Marcelo d'Almeida
md@id.uff.br

Algoritmos de Extração de Regras com Taxonomias

Survey - Mineração de Dados (2015.1)

Universidade Federal Fluminense

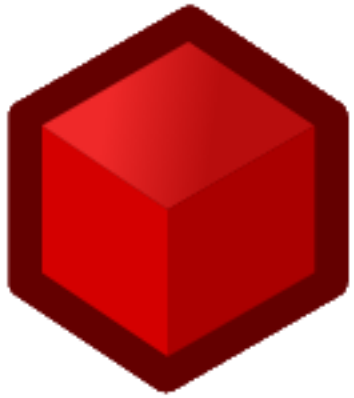
CCD-UFF
Clube de Ciência de Dados

Marcelo d'Almeida
md@id.uff.br

Knowledge-based Association Rules Mining

Incorporando conhecimento do domínio na extração de Regras de Associação

Survey - Mineração de Dados (2015.1)



CCD-UFF
Clube de Ciência de Dados

Conteúdo

- Conhecimento prévio
  - Regras de Associação
  - Regras de Associação com Taxonomia
  - Representação do Conhecimento
- Diversas nomenclaturas e variações
- Tipos de abordagens disponíveis
- Primeiras abordagens
- Abordagens adicionais
- Conclusão

Conhecimento prévio

Regras de Associação

[9]

- Representa um padrão de relacionamento entre itens de dados de um domínio de aplicação que ocorre com uma determinada frequência na base de dados
- Base de Dados Transacional
- Suporte(X) = T_X / T, Suporte(R) = T_{XUY} / T e Confiança(R) = T_{XUY} / T_X, onde

  X e Y são conjuntos de itens;

  T_X é o número de transações que incluem os elementos de X;

  T é o número de transações da base;

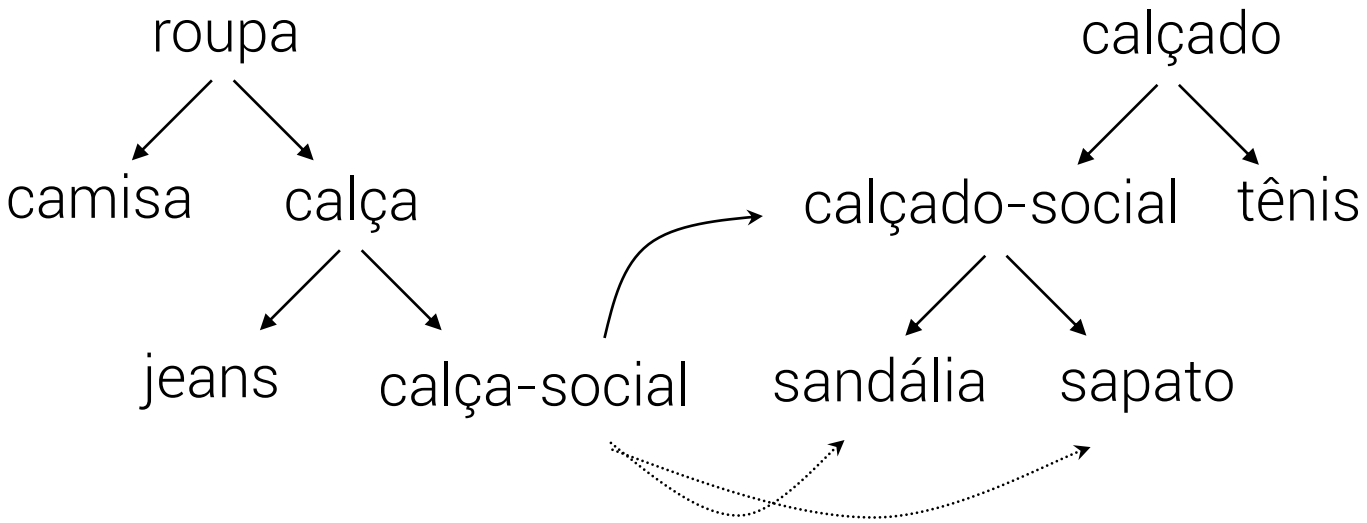
  R é a regra X ⇒ Y;

  T_{XUY} é o número de transações que incluem os elementos de X U Y.

