-processamento e indexação, a qual classifica entidades no texto utilizando conhecimentos de linguística computacional para manipular palavras segundo sua classe gramatical, fazendo inferências sobre os limites de entidade e sua classificação, a aplicação de algoritmos e a análise dos resultados. (PINHEIRO, 2008, p.300)

Antes de compreender melhor as etapas da mineração de textos se torna de grande valia definir alguns conceitos básicos que reger o processo de aquisição de conhecimento. Quilici-Gonzalez e Zampirolli (2014) apresentam algumas definições importantes, o primeiro conceito é o de dado, que para os autores deve ser entendido como um fato registrado sem a necessidade de elaboração, outra importante conceituação é sobre informação, embora não haja um consenso estabelecido sobre ela, os autores marcam uma diferença em que a informação indica a existência de uma relação entre os dados produzindo algum significado. E por último eles consideram o conhecimento como um passo adiante, pois é compreendido como o resultado de uma análise de informações úteis a um propósito.

Pinheiro (2008) apresenta as etapas da descoberta do conhecimento em bases de dados que podem ser adaptadas para a mineração de textos. Segundo o autor deve iniciar-se com a compreensão dos objetivos e o conhecimento que se pretende adquirir ao final do processo, em seguida coletar os dados necessários a atingir os objetivos, no terceiro momento a limpeza e tratamento dos dados. No que diz respeito ao tratamento de dados no caso de textos podemos ver como exemplo desta etapa: correções ortográficas, adequação as normas ortográficas dos textos, corrigir as palavras que estão separadas por hífen em decorrência de translineação. Retomando as etapas, a quarta delas é reduzir os dados para poder focar nos que podem encaminhar ao conhecimento necessário, em

seguida a escolha do método de mineração que apresentará o resultado, e por fim, a sétima etapa consiste na interpretação dos resultados obtidos.

Existem diversos softwares capazes de minerar textos, tantos pagos quanto gratuitos e de código aberto. Dentre os pagos destacam-se o SAS Text Miner, Lexalytics, IBM Watson, WordStat e Sysomos, já entre as soluções gratuitas merecem menção OpenNLP, GATE, NLTK e AIKA. Estes softwares podem realizar diversas ações como, por exemplo, detecção de idioma, classificar, detectar frases, separar frases ou palavras, analisar as palavras semanticamente, identificar nomes próprios, calcular a frequência que uma palavra ocorre no texto e outras funcionalidades.

As consequências da utilização dessas ferramentas são muito interessantes, pois, as descobertas provenientes de procedimentos de mineração de textos podem produzir conhecimentos práticos com seus resultados diretamente ou ainda podem servir de alicerce para investigações de outros pesquisadores que não possuem acesso a todo material analisado ou conhecimento metodológico para realizar o processamento dos textos.

TRANSDISCIPLINARIDADE

O diálogo entre as disciplinas de áreas distintas tem sido um caminho seguido para a ciência uma vez que os problemas do mundo estão cada vez mais complexos, Bicalho e Oliveira (p. 89, 2011) afirmam essa condição: nesse fragmento:

> Para o desenvolvimento das novas disciplinas científicas tornou-se imprescindível utilizar abordagens e metodologias que possibilitem alcançar resultados decorrentes da interação com outras disciplinas, em diferentes níveis e formatos

É válido apontar o exemplo da Neurociência, que se estabelece como uma disciplina com grande reconhecimento de sua importância, e tem por sua constituição característica ser uma área de saber composta por diversas disciplinas do conhecimento: a Medicina, a Farmácia, a Química, a Psicologia, a Biologia entre outras, visto a complexidade do funcionamento do cérebro humano que necessita um diálogo que transcenda as barreiras de cada disciplina.

Bicalho e Oliveira (2011) afirmam que existem diversas nomenclaturas que descrevem as possíveis interseções entres campos do saber, entretanto, os mais

utilizados são os conceitos de interdisciplinaridade, multidisciplinaridade e transdisciplinaridade. Ainda segundo os autores o que difere essas três nomeações está relacionado aos níveis de integração entre as disciplinas, onde no campo multidisciplinar existe uma coordenação paralela dos pontos de vistas das áreas do saber, na interdisciplinaridade acontece uma maior integração que caminha para a convergência e no campo da transdisciplinaridade que é considerado o mais profundo onde há a fusão ou unificação das perspectivas.

