UNIVERSIDADE FEDERAL DO CEARÁ
Campus de Sobral

Curso de Engenharia da Computação

Ken Esparta Ccorahua

EXTRAÇÃO, CLASSIFICAÇÃO E RECONHECIMENTO DE PADRÕES DE INFORMAÇÃO CONTIDA EM ARTIGOS ACADÊMICOS UTILIZANDO MINERAÇÃO WEB DE CONTEÚDO E MINERAÇÃO WEB DE USO

Sobral – CE

2016

Ken Esparta Ccorahua

EXTRAÇÃO, CLASSIFICAÇÃO E RECONHECIMENTO DE PADRÕES DE INFORMAÇÃO CONTIDA EM ARTIGOS ACADÊMICOS UTILIZANDO MINERAÇÃO WEB DE CONTEÚDO E MINERAÇÃO WEB DE USO

Memorial de Monografia apresentado à disciplina de Seminário de Monografia do Curso Engenharia da Computação como requisito parcial para obtenção do grau de Engenheiro da Computação na Universidade Federal do Ceará, *campus* Sobral.

Orientador

Prof. Me. Fernando Rodrigues de Almeida Júnior

Coorientador

Prof. Dr. Márcio André Baima Amora

Sobral − CE, julho de 2016

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A World Wide Web (WWW) ou comumente denominada Web, vem crescendo nos últimos vinte anos. Aplicações Web tornam-se populares e um dos motivos é que o navegador permite executá-las a partir de qualquer tipo de dispositivo, em outras palavras, o sistema operacional não influencia a aplicação no tempo de execução.

Tanto dados estruturados como dados não estruturados trafegam pela Web. O próprio documento HTML renderizado no navegador é uma fonte de dados tais como endereços Web, informações pessoais, meta-tags, folhas de estilos, arquivos XML, arquivos JSON, entre outros. Não obstante, os dados que estão contidos em um documento HTML possuem informações relevantes que podem ser mineradas. É nesse contexto que se aplica o processo de mineração de dados.

Existem também os Web logs, que são as informações armazenadas através do tempo relacionadas ao servidor que estão em ordem cronologicamente descendente. Para minerar os dados no Web log, utiliza-se a mineração Web de uso. No presente trabalho utiliza-se o Web log de artigos da Wikipédia.

A teoria apresentada no trabalho realiza-se através de uma rede de Petri, de modo que os estados são as áreas da Ciência da Computação a serem atingidas e as transições são as técnicas e o embasamento teórico requeridos para passar ao estado seguinte. Começa-se pela Soft Computing (SC), que é uma área das Ciências da Computação que permite resolver problemas com uma abordagem orientada à tolerância de imprecisões. Em seguida, aborda-se a teoria relacionada com o aprendizado de máquina, que, juntamente com as técnicas de descobrimento de conhecimento, dão origem às técnicas de mineração de dados. Finalmente, apresenta-se a mineração Web de conteúdo (ou Web Content Mining) e mineração Web de uso (ou Web Usage Mining), cujas técnicas são utilizadas no trabalho.

Existe o problema da desconfiança do conteúdo nos artigos da Wikipédia e também de outros na internet. A proposta neste trabalho é o desenvolvimento de uma aplicação Web de busca na internet que tem como objetivo principal a mineração de informação em artigos acadêmicos se baseando em Web logs da Wikipédia e, finalmente, mostrando como resultado ao usuário endereços Web de artigos de confiança.

A mineração será dada em dois níveis. No primeiro nível tem-se a mineração de uso nos Web logs de artigos da Wikipédia e no segundo nível, a mineração de conteúdo de

outros sites nos quais também são oferecidos artigos acadêmicos. O resultado da pesquisa que o usuário faz na aplicação Web final é a junção da mineração de uso de Web logs da Wikipédia e a mineração de conteúdo de artigos acadêmicos externos a ela.

2 Objetivo

2.1 Objetivo geral

Extrair, classificar e reconhecer padrões nas informações contidas em artigos acadêmicos, a partir de uma busca efetuada pelo usuário do sistema proposto no trabalho, tendo como base os Web logs da Wikipédia e utilizando técnicas de mineração Web de conteúdo e mineração Web de uso.

2.2 Objetivos específicos

- Mostrar ao usuário final que os resultados da busca efetuada são artigos de confiança do ponto de vista acadêmico.
- Utilizar técnicas de classificação de conteúdo de artigos Web da Wikipédia no aprimoramento da utilização de algoritmos de aprendizado de máquina.
- Utilizar técnicas de aprendizado de máquina para reconhecer padrões a partir de logs gerados pelos servidores da Wikipédia.
- Aplicar a fundamentação teórica relacionada com redes neurais para ajudar a geração de conhecimento dos dados extraídos de artigos acadêmicos.
- Implementar o processo de Mineração de dados CRISP-DM, proposto em (CHAPMAN et al., 2000), para gerar um modelo sólido que garanta o funcionamento da mineração de conteúdo e de uso nas páginas da Wikipédia.
- Mostrar que a Web é uma poderosa ferramenta para adquirir conhecimento relevante relacionado com artigos acadêmicos de sites semelhantes à Wikipédia.