Regras de Associação com Taxonomia

[9]

- Algumas regras podem não atender o suporte desejado. Porém, a generalização destas regras podem torna-las relevantes.
- Uma regra entre taxonomias pode relacionar itens de diferentes níveis.



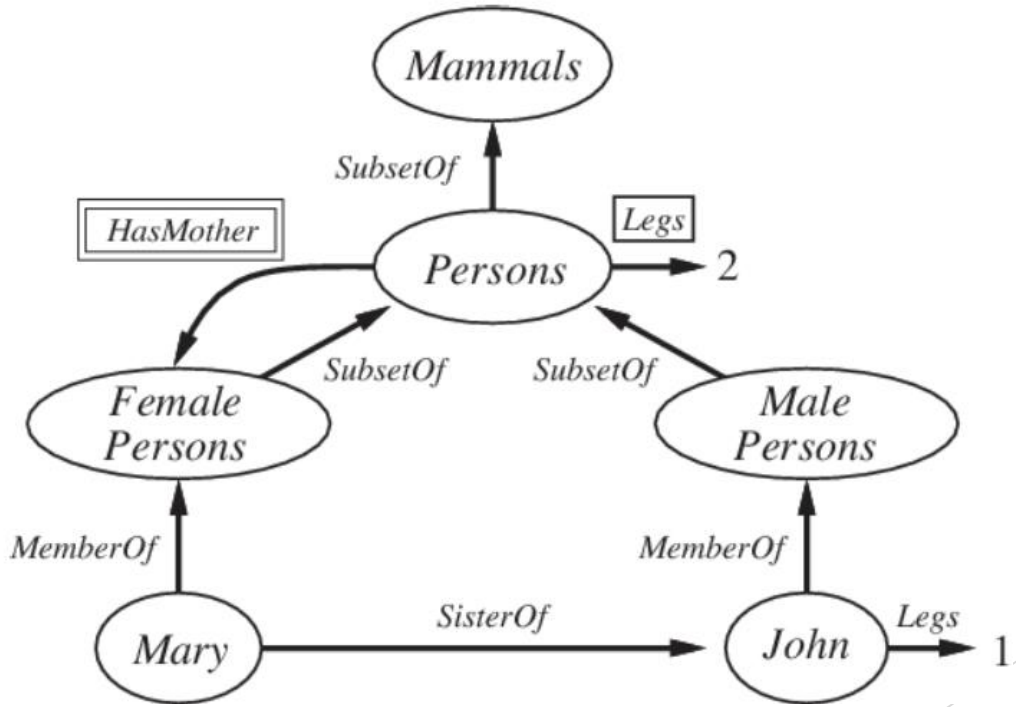
{calça-social} ⇒ {sandália}
{calça-social} ⇒ {sapato}