Um dos grandes problemas de se tentar obter conhecimento através de bases textuais é devido a linguagem humana ter um alto grau de complexidade e a variação em consequência da criatividade humana. Bird, Klein e Loper (2009) apontam para os desafios de se trabalhar com o processamento de linguagem natural, pois diferente das linguagens artificiais como as de programação ou matemáticas, as linguagens naturais como o Inglês, Português e Hindi, evoluíram durante anos de gerações por gerações tornando muito difícil codificar todas as suas regras, entretanto as ferramentas possuem utilidades muito interessantes, pois, podemos analisar uma série de textos e obter informações que seria praticamente impossível de processar sozinho como qual a palavra que aparece com maior frequência, qual verbo foi mais utilizado, qual o nome próprio ou país que é mais citado ou comparar e identificar diferentes estilos de escrita.

Sendo assim, para dar conta da complexidade da obtenção de informações na mineração de textos é necessário recorrer a utilização de diversas áreas do campo da Tecnologia da Informação como: tecnologia de banco de dados, computação paralela, computação distribuída, reconhecimento de padrões, aprendizagem de máquina. Além disso, é necessário utilizar o conhecimento proveniente de outras áreas, segundo Aranha e Passos (2006) os saberes envolvidos na mineração de textos não se restringem somente a Informática, mas também atuam o campo da Estatística, Linguística e da Ciência Cognitiva.

CONHECIMENTO ALÉM DA LEITURA

Cabe ressaltar neste momento uma característica que a mineração de textos coloca em cena por proporcionar conhecimento do material textual que ultrapassa a barreira do conhecimento adquirido pela leitura. Através do processamento dos textos é

possível calcular informações em minutos que um indivíduo levaria muito tempo para realizar, como quais foram as palavras que mais apareceram num texto ou qual o sentimento (positivo, neutro ou negativo) do autor para cada frase. Estamos diante de uma grande mudança nas possibilidades de se extrair conhecimento de fontes textuais, como retrata o parágrafo a seguir:

> Foi criada uma nova disciplina acadêmica chamada 'Culturonomics': lexologia computacional que tenta compreender o comportamento humano e tendências culturais por meio da análise quantitativa de textos(MAYER-SCHONBERGER; CUKIER, 2013, p.58)

Um exemplo dessa nova cultura é a ferramenta Ngran Viewer que possibilita consultar uma palavra e com que frequência tem sido utilizada nos livros digitalizados pelo banco de dados do Google Books, permitindo o conhecimento que vem além daquele adquirido pela leitura, pois é possível observar tendências da cultura que são expressas através do uso das palavras. Um exemplo da possibilidade de aplicação desta ferramenta é acompanhar a utilização de palavras correlatas numa série temporal, como podemos observar na figura 1 – nós últimos anos do século XX vemos o crescimento da utilização da palavra "transdisciplinaridade":

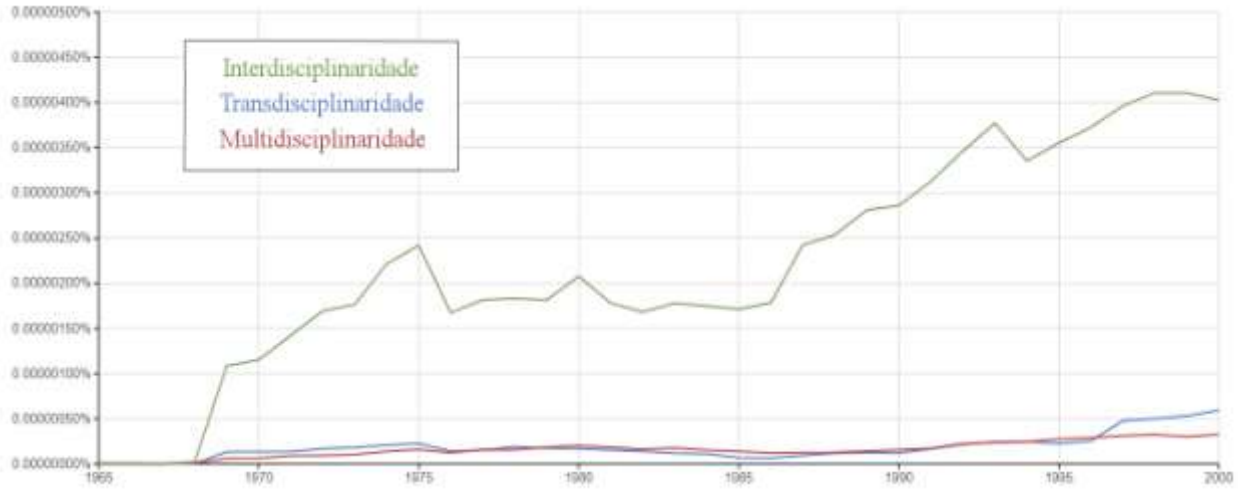

Figura 1 - Frequência da ocorrência das palavras interdisciplinaridade, multidisciplinaridade e transdisciplinaridade nos livros do Google Books de língua inglesa de 1965 a 2000.