3 Justificativa

As aplicações Web estão sendo amplamente utilizadas pelas empresas principalmente porque são multiplataforma e podem ser acessadas a partir de qualquer lugar do mundo

desde que exista uma conexão com a internet. Computação na nuvem como: Amazon Web Service, Heroku, Digital Ocean, entre outras, estão crescendo no mercado. O próprio trabalho utiliza os serviços da Digital Ocean para alocar os códigos de teste e do produto final.

Assim como as aplicações Web estão crescendo, o mercado relacionado com comércio eletrônico também. É imprescindível que uma empresa que vende produtos e serviços através de uma aplicação Web tenha dados das atividades que seus clientes realizam quando usam o aplicativo e isso pode ser feito utilizando a mineração Web de uso. Por outro lado, é necessário também que a empresa obtenha informação a partir de aplicações Web de outras empresas concorrentes para oferecer melhores ofertas aos clientes. Essa obtenção de informação é realizada utilizando técnicas de mineração Web de conteúdo.

A lógica proposta no parágrafo anterior aplica-se também no trabalho, porém não possui abordagem comercial. O usuário final busca artigos utilizando o sistema Web, que é proposto no trabalho, o qual indexa outras páginas de artigos Web relacionadas com a busca utilizando mineração Web de conteúdo. O resultado é uma busca que retorna um conjunto de endereços Web com artigos de confiança considerando, além dos Web logs da Wikipédia, a informação de outros artigos acadêmicos.

4 Trabalhos relacionados

No artigo acadêmico intitulado *"E-commerce Web page classification based on automatic content extraction"* referenciado em (PETPRASIT; JAIYEN, 2015), propõe-se o método Markov Random Field para reduzir o número de caraterísticas no conjunto de dados que foram extraídos da Web utilizando o método Subject Detection and Density. Em seguida, faz-se uma comparação de performance entre as redes neurais Radial Basis Function (RBF) e Support Vector Machine com o método Naive Bayes. Utiliza-se, na aplicação Web proposta, a rede neural RBF para poder minerar os web logs da Wikipédia. Não se tem uma visão definida sobre a rede neural a ser projetada porque o problema a ser resolvido não foi analisado com detalhe.

No artigo acadêmico intitulado *"E-commerce website ranking using Semantic Web mining and neural computing"* referenciado em (VERMA et al., 2015), utilizam-se redes neurais, mineração Web de uso e de conteúdo para desenvolver uma aplicação Web de determinação de prioridade de diferentes páginas Web de comércio eletrônico. A implementação das técnicas de mineração Web de uso e de conteúdo no presente trabalho, são inspiradas nas técnicas propostas neste artigo.

No artigo acadêmico intitulado *"A framework for building Web mining applications in the world of blogs: A case study in product sentiment analysis"* referenciado em (COSTA et al., 2012), desenvolve-se um framework desenvolvido na linguagem Java™ que utiliza os serviços da mineração Web semântica para ajudar a usuários de blogs na busca efetiva de informação relevante no mundo dos blogs. Esse framework utiliza uma interface de aplicação, um controlador de mineração de dados na Web (Web *crawler*) e um extrator de textos. No presente trabalho, o Web *crawler* e o extrator de textos serão implementados. Por outra parte, a interface de usuário será desenvolvida para que seja executada na Web.

No artigo acadêmico intitulado *"Web Classification Mining Based on Radial Basis Probabilistic Neural Network"* referenciado em (GAO; TIAN, 2009), utiliza-se o método K-Nearest Neighbor (KNN) modificado para treinar a rede neural do tipo Radial Basis Probabilistic; os dados utilizados para classificação e treinamento foram extraídos de mil páginas Web utilizando a mineração Web de conteúdo. Será utilizado, no presente trabalho, o método KNN para poder agrupar os dados dos Web logs da Wikipédia.

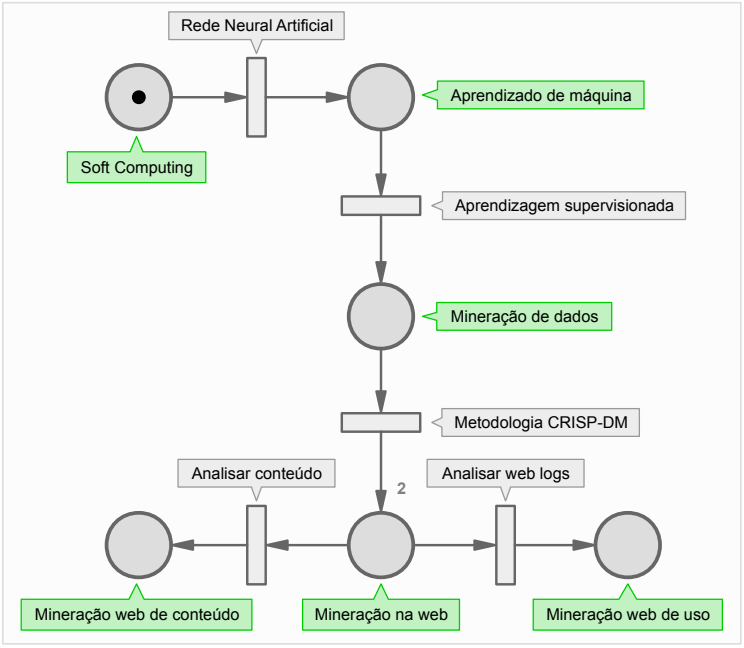
No artigo acadêmico intitulado *"Structured Web information extraction using repetitive subject pattern"* referenciado em (THAMVISET; WONGTHANAVASU, 2012) propõe-se uma técnica semi supervisionada que cria um padrão de extração de dados de páginas Web de comércio eletrônico. O sistema é composto por um extrator de dados e um wrapper que traduz os conteúdos da Web para tabelas do tipo relacionais. O extrator de dados das páginas Web do presente trabalho será inspirado no extrator deste artigo.