{calça-social} ⇒ {calçado-social}

Representação do Conhecimento

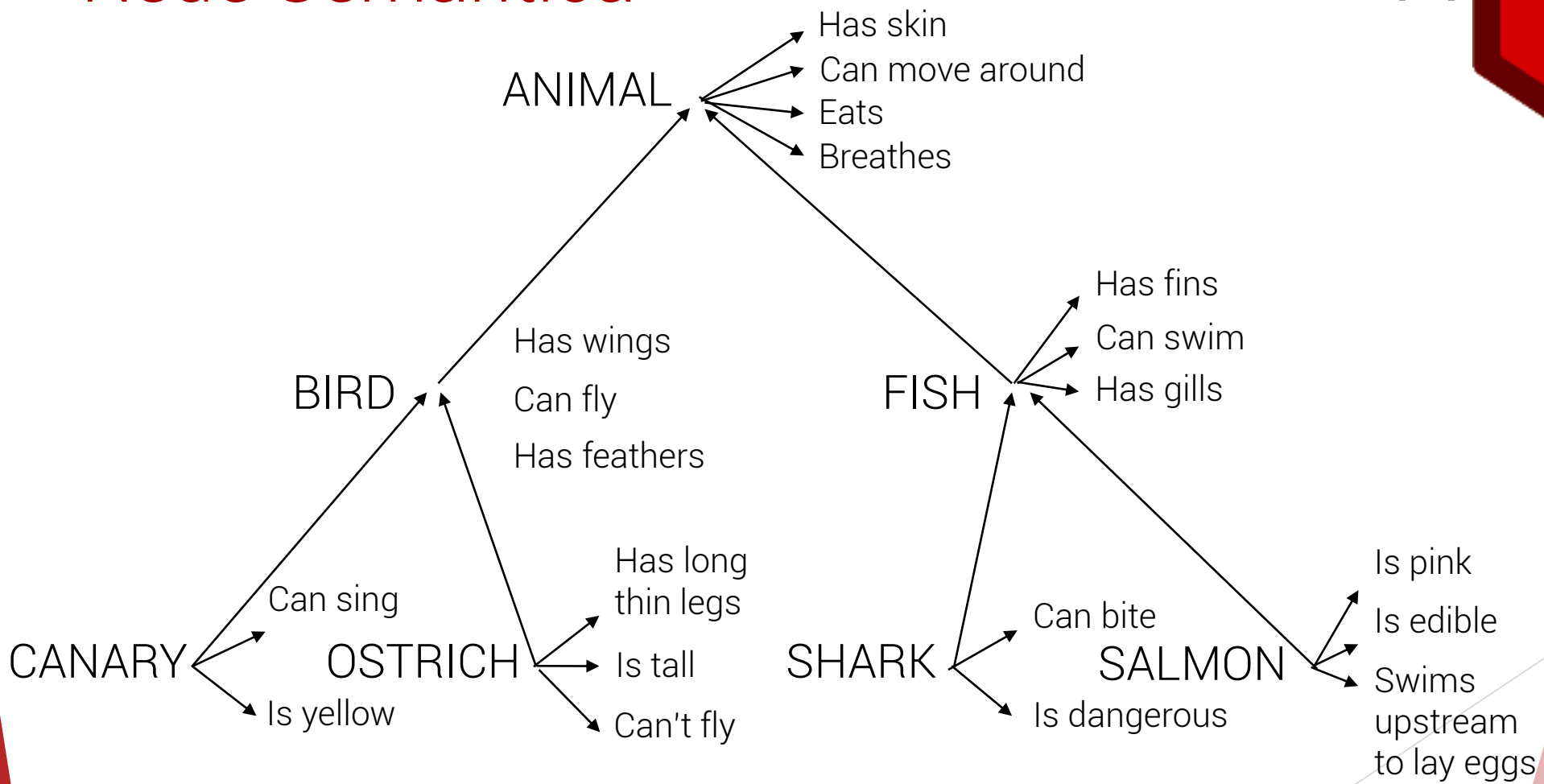
[10]

- Estabelecer um relacionamento entre o conhecimento humano e sua representação, por meio de formalismos
  - Redes Semânticas
  - Grafos Semânticos
  - Ontologias
  - ...



