UNIVERSIDADE FEDERAL DO CEARÁ
Campus de Sobral

Curso de Engenharia da Computação

Ken Esparta Ccorahua

EXTRAÇÃO, CLASSIFICAÇÃO E RECONHECIMENTO DE PADRÕES DE INFORMAÇÕES CONTIDAS 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 INFORMAÇÕES CONTIDAS 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

Lista de abreviaturas e siglas

CRISP-DM	Cross-Industry Standard Process for Data Mining
EC	Evolutionary Computing
FS	Fuzzy Systems
HC	Hard Computing
KDD	Knowledge Discovery From Data
ML	Machine Learning
NN	Neural Networks
PR	Probabilistic Reasoning
SEMMA	Sample, Explore, Modify, Model and Assess

Sumário

1 Introdução

2 Objetivo

2.1 Objetivo geral

2.2 Objetivos específicos

- Mostrar

3 Justificativa

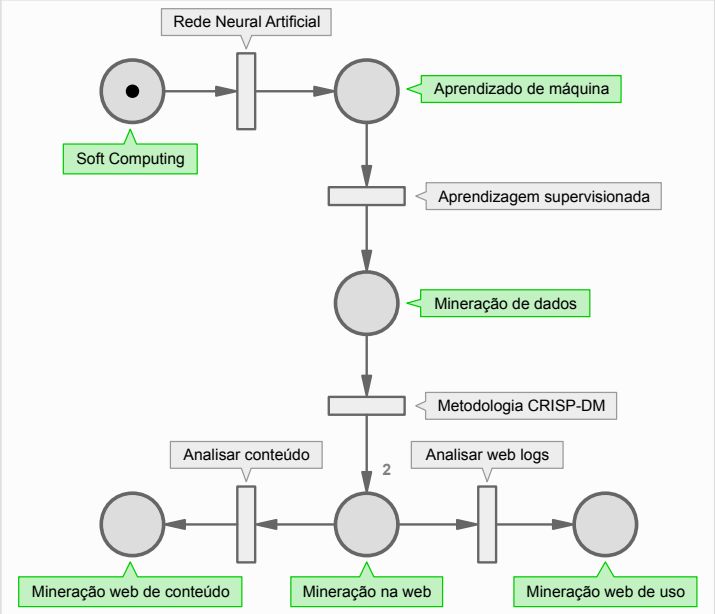
4 Trabalhos relacionados

5 Fundamentação teórica

A seguir, apresentam-se breves definições sobre os assuntos teóricos relacionados com a monografia, utilizando-se uma abordagem *top-down* como pode-se observar na (FIGURA 1), exibe-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 que são adotadas para que a ficha possa 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 1: *Estruturação do conhecimento para conseguir as ferramentas teóricas e técnicas relacionadas com o presente trabalho.*



Fonte: Autoria própria.

5.1 Soft computing

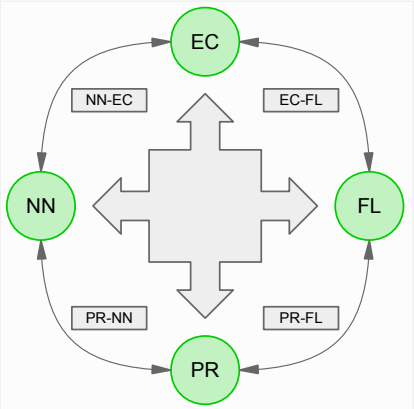
Existem dois tipos de paradigmas computacionais, a Soft Computing (SC) e a Hard Computing (HC), ambos os termos foram estabelecidos pela primeira vez pelo professor L. A. Zadeh no ano 1996. O HC trata sobre modelos precisos em que as soluções são atingidas imediatamente. Por outro lado, a SC lida com 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 é composta por técnicas como as Redes Neurais (Neural Networks), Computação Evolucionária (Evolutionary Computing), Sistemas Difusos (Fuzzy Systems) e Raciocínio Probabilístico (Probabilistic Reasoning), como mostrado na (FIGURA 2).

Figura 2: *Principais componentes da família da SC.*

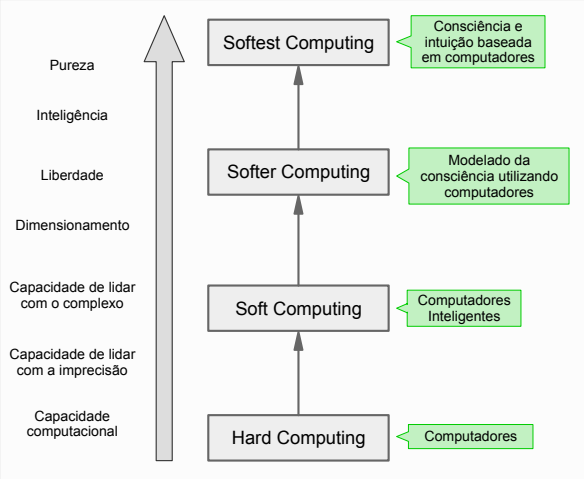


Fonte: Adaptado de (MANKAD, 2013).

As técnicas mostradas na (FIGURA 2), nem sempre estão isoladas. Elas podem ser combinadas 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 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 3: *Desenvolvimento da SC.*



Fonte: Adaptado de (CHATURVEDI, 2008).