5 Fundamentação teórica

Apresentam-se breves definições sobre os assuntos teóricos relacionados com a monografia, utiliza-se uma abordagem *top-down* como pode-se observar na (FIGURA 5.1), mostra-se uma rede de Petri que representa a montagem teórica do presente trabalho, na qual o estado inicial é o mais genérico e o estado final é o mais específico.

Na rede de Petri proposta, observa-se também que cada estado representa uma área das Ciências da Computação a qual possui ferramentas teóricas as quais serão adotadas para que a ficha posa ativar as transições. Cada transição representa uma ferramenta que será adquirida do estado anterior antes de passar ao próximo. Em outras palavras, a ficha pode passar somente ao estado seguinte quando uma ferramenta específica do estado anterior é definida para ser utilizada.

Figura 5.1: *Estruturação do conhecimento para conseguir as ferramentas teóricas e técnicas relacionadas com o presente trabalho.*



Fonte: Autoria própria.

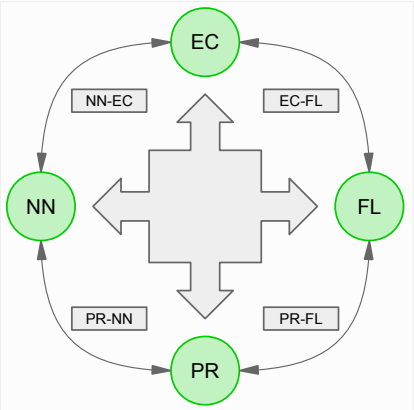
5.1 Soft computing

Existem dois tipos de paradigmas computacionais, a Soft Computing (SC) e a Hard Computing (HC), ambos os termos foram estabelecidos pela primeira vez pelo professor L. A. Zadeh no ano 1996. O HC trata sobre modelos precisos em que as soluções são atingidas imediatamente, por outra parte, a SC lida com modelos de aproximações e dá soluções a problemas complexos (SIVANANDAM; DEEPA, 2004). Como pode-se perceber, a HC é basicamente a computação convencional de modo que a solução dos problemas baseia-se nos princípios da precisão. Em contraste, o paradigma da SC trata sobre solucionar problemas utilizando modelos imprecisos que possuem certa porcentagem de aproximação.

Cataloga-se a SC, em (MAIMON; ROKACH, 2007), como uma coleção de novas técnicas em inteligência artificial que exploram a tolerância para a imprecisão, incerteza, verdade parcial e manipulação de não linearidades para poder alcançar rastreabilidade, robustez e soluções de menor custo comparado com os métodos da HC, ou seja apresenta-se uma coleção de ferramentas aptas para minerar a Web porque esta encaixa-se nas definições de imprecisão, incerteza e veracidade duvidosa.

A SC está composta por técnicas como a rede neural (Neural Network), computação evolucionária (Evolutionary Computing), sistemas difusos (Fuzzy Systems) e raciocínio probabilístico (Probabilistic Reasoning), como mostrado na (FIGURA 5.2).

Figura 5.2: *Principais componentes da família da SC.*

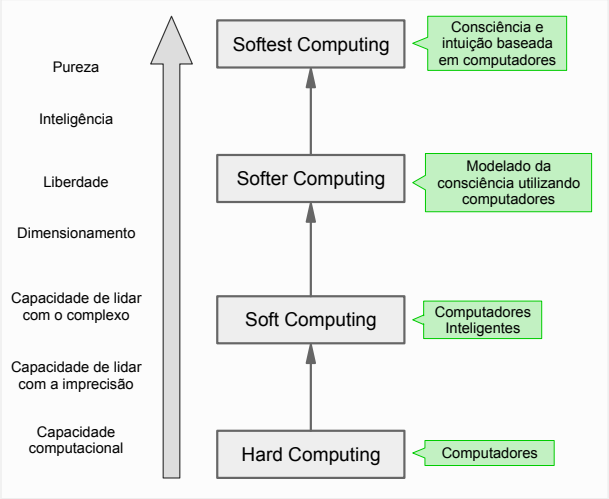


Fonte: Adaptado de (MANKAD, 2013).

As técnicas mostradas na (FIGURA 5.2), nem sempre estão isoladas. Elas podem ser aplicadas ao mesmo tempo para resolver um mesmo problema, e isso denomina-se hibridação.

Em (ZADEH, 1996) define-se a SC como uma nova abordagem da computação a qual é análoga à habilidade marcante da mente humana para pensar e aprender sobre um entorno incerto e impreciso. A SC, segundo (CHATURVEDI, 2008), é uma vertente da computação na qual pretende-se construir máquinas inteligentes e sábias. Pode-se observar na (FIGURA 5.3) o desenvolvimento da SC, começando pela computação convencional, até chegar ao extremo de pensamento puro por parte de um computador, ou seja, o objetivo final da SC é projetar e desenvolver um computador que possa agir de forma semelhante aos seres humanos.

Figura 5.3: *Desenvolvimento da SC.*



Fonte: Adaptado de (CHATURVEDI, 2008).