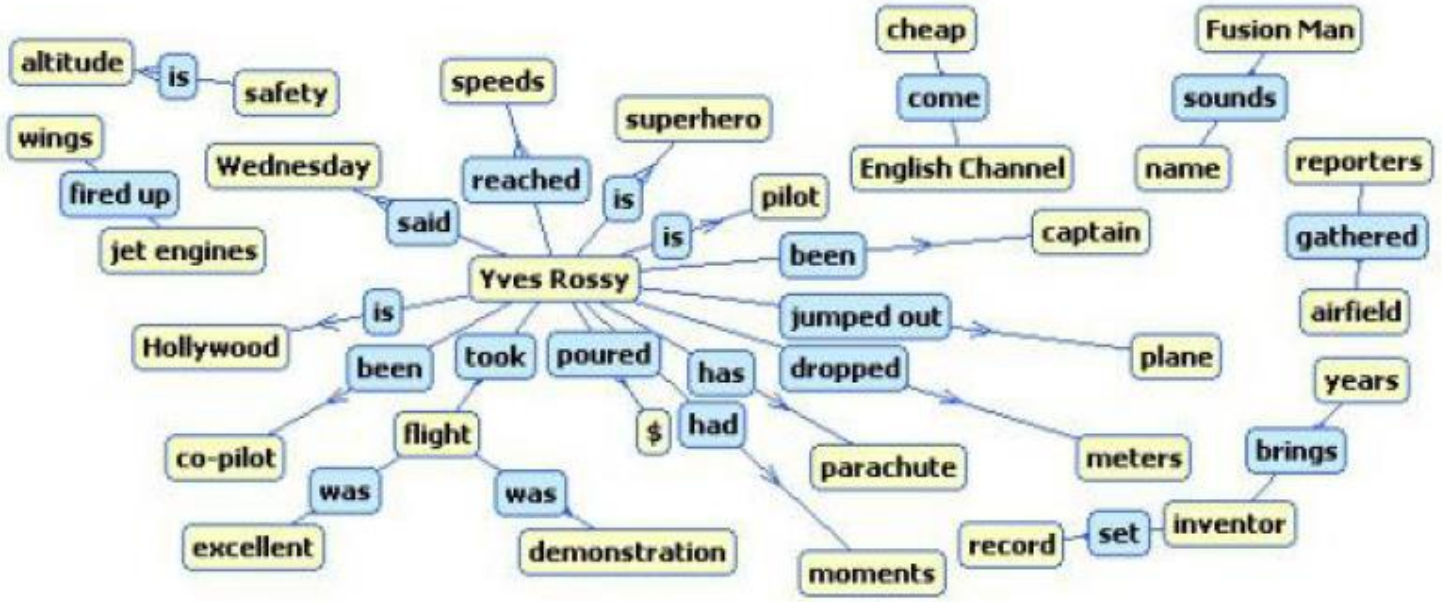
Representação do Conhecimento: Rede Semântica

[10]



Representação do Conhecimento: Grafo Semântico

[10]



Representação do Conhecimento: Ontologia

[10]

Vehicle world
- type
  - ground vehicle
  - ship
  - air craft
- function
  - to carry persons
  - to carry freights
- attribute
  - power
  - size
- component
  - engine
  - body
- traffic system
- ...

Diversas nomenclaturas e variações

Diversas nomenclaturas e variações

- Generalized Association Rules
  - Preknowledge-based Generalized Association Rules
  - …
- Multi-Level Association Rules
  - Cross-Level Association Rules
  - Constraint-based Multi-level Association Rules
  - …

Tipos de abordagens disponíveis

Tipos de abordagens Disponíveis

- Referente à etapa do processo de mineração no qual é aplicado
  - Pré-processamento
  - Processamento
  - Pós-processamento
  - Auxiliar
- Referente ao tipo de representação do conhecimento utilizado
  - Taxonomias
  - Taxonomias Fuzzy
  - Ontologias
  - Grafos de Conhecimento
  - DataCubes

Primeiras abordagens

‘Basic’ Algorithm

[1]

- “An obvious solution to the problem is to add all ancestors of each item in a transaction to the transaction, and then run any of the algorithms for mining association rules on these ‘extended transactions’. However, this ‘Basic’ algorithm is not very fast”

Cumulate, Stratify, Estimate and EstMerge Algorithms [1]

- Várias otimizações são realizadas para deixar a ideia intuitiva mais rápida
- “Run 2 to 5 times faster than Basic (and more than 100 times faster on one real-life dataset)”

Cumulate Algorithm

[1]

- Estratégias utilizadas:
  - “1. Filtering the ancestors added to transactions."
  - “2. Pre-computing ancestors."
  - “3. Pruning itemsets containing an item and its ancestor."

Stratify Algorithm

[1]

- Utiliza a noção de estratificação
  - Se um itemset mais generalizado não atender o suporte mínimo, algum itemset 'derivado' (algum elemento substituído por uma de suas especializações) também não atenderá e por isso não precisa ser considerado.

Estimate Algorithm

[1]

- Utiliza amostragem para estimar o suporte de cada elemento
  - Em cada passada, ele só calcula o suporte caso tenha o suporte esperado mínimo ou seja descendente de um que tenha suporte esperado mínimo (pois é a única forma de saber se ele atende ou não)
  - Realiza uma segunda passagem pra considerar os descendentes daqueles que ele tinha classificado de forma equivocada em relação ao mínimo
- As otimizações usadas no Cumulate também são válidas

EstMerge Algorithm

[1]

- Baseado no algoritmo Stratify e Similar ao algoritmo Estimate
  - Também usa amostragem para aumentar a eficácia (assim como o algoritmo Estimate).
  - Porém ele não faz uma segunda passagem. Ele, na primeira passagem, já considera como se o descendente tivesse o mínimo.

Abordagens adicionais

Generalization of Association Rules using Taxonomies – (GART Algorithm) [5]

- "A method that can help the analysis of the association rules is the use of taxonomies in the step of post-processing knowledge."
- "The user looks to a small set of rules without taxonomies, builds some taxonomies and then uses the algorithm to generalize the association rules, pruning the original rules that are generalized."

Generalized Association Rule Post-Processing Approach – (GARPA) [6]

- "It is important to note that our approach (GARPA) has basically five differences from the approach proposed by (Domingues and Rezende, 2005):"
  - "(a) generalization does not occur on only one side of the rule, but also on both sides;"
  - "(b) generalization does not only occur among the rules, but also among the items of the rules;"
  - "(c) it is not necessary that there is one specialized rule for each of the items contained in the taxonomy;"
  - "(d) generalization occurs even if one rule possesses more than one item with the same ancestor;"
  - "(e) a generalized rule will be valid only if its support/confidence is higher than t% of the highest value of the same measure in its specialized rules."

Ontology-Driven Association Rule Extraction

[7]

- “The focus of our approach is the introduction of the expressive power of ontologies for constraint-based multilevel association rule mining.”
- “Our approach follows the research line proposed by the cited works, nevertheless it introduces three main differences:”
  - “(i) we employ an ontology to represent an item taxonomy; “
  - “(ii) constraints can be defined on the basis of specific properties of the items; “
  - “(iii) by using an ontology instead of a taxonomy, a new item property or a concept can be added without re-engineering the (meta-data) representation model or the relational database.”

Improving Classification Models with Taxonomy Information (G−L³ Classifier)[3]

- "Note that, in general, more complex semantics-based models (e.g., ontologies) may also contain knowledge fruitful for classification models."
- "This work focuses on integrating taxonomy models in classification model learning"
- "We present an extended version of a state-of-the-art associative classifier [19] in which taxonomy information is pushed into the classification learning process to drive the extraction of generalized (high level) rules [9] instead of traditional ones [21]."
- "The proposed approach has the great advantage to be independent from the applied classifier type."

Updating Generalized Association Rules with Evolving Fuzzy Taxonomies (FDiff_ET and FDiff_ET* Algorithms)

[2]

- "Mining generalized association rules with fuzzy taxonomic structures has been recognized as an important extension of generalized associations mining problem. To date most work on this problem, however, required the taxonomies to be static, ignoring the fact that the taxonomies of items cannot necessarily be kept unchanged."

Generalized Association Rule Mining Using Genetic Algorithms

[8]

- "An interactive Association rule mining system is designed using a combination of genetic algorithms and a modified a priori based algorithm."
- "The association rule mining problem is modeled as a multi-objective combinatorial problem which is solved using genetic algorithms."
- "This is achieved as a combination of an a−priori based algorithm for finding frequent itemsets with multi-objective evolutionary algorithms (MOEA) with emphasis on genetic algorithms (GA)."
- "The main motivation for using GAs is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining tasks."