Até o presente momento a ferramenta só disponibiliza dados até o ano 2000, mas é bastante plausível que essa tendência tenha continuado a crescer, justamente por assistirmos uma mudança na produção de conhecimento em que as áreas do saber tendem a dialogar mais na resolução de problemas. Os dados utilizados por esta ferramenta consistem no banco de livros digitalizados pela Google no idioma inglês, pois a plataforma ainda não liberou o conteúdo dos livros de língua portuguesa. Esta ferramenta demonstra claramente só existir devido à contribuição dos diversos campos de saberes, graças ao desenvolvimento técnico nos equipamentos de digitalização dos campos da engenharia e informática permitiram a digitalização de vários livros, e uma vez com os dados armazenados foi possível criar algoritmos que pudessem processar e quantificar a frequência das palavras com bases estatísticas.

CONSIDERAÇÕES FINAIS

Com o avanço observado nas últimas décadas, tanto na capacidade de processamento dos computadores quanto nas ferramentas disponíveis para mineração de textos, somado ao fato de estarmos vivendo numa época em que a quantidade de dados digitais produzidos e disponíveis para pesquisa é enorme, podemos esperar uma grande evolução na área de mineração de textos e uma grande produção científica utilizando esta metodologia. São inimagináveis as consequências que a mineração de texto no momento em que será passível de processar grande parte da comunicação interpessoal, o parágrafo a seguir retrata essa incapacidade de previsão:

> Com o estabelecimento da Internet como meio de disponibilização de informações, na forma de textos não estruturados, das mais variadas naturezas, é temeroso estabelecer os limites das aplicações da mineração de texto.(BEZERRA; GUIMARÃES, 2014)

Entretanto, há de convir, que por mais que os avanços sejam motivadores, ainda existe um grande lapso o qual requer muito desenvolvimento, tanto técnico quanto metodológico, para que as máquinas atinjam o patamar do ser humano no que se refere a compreensão textual, dada a complexidade da aprendizagem e o processo criativo humano.

REFERÊNCIAS BIBLIOGRÁFICAS

ARANHA, C.; PASSOS, E. **A Tecnologia de Mineração de Textos.** In: RESI-Revista Elerônica de Sistemas de Informação, FAECLA, Campo Largo, Paraná, Nº2, 2006.

BEZERRA, C. A.; GUIMARÃES, A. J.R. **Mineração de texto aplicada às publicações científicas sobre gestão do conhecimento no período de 2003 a 2012.** Perspectivas em Ciência da Informação, v. 19, n. 2, p. 131-146, 2014.

BICALHO, L, OLIVEIRA, M. **Aspectos conceituais da transdisciplinaridade e a pesquisa em ciência da informação.** Informação e sociedade: estudos, João Pessoa, v. 21, n. 2, p. 87-102, maio/ago. 2011.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the Natural Language Toolkit.** Sebastopol: O'Reilly, 2009.

KUMAR, V.; STEINBACH, M.; TAN P.N. **Introdução ao data mining: mineração de dados.** Rio de Janeiro: Editora Ciência Moderna, 2009.

MAYER-SCHONBERGER, V.; CUKIER, K.. **Big data: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana.** Rio de Janeiro: Elsevier, 2013.

PINHEIRO, C. **Inteligência analítica – mineração de dados e descoberta de conhecimento.** Rio de Janeiro: Editora Ciência Moderna, 2008.

QUILICI-GONZALEZ, J. A., ZAMPIROLLI, F. A. Sistemas **Inteligentes e Mineração de Dados.** Santo André: Triunfal Gráfica e Editora, 2014.

REFERÊNCIAS ICONOGRÁFICAS

Figura 1 – GOOGLE. Imagem adaptada de pesquisa realizada na ferramenta Ngram Viewer. Em: <https://books.google.com/ngrams>. Acesso em 10 agosto de 2017.

SOBRE O AUTOR:

Daniel Costa Vianna Mucicolo é mestre em Psicologia pela Universidade Federal Rural do Rio de Janeiro. Possui pós-graduação Latu Sensu em Psicologia Junguiana do Centro Universitário Hermínio da Silveira (Uni-IBMR). Possui titulação de Psicólogo e Bacharel em Psicologia pela Universidade Federal Fluminense (2011). Atualmente é docente no curso de Psicologia da Universidade do Contestado no Campus de Canoinhas/SC. Tem como área de interesse de pesquisa o campo da cibercultura e as potencialidades dos meios de tecnologia de informação e comunicação na interação humana e na produção de conhecimento.