Para conseguir ativar a transição na rede de Petri da (FIGURA 5.1) é necessário estabelecer a rede neural como técnica para poder passar ao estado seguinte que é o aprendizado

de máquina.

Uma rede neural artificial (ou simplesmente rede neural) é um modelo matemático simplificado de uma rede de neurônios biológicos. A unidade fundamental de uma rede neural é o neurônio que está interconectado com outros. Essas redes executam em paralelo uma tarefa global comum e possuem a habilidade de aprendizagem. Explica-se em (MANKAD, 2013) que uma consequência da aprendizagem dos neurônios de uma rede neural é a aquisição de conhecimento de modo a torná-lo disponível para o uso. As caraterísticas básicas de uma rede neural são o paralelismo inerente, acesso à informação local, a semelhança entre suas componentes e o aprendizado incremental.

5.2 Aprendizado de máquina

A ficha encontra-se no seguinte estado, contendo já a teoria de uma rede neural. Nesta seção trata-se sobre as técnicas que serão adotadas do aprendizado de máquina (Machine Learning).

O aprendizado automático ou aprendizado de máquina é um programa de computador que pode "aprender" de um conjunto de entradas disponíveis. Define-se, em (MURPHY, 2012), que o aprendizado de máquina é um conjunto de métodos que podem detectar automaticamente padrões em um determinado conjunto de dados e, posteriormente, utilizar esse padrões descobertos para predizer dados futuros. Aprendizado é, grosseiramente falando, o processo de converter experiência em habilidade ou conhecimento (SHALEV-SHWARTZ; BEN-DAVID, 2014). Aprender a partir de dados é o conceito fundamental do ML.

O foco do ML é a modelagem do aprendizado e a adaptação (atividades de animais e humanos) num computador. Os métodos do ML são denominados "sub-simbólicos" porque não existem símbolos ou manipulação deles envolvidos, em contraste com a Inteligência Artificial, em que o computador manipula símbolos que refletem o entorno (processo simbólico) (MARSLAND, 2015).

Tomando como ponto de partida que as máquinas aprendem a partir de dados, então, em (MARSLAND, 2015) a ML trata sobre como fazer computadores modificarem ou adaptarem as suas ações de modo que estas se tornem mais precisas, a precisão é medida por quão bem a escolha de ações refletem as escolhas corretas.

Em que momento precisa-se do ML?, responder esta pergunta é fundamental porque justifica a utilização dos conceitos de ML no trabalho. A resposta é dada em (SHALEV-

SHWARTZ; BEN-DAVID, 2014), sugere-se que os conceitos de ML devem ser utilizados em tarefas realizadas por humanos ou animais. De acordo com isso, precisa-se do ML para poder modelar os algoritmos deste trabalho; de modo que utilizam-se dados gerados pela interação humana com a Web a partir de buscas.

De acordo com (MURPHY, 2012), o aprendizado de máquina divide-se, usualmente, em aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. Na abordagem supervisionada ou preditiva do ML, tem-se como objetivo o aprendizado mapeando as saídas a partir das entradas, dado um conjunto de treinamento. A segunda abordagem de ML, a não supervisionada ou descritiva, tem como objetivo encontrar padrões e correlações entre os dados. Finalmente, a abordagem de aprendizado por reforço, trata-se do modo de agir ou se comportar diante uma situação de recompensa ou de punição.

A atividade de classificar e extrair conhecimento a partir do conteúdo de páginas Web da Wikipédia justifica a escolha da utilização das ferramentas oferecidas pela abordagem supervisionada do aprendizado de máquina. Então, a ficha da rede de Petri passa para o seguinte estado sabendo que até esta subseção, juntando os conceitos revisados, está sendo utilizada uma rede neural com aprendizagem supervisionada.

5.3 Mineração de dados

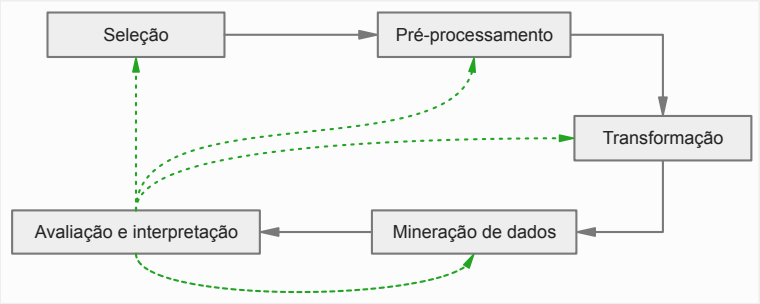
A mineração de dados (data mining) é definida, segundo (ZAKI; JR., 2014), como o processo de descoberta intuitiva de novos padrões de interesse, assim como modelos preditivos, descritivos e compreensíveis.

Em (RAJARAMAN; LESKOVEC; ULLMAN, 2012) descreve-se a mineração de dados como a descoberta de modelos a partir de dados. Existem duas abordagens: a estatística e a do aprendizado de máquina. Para o presente trabalho utiliza-se a abordagem de aprendizado de máquina, esta decisão justifica-se na (SUBSEÇÃO 5.2).