COmplexity Guided Association Rule Extraction (COGARE Algorithm) [4]

- “Our aim in this paper is to propose a new method to extract association rules from a fuzzy multidimensional model that can represent and manage imprecision in different aspects: COGARE.”
- “This method extracts inter-dimensional association rules and tries to reduce the complexity of the obtained rules using the fuzzy concepts defined in the dimensions and hierarchies.”

Conclusão

Conclusão

- A Representação do Conhecimento escolhida será a que for adequada às necessidades e restrições do domínio, bem como ao tipo, quantidade e qualidade da informação (conhecimento) disponível
- Se tem estudado como se melhorar o desempenho dos algoritmos e qualidade das regras geradas, mas também como atender outras características, como lidar com incertezas, por exemplo

Referências

- [1] Ramakrishnan Srikant, Rakesh Agrawal, "**Mining generalized association rules**", Future Generation Computer Systems 13 (1997) 161-180
- [2] Wen-Yang Lin • Ja-Hwung Su • Ming-Cheng Tseng, "**Updating generalized association rules with evolving fuzzy taxonomies**", Soft Comput (2012) 16:1109–1118
- [3] Luca Cagliero, Paolo Garza, "**Improving classification models with taxonomy information**", Data & Knowledge Engineering 86 (2013) 85–101
- [4] Nicolas Marin, Carlos Molina, José M. Serrano, and M. Amparo Vila, "**A Complexity Guided Algorithm for Association Rule Extraction on Fuzzy DataCubes**", IEEE Transactions on Fuzzy Systems, vol. 16, no. 3, june (2008)

Referências

- [5] Marcos Aurélio Domingues and Solange Oliveira Rezende, "Using Taxonomies to Facilitate the Analysis of the Association Rules", Proc. Second Int'l Workshop Knowledge Discovery and Ontologies, held with ECML/ PKDD, pp. 59-66, (2005).
- [6] Veronica Oliveira de Carvalho, Solange Oliveira Rezende, Mário de Castro, "Obtaining and Evaluating Generalized Association Rules", Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS), vol. 2 - Artificial Intelligence and Decision Support Systems, 310–315. (2007)
- [7] Andrea Bellandi, Barbara Furletti, Valerio Grossi, and Andrea Romei, "Ontology-Driven Association Rule Extraction: A Case Study", Proc. Workshop Context and Ontologies: Representation and Reasoning, pp. 1-10, (2007).
- [8] Peter P. Wakabi-Waiswa, Venansius Baryamureeba and K. Sarukesi, "Generalized Association Rule Mining Using Genetic Algorithms", International Journal of Computing and ICT Research, Vol. 2 No. 1, pp:59-69, (2008)

Referências

- [9] Prof. Alexandre Plastino, Aula 07 – Mineração de Dados, "**Regras de Associação**" (2015)
- [10] Prof. Aline Paes, Aula 13 – Inteligência Artificial, "**Representação de Conhecimento**" (2015)

Esta apresentação possui um documento associado

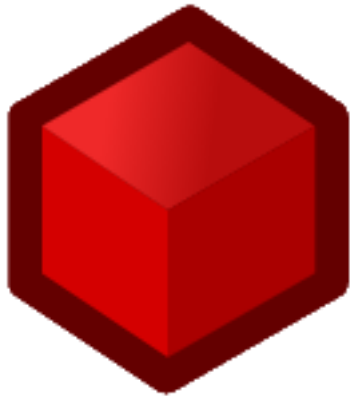
Veja mais em: github.com/ClubedeCienciadeDadosUFF/CCDrepository

“The difference between ordinary and extraordinary is that little extra.” (Jimmy Johnson)

Dúvidas?

Sugestões?

Marcelo d'Almeida
md@id.uff.br



CCD-UFF
Clube de Ciência de Dados