BIG DATA E DATA SCIENCE

Luís Cavique,
Universidade Aberta

1. Introdução

A informatização dos serviços, desde as sofisticadas transações em bolsa à simples compra de um café, associada às redes sociais e aos dispositivos móveis (*tablets*, *smart-phones*) produzem uma enorme quantidade de dados. Para além da quantidade de dados, a taxa de atualização desses mesmos dados é também enorme. Atualmente, em cada 10 minutos são gerados mais dados do que todos os dados gerados desde a pré-história até ao ano de 2003.

Por outro lado, a capacidade de processamento também tem tido aumentos significativos. Nos últimos 40 anos a capacidade de integração dos circuitos integrados permitiu duplicar em cada dois anos a capacidade de processamento [14], aumentar a capacidade de armazenamento e reduzir o respetivo preço. A lei de Moore tem-se verificado nos últimos 40 anos, permitindo um aumento no processamento na ordem de 2^{20}.

Para ter uma noção do aumento da capacidade, dada a dificuldade dos humanos para compreender o significado de um crescimento exponencial, vamos usar o exemplo do tempo de viagem de Lisboa ao Porto. Suponhamos que uma viagem de Lisboa ao Porto, há 40 anos atrás, decorria em média em 6 horas. Se a evolução dos transportes fosse tão grande como nos computadores, o tempo da viagem de Lisboa ao Porto, hoje em dia, teria a duração de 2 centésimos de segundo ($6x60x60/2^{20}$).

O grande volume de dados compensado pelo aumento da capacidade de processamento tem originado novos conceitos, como o *Big Data* e a criação de novas profissões como os *data scientists*, apelidada pela Harvard Business Review como a profissão mais *sexy* do século XXI.

2. *Big Data*

Com o advento da web 2.0 (a web das pessoas) associada aos dispositivos móveis e à *internet of things*, as clássicas aplicações empresariais foram largamente ultrapassadas em volume de dados.

Num estudo realizado em 2012, o valor estimado de informação no planeta foi de 2,8 ZB (zetabytes, 10^{21} bytes). A mudança de escala no volume de dados e na sua taxa de atualização deu origem ao que genericamente se chama de *Big Data*.

Ao nome *Big Data* está associada a sigla 3V: volume, velocidade de atualização e variedade dos formatos. Alguns autores incluem um quarto V de valor ou veracidade dos dados.

Dos 2,8 ZB existentes, 85% são dados não estruturados, ou seja, *medias* como o vídeo, fotografia e som. Dos restantes 15%, de dados formatados e de texto, só 3% são analisados. Concluímos assim que só a pequena percentagem de 0,45% dos dados do planeta são objeto de análise. Tal como na astrofísica, onde a matéria negra contém eventualmente várias explicações para a origem do universo, os 99,55% dos dados não analisados são apelidados de *dark data*.

Com o surgimento de novos formatos de dados estruturados surgiu dentro do *Big Data* o conceito do NoSQL [3]. O NoSQL, ou Notonly SQL, permite o armazenamento, tratamento e consulta de dados de forma muito eficiente. As soluções NoSQL estão divididas em alguns grupos:

- Armazenamento de chave/valor, como Voldemort da LinkedIn.
- Armazenamento de super-colunas, como HBase ou Cassandra do Facebook.
- Armazenamento de documentos, como XMLdatabase ou MongoDB.
- Armazenamento de grafos, como HyperGraphDB ou ArangoDB.
- Armazenamento orientado a objetos, como Db4object.

Tendo como ponto de partida o modelo relacional e a linguagem declarativa SQL (*Structured Query Language*), utilizada na maior parte das bases de dados das empresas, o NoSQL apresenta-se como a alternativa para lidar com grandes volumes de dados.

As estruturas das soluções NoSQL foram simplificadas relativamente ao modelo relacional e garantem a consulta da informação de forma muito eficiente, com complexidades algorítmicas de ordem O(1).

Em NoSQL a complexidade máxima deve ser da ordem O(N), sendo N a dimensão do ficheiro. Ao contrário do SQL a operação de junção de tabelas não existe, dada a sua elevada complexidade de ordem $O(N^2)$ para o pior caso.

Para a agregação de dados é utilizado o conceito de *MapReduce*, implementado em duas fases. A função do operador *Map* seleciona os dados em subgrupos. A operação *Reduce* agrega a informação de cada subgrupo. A complexidade algorítmica no pior caso será de duas vezes O(N).

O conceito de *Big Data* traz um conjunto de novos desafios para lidar com grandes volumes de dados, tanto para as empresas como para a comunidade científica. O desenvolvimento de novos algoritmos é crítico já que as complexidades algorítmicas são de preferência de ordem O(1) e nunca devem exceder a ordem O(N).

Por consequência, o *Big Data* cria novas oportunidades na tomada de decisão baseada em dados, *data driven decisions*. Tal como refere Peter Norvig, diretor da Google Research, "nós não temos melhores algoritmos; nós temos mais dados" [10].