Para conseguir ativar a transição na rede de Petri da (FIGURA 1) é necessário estabelecer a rede neural como técnica para poder passar ao estado seguinte que é o aprendizado de máquina.

5.1.1 Computação Evolucionária

5.1.2 Sistemas Difusos

5.1.3 Raciocínio Probabilístico

5.1.4 Rede Neural Artificial

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, ou ML).

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 esses 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.

Saber em que momento precisa-se do ML é fundamental porque justifica a utilização dos conceitos de ML no trabalho. 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 de 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.

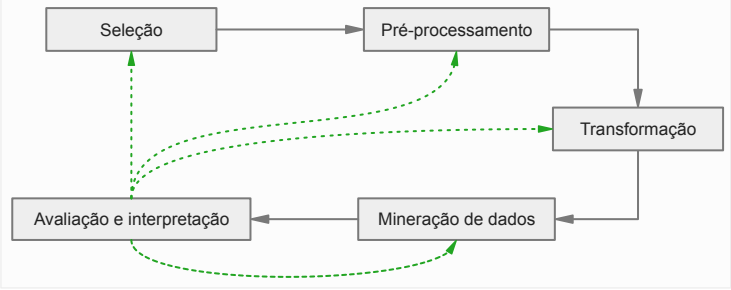
Em (RAJARAMAN; LESKOVEC; ULLMAN, 2012) descreve-se a mineração de dados como a descoberta de modelos a partir de dados. Existem duas abordagens: a estatística e a do aprendizado de máquina. Para o presente trabalho utiliza-se a abordagem de aprendizado de máquina, esta decisão justifica-se na (SUBSEÇÃO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

5.3.1 A metodologia 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, e possui cinco passos de desenvolvimento que se podem observar na (FIGURA 4).

Figura 4: *Uma visão geral dos passos que compõem o processo KDD.*

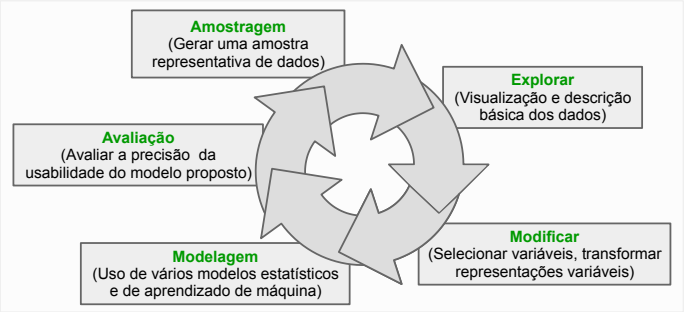


Fonte: Adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

5.3.2 A metodologia 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).

Figura 5: *Metodologia SEMMA proposto pela SAS.*



Fonte: Adaptado de (BINUS, 2014).

5.3.3 A metodologia 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 ferramenta que está constituída por seis processos cíclicos como se mostra na (FIGURA 6). Segundo a pesquisa de (PIATETSKY, 2014) no site KDnuggets, pode-se observar na (TABELA 1) que o CRISP-DM é o método mais utilizado nos anos 2007 e 2014.

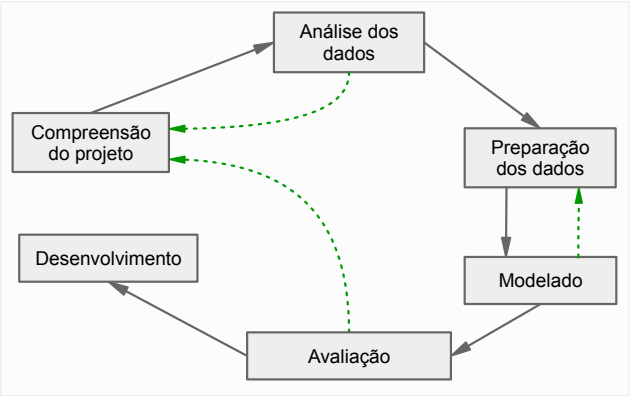
[1]Statistical Analysis System, é o nome de uma empresa pioneira em Business intelligence. Disponível em <http://www.sas.com>

[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>

Figura 6: *Metodologia CRISP-DM, utilizada no presente trabalho.*



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

Tabela 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
Criada por mim	19.0	27.5
SEMMA	13.0	8.5
Outra	4.0	8.0
Processo KDD	7.3	7.5
Criada pela organização onde trabalho	5.3	3.5
Específica de domínio	4.7	2.0
Nenhuma	4.7	0.0

Fonte: Adaptado de (PIATETSKY, 2014).

Para passar ao seguinte estado da rede de Petri estabelecida, de acordo com a (FIGURA 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 **??** vê-se a aplicação desta metodologia a partir do ponto de vista da proposta do trabalho.

5.4 O software Weka

6 Materiais e métodos

Utiliza-se o dataset de Gas disponivel em [].

Possui os atributos como mostra na figura

6.1 Compreensão do projeto

Descrição sobre o artigo relacionado.

6.2 Análise dos dados

6.3 Preparação dos dados

6.4 Modelagem

6.5 Avaliação

6.6 Desenvolvimento

7 Resultados

8 Conclusões e recomendações

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.

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

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

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.

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.

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.