Uma analogia da mineração de dados faz-se, em (HAN; PEI, 2012), com a obtenção de ouro das rochas em uma mina. Ademais, enfatiza-se que o nome apropriado para mineração de dados é “conhecimento minerado a partir de dados”. Segundo (HAN; PEI, 2012), a descoberta de conhecimento a partir de dados (ou KDD do inglês knowledge discovery from data) é sinônimo de mineração de dados, não obstante em (SHAFIQUE; QAISER, 2014) coloca-se o KDD, juntamente com o CRISP-DM e o SEMMA, como um tipo de processo relacionado com a mineração de dados, seguidamente, expõe-se uma breve descrição sobre eles:

KDD É um modelo de processo que consiste na extração de conhecimentos escondidos a partir de um banco de dados. Deve ser um processo interativo e iterativo, possui cinco passos de desenvolvimento que se podem observar na (FIGURA 5.4).

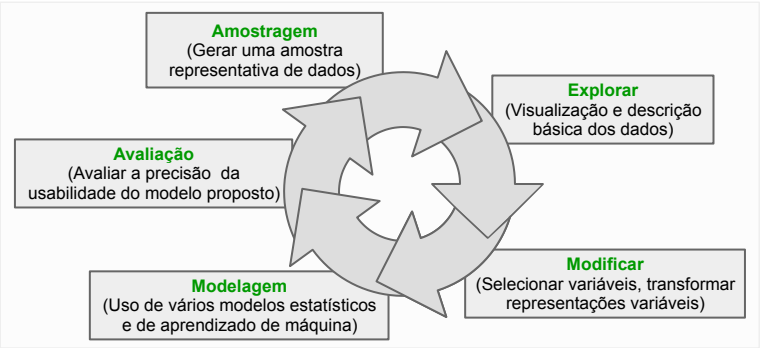
Figura 5.4: *Uma visão geral dos passos que compõem o processo KDD.*



Fonte: Adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

SEMMA Do acrônimo Sample, Explore, Modify, Model and Assess; que foi desenvolvido pelo instituto SAS[1]. Foca-se basicamente no desenvolvimento e manutenção de projetos de mineração de dados. Consiste em um ciclo altamente iterativo de cinco passos como pode-se verificar na (FIGURA 5.5).

Figura 5.5: *Metodologia SEMMA proposto pela SAS.*



Fonte: Adaptado de (BINUS, 2014).

CRISP-DM É um processo de mineração de dados que significa Cross-Industry Standard Process for Data Mining. Segundo (CHAPMAN et al., 2000), o processo foi concebido no ano 1996 pelas empresas DaimlerChrysler[2], SPSS[3] e NCR[4]. É uma

[1] Statistical Analysis System, é o nome de uma empresa pioneira em Business intelligence. Disponível em <http://www.sas.com>

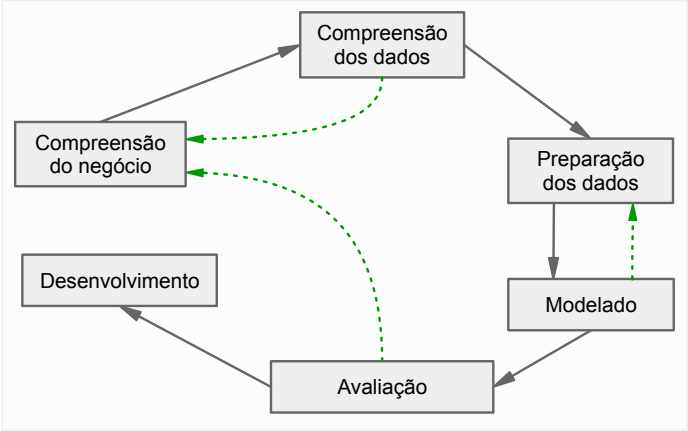
[2] É um fabricante de automóveis de passageiros e veículos comerciais. Disponível em <https://www.daimler.com>

[3] É um software para análise estatística de dados, disponível em <http://www.ibm.com/analytics/us/en/technology/spss/>

[4] É uma empresa de tecnologia especializada em produtos para o varejo e setores financeiros, disponível em <http://www.ncr.com>

ferramenta que está constituída por seis processos cíclicos como se mostra na (FIGURA 5.6). Segundo a pesquisa de (PIATETSKY, 2014) no site KDnuggets, pode-se observar na (TABELA 5.1) que o CRISP-DM é o método mais utilizado nos anos 2007 e 2014.

Figura 5.6: *Metodologia CRISP-DM, utilizada no presente trabalho.*



Fonte: Adaptado de (OLSON; DELEN, 2008).

Para passar ao seguinte estado da rede de Petri estabelecida, veja-se (FIGURA 5.1), é necessário adotar um tipo de processo de mineração de dados. Como é mencionado em (OLSON; DELEN, 2008), nas três metodologias que foram abarcadas, não há obrigação de seguir de forma rígida seus respectivos passos. Afinal, optou-se por seguir a metodologia CRISP-DM. Na (SEÇÃO 6) vê-se a aplicação desta metodologia a partir do ponto de vista da proposta do trabalho.

Tabela 5.1: *Pesquisa realizada com 200 pessoas sobre qual é a metodologia favorita utilizada nos projetos de mineração de dados, nos anos 2007 e 2014.*

Metodologia	**Ano 2007 (%)**	**Ano 2014 (%)**
CRISP-DM	42.0	43.0
Própria	19.0	27.5
SEMMA	13.0	8.5
Outra	4.0	8.0
Processo KDD	7.3	7.5
Minha organização	5.3	3.5
Específica de domínio	4.7	2.0
Nenhuma	0.0	4.7

Fonte: Adaptado de (PIATETSKY, 2014).

5.4 Mineração na Web

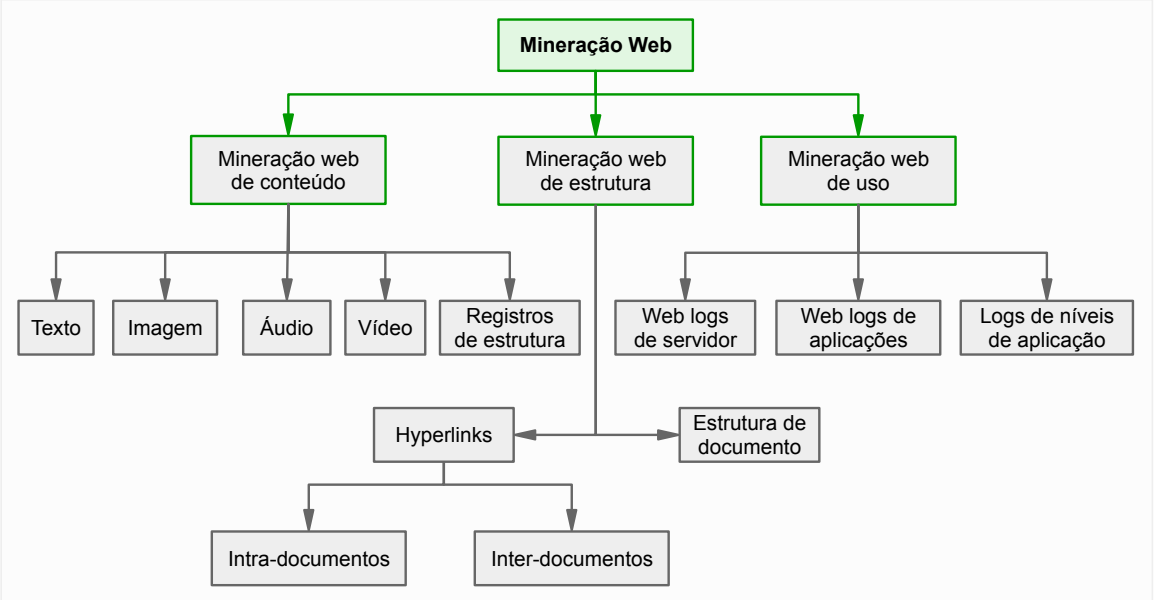
A mineração na Web ou Web mining, é um conjunto de ferramentas adicionais à mineração de dados que, como o nome indica, permitem a descoberta de conhecimento utilizando recursos obtidos da Web. Em (LIU, 2011) explica-se que a mineração a na Web é descobrir informação útil e relevante a partir de estruturas de hiper-referências, conteúdo das páginas e dados de uso de clientes. Justamente essas três fontes de extração de informação fazem referência às três categorias de mineração na Web mostradas a seguir, veja-se na (FIGURA 5.7):

Conteúdo Ou Web content mining (WUM), é a extração de informação valiosa a partir de informação alocada em páginas Web.

Estrutura Ou Web structure mining (WSM), nesta técnica considera-se uma página Web como um nó e as referências entre elas como as arestas que as juntam.

Uso Ou Web usage mining (WUM), refere-se ao descobrimento de padrões de acessos do usuário ao servidor, ou seja extração de padrões de Web logs.

Figura 5.7: *Taxonomia da mineração na Web.*



Fonte: Adaptado de (PATEL; CHAUHAN; PATEL, 2011).

Pretende-se utilizar as ferramentas fornecidas pela mineração Web de conteúdo e mineração Web de uso. Esta escolha deve-se a dois fatores; o primeiro é a utilização de Web logs (obtidos de `<https://dumps.wikimedia.org/other/pagecounts-raw/>`) para melhorar a experiência do usuário na busca por conteúdo na Wikipédia, e o segundo fator,

é a utilização de métodos de Web scraping que se aplicam em páginas de artigos acadêmicos semelhantes à Wikipédia. Web scraping é um conjunto de técnicas computacionais que são utilizadas para obter informação diretamente de páginas Web utilizando requisições HTTP.

Finalmente, chegando neste ponto, a estruturação teórica pára e em consequência uma ficha da rede de Petri fica no estado da mineração Web de conteúdo e a outra, na mineração Web de uso. Com a parte teórica consolidada, na subsecção seguinte, explica-se sobre o software que é utilizado na parte prática da aplicação Web proposta no trabalho.

5.5 Bibliotecas de Python utilizadas

Esta secção não tem como intenção explicar sobre as funcionalidades e recursos básicos linguagem Python, senão, a intenção é expor sobre as bibliotecas necessárias que se utilizam para conseguir o objetivo estabelecido na presente monografia.

Scrapy, Requests e Beautiful Soup são três bibliotecas open source utilizadas para realizar a descoberta de informação na Web de artigos relacionados com a Wikipédia, um resumo sobre a funcionalidade dessas bibliotecas são mostradas seguidamente:

Scrapy É um framework escrito em Python para fazer rastreadores Web (Web crawling) mediante a criação de "spiders" que são robôs que coletam informação de uma determinada página Web.

Requests É uma biblioteca HTTP para Python, que faz requisições Web e extração de conteúdo de uma página Web, essa extração é direta, ou seja não existem filtros de tags. Para a filtragem utiliza-se a biblioteca Beautifull Soup.

Beautiful Soup É uma biblioteca de Python que serve, principalmente, para extrair dados de documentos HTML e XML. A extração de dados é realizada utilizando expressões regulares.

Para a etapa de análise dos dados coletados de Web logs da Wikipédia, veja-se na (FIGURA 5.8), utilizam-se as seguintes bibliotecas open source: Scipy, Numpy, Matplotlib e Jupyter. Uma breve explicação sobre o funcionamento daquelas bibliotecas é exposta em seguida:

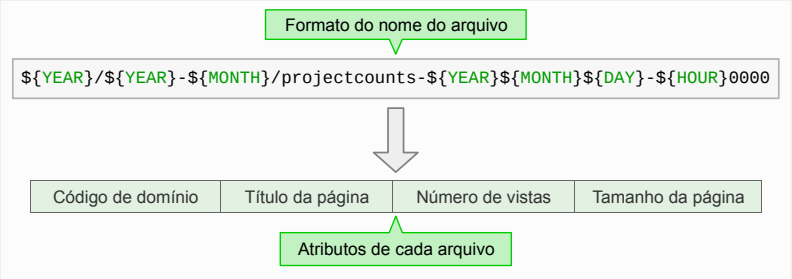
SciPy Coleção de ferramentas programadas em Python focadas para cientistas da computação e analistas.

NumPy Ferramentas oferecidas para resolver problemas computacionais relacionados com álgebra linear e transformadas de Fourier.

Matplotlib É uma biblioteca de Python para fazer gráficos em duas dimensões. Utiliza-se geramente com a feramenta Jupyter para aplicações Web relacionadas com a produção de gráficos semelhantes à Mathematica®

Jupyter É uma aplicação Web que permite criar e compartilhar documentos que tenham código, equações, visualizações e textos explicativos. Inclui ferramentas como limpeza de dados, simulações numéricas, modelagem estatística e aprendizado de máquina.

Figura 5.8: *Formato geral de um arquivo de Web log obtido da Wikipédia.*



Fonte: Autoria própria.

O resultado final é uma plataforma Web produzida inteiramente na linguagem Python utilizando o framework Django. São implementadas, na plataforma, as rotinas para a obtenção de dados, pré-processamento, agrupamento, análise, obtenção de conhecimento e reconhecimento de padrões de artigos relacionados com a Wikipédia e os Web logs dos servidores desta.

6 Materiais e métodos

Como foi exposto na (SUBSEÇÃO 5.5), utiliza-se a linguagem Python (versão 2.7) com as suas respectivas bibliotecas para realizar todas as tarefas descritas na secção 4.5. O programa final produzido será executado num servidor Web fornecido por Digital Ocean (serviço de computação na nuvem) com a interação do usuário final.

Mostram-se, como explicado na (SUBSEÇÃO 5.3), as etapas do desenvolvimento do trabalho utilizando a metodologia CRISP-DM:

Compreensão do negócio Deseja-se aprimorar as buscas de artigos acadêmicos dos usuários baseando-se na mineração de Web logs da wikipédia e na mineração de páginas Web

relacionadas. Utiliza-se a técnica de mineração de dados porque as informações das páginas Web maiormente tem informações irrelevantes.

Compreensão dos dados Os dados vem de duas fontes: Web logs da Wikipédia, que são dados homogêneos e da mineração Web de conteúdo de artigos acadêmicos, que são dados heterogêneos.

Preparação dos dados Aplicam-se técnicas de filtragem nos bancos de dados de Web logs da Wikipédia e os que foram obtidos a través da mineração Web de modo que o "ruido" não desejado é eliminado. Também aplicam-se técnicas de agrupamento de dados, tudo isso é para poder otimizar o treinamento da rede neural que é modelada no seguinte passo.

Modelagem Nesta etapa, projeta-se uma rede neural adequada ao problema a ser resolvido. A rede neural é do tipo RBF (Radial Basis Function) e é implementada tendo como referência o artigo (PETPRASIT; JAIYEN, 2015).

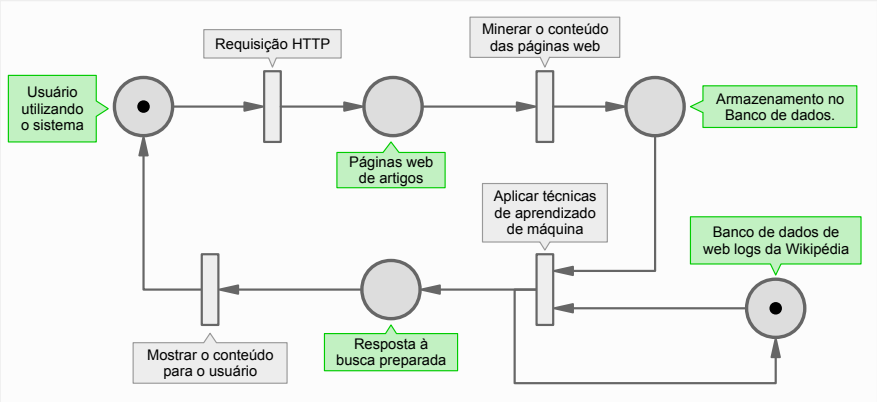
Avaliação Esta etapa realiza-se juntamente com a da Modelagem, é dizer, enquanto se projeta uma rede neural ela é testada para verificar o cumprimento dos requisitos de solução do problema.

Desenvolvimento Finalmente, o desenvolvimento da aplicação Web tendo em conta o anterior mencionado. Destacando que o desenvolvimento realiza-se utilizando a técnica TDD.

Depois de haver estabelecido a metodologia, estabelecem-se as atividades da aplicação Web proposta no trabalho, veja-se na (FIGURA 6.1). Primeiramente, o usuário faz uma requisição HTTP no momento de fazer uma busca usando o sistema Web proposto, essa ação ativará a busca em vários sites (relacionados com artigos acadêmicos), essa busca é realizada utilizando os métodos de Web scraping. Para extrair dados da Web são utilizados os algoritmos Subject Detection e Node Density propostos em (PETPRASIT; JAIYEN, 2015), em seguida, aplica-se a técnica de mineração Web de conteúdo. Os dados obtidos a partir da mineração Web são armazenados em um banco de dados não relacional (neste caso MongoDB) e analisam-se, juntamente com os dados dos Web logs da Wikipédia, utilizando uma técnica de rede neural com aprendizagem supervisionada. Finalmente, a resposta preparada é mostrada para o usuá-

rio, dita resposta contém links de páginas Web de artigos acadêmicos que estão relacionados com as palavras-chave da busca.

Figura 6.1: *Atividades da aplicação Web proposta no trabalho.*



Fonte: Autoria própria.

Na (TABELA 6.1) mostra-se uma previsão de tempo para poder realizar todas as tarefas. Lembrando que o processo CRISP-DM, veja-se (FIGURA 5.6), é cíclico e iterativo.

Tabela 6.1: *Cronograma que se segue no trabalho de conclusão do curso.*

Atividade	**Data**
Compreensão do negócio	01/06/2016 ⟶ 14/07/2016
Compreensão dos dados	15/07/2016 ⟶ 02/08/2016
Preparação dos dados	03/08/2016 ⟶ 14/08/2016
Modelado, Avaliação, desenvolvimento e TDD	15/08/2016 ⟶ 15/09/2016

Fonte: Autoria própria.

Referências

BINUS. *Processes in Data Mining*. 2014. Disponível em: <http://sisbinus.blogspot.com.br/2014/11/processes-in-data-mining.html>. Acesso em: 10 jun. 2016.

CHAPMAN, P. et al. *CRISP-DM 1.0: Step-by-step data mining guide*. [S.l.], 2000.

CHATURVEDI, D. K. *Soft Computing: Techniques and its Applications in Electrical Engineering*. [S.l.]: Springer, 2008.

COSTA, E. et al. A framework for building web mining applications in the world of blogs: A case study in product sentiment analysis. *Expert Systems with Applications*, 2012.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI MAGAZINE*, 1996.

GAO, M.; TIAN, J. Web classification mining based on radial basis probabilistic neural network. *International Workshop on Database Technology and Applications*, 2009.

HAN, J.; PEI, M. K. J. *Data Mining Concepts and Techniques*. 3rd. ed. [S.l.]: ELSEVIER, 2012.

LIU, B. *Web Data Mining*. [S.l.]: Springer, 2011.

MAIMON, O.; ROKACH, L. *Soft Computing for Knowledge Discovery and Data Mining*. [S.l.]: Springer, 2007.

MANKAD, K. B. *A Genetic-Fuzzy Approach to Measure Multiple Intelligence*. Dissertação (Mestrado) — Sardar Patel University, 2013.

MARSLAND, S. *Machine Learning: An Algorithmic Perspective*. [S.l.]: CRC Press, 2015.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. [S.l.]: Massachusetts Institute of Technology, 2012.

OLSON, D. L.; DELEN, D. *Advanced Data Mining Techniques*. [S.l.]: Springer, 2008.

PATEL, K. B.; CHAUHAN, J. A.; PATEL, J. D. Web mining in e-commerce : Pattern discovery , issues and applications. *International Journal of P2P Network Trends and Technology-*, 2011.

PETPRASIT, W.; JAIYEN, S. E-commerce web page classification based on automatic content extraction. *International Joint Conference on Computer Science and Software Engineering*, 2015.

PIATETSKY, G. *CRISP-DM, still the top methodology for analytics, data mining or data science projects*. 2014. Disponível em: <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. Acesso em: 10 jun. 2016.

RAJARAMAN, A.; LESKOVEC, J.; ULLMAN, J. D. *Mining of Massive Datasets*. [S.l.]: Stanford University, 2012.

SHAFIQUE, U.; QAISER, H. A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 2014.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding Machine Learning: From Theory to Algorithms*. [S.l.]: Cambridge University Press, 2014.

SIVANANDAM, S. N.; DEEPA, S. N. *Principles of S.oft Computing*. [S.l.]: Wiley, 2004.

THAMVISET, W.; WONGTHANAVASU, S. Structured web information extraction using repetitive subject pattern. *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2012.

VERMA, N. et al. E-commerce website ranking using semantic web mining and neural computing. *Procedia Computer Science*, 2015.

ZADEH, L. A. The roles of soft computing and fuzzy logic in the conception, design and deployment of intelligent system. *IEEE Intelligent Systems*, 1996.

ZAKI, M. J.; JR., W. M. *Data Mining and Analysis Fundamental Concepts and Algorithms*. [S.l.]: Cambridge University Press, 2014.