3. *Data Science*

Data Science, é o atual termo para a ciência que analisa dados, combinando a estatística com *machine learning/data mining* e tecnologias de base de dados, para responder ao desafio que o *Big Data* apresenta.

O termo criado na década de 2010, *Data Science*, corresponde aquilo que nos anos de 1970 se apelidava de *Decision Support Systems*, DSS, nos anos 80 aos *Executive Information Systems*, EIS, nos anos 90 aos *Online Analytical Processing*, OLAP, e nos anos de 2000 ao *Business Intelligence*, BI [10].

3.1 Base de Dados *versus Data Mining*

As questões colocadas que têm resposta numa Base de Dados são semelhantes às questões colocadas ao analista de *Data Mining*.

Em Base de Dados pretende-se por exemplo:

- Identificar os clientes que compraram mais de 1000 euros.
- Identificar os dois produtos mais vendidos.
- Identificar os 10 clientes com mais reclamações.

Enquanto que em *Data Mining* procura-se:

- Identificar os grupos de clientes com hábitos de compra idênticos (*clustering*).
- Encontrar o produto X que é adquirido com o produto Y (regras associativas).
- Encontrar os atributos que levam os clientes a reclamar (classificação).

Embora as questões sejam semelhantes, nas Bases de Dados é apresentado um padrão (e.g. consulta SQL) e são devolvidos dados, por outro lado, em *Data Mining* são fornecidos os dados e pretende-se extrair padrões.

3.2 Macro e Micro padrões

Data Science é um processo computacional para descobrir "padrões". Padrão é uma forma com uma configuração específica e facilmente reconhecível, que se caracteriza por uma regularidade, repetição de partes e acumulação de elementos. Por exemplo, uma duna criada pela ação do vento é composta por várias camadas de areia e tem uma configuração reconhecível.

Os micro-padrões correspondem a pequenas percentagens de dados; por exemplo nas regras associativas, uma medida de suporte que apresenta valores suporte ≥ 5%, sendo escolhidas as regras com maior *confidence* (ou probabilidade condicionada). Por outro lado, os macro-padrões envolvem uma grande percentagem, ou a totalidade, dos dados; por exemplo na modelação com regressão são utilizados todos os dados disponíveis. Os micro-padrões caracterizam-se por

elevada confiança (*confidence*) e os macro-padrões por elevado suporte.

Existem outros exemplos de micro-padrões: nos problemas de *sequence/episode mining* com suporte maior ou igual a 1%; no problema de classificação, ao utilizar *decision trees*, cada ramo da árvore corresponde a uma pequena percentagem dos dados; ainda no problema de classificação ao utilizar o *k-nearest neighbor* a comparação que é efetuada é com o reduzido número de *k* elementos. Quanto aos macro-padrões, em problemas como regressão, teste de hipóteses, *clustering* ou redução de atributos, todos os dados são tidos em consideração.

A origem desta dicotomia na análise de dados remonta aquando do aparecimento do *Data Mining*, hoje uma área madura, mas que tinha inicialmente uma conotação negativa com os nomes de *data snooping* (bisbilhotando) e *data fishing*, onde o objetivo era explorar e/ou espiar subconjuntos de dados.

Leo Breiman em 2001 [5] já tinha referido as duas culturas na modelação de dados. A cultura dos micro-padrões corresponde à procura de pequenas percentagens de dados com eventual utilidade ou interesse. Esta abordagem tem tido, até à data, um grande apoio dos grandes decisores dos EUA em projetos de mais de 1.000.000 dólares. A cultura dos macro-padrões utiliza a totalidade dos dados, tem origem na matemática e na estatística e conta com projetos vinte vezes menores que os anteriores.

No atual paradigma de *Big Data*, em que as complexidades algorítmicas não devem exceder O(N), grande parte dos algoritmos de *machine learning/data mining* são desadequados. A reutilização das métricas da estatística combinada com a tecnologia de base de dados faz anunciar a reconciliação das duas culturas na modelação de dados na recente *Data Science*.

3.3 *Data Scientist*

Data Scientist é apelidada como a profissão mais sexy do século XXI [11]. Por *data scientist* entende-se "alguém melhor em estatística que um engenheiro informático e alguém melhor em programação do que um matemático". A maior parte dos programadores não se querem envolver em conceitos matemáticos e da mesma forma os estatísticos não aceitam programar em SQL, R ou Python.

O *data scientist* será assim alguém que saiba diferenciar um teste de hipóteses *t-student* de um qui-quadrado, ao mesmo tempo que sabe ver a diferença entre um algoritmo polinomial de ordem O(N) e de $O(N^2)$.

As duas culturas na modelação de dados criaram especialistas que atualmente são obrigados a compatibilizar esforços. A Investigação Operacional encontra-se numa posição privilegiada já que sempre combinou os conceitos da matemática com a sua aplicação nas ciências informáticas.

4. Redução da Dimensionalidade

Em ambientes *Big Data* o volume é grande, dinâmico e não estruturado. Por outro lado, não existem algoritmos disponíveis para responder a este desafio. O volume de dados não analisados (*dark data*) é ao mesmo tempo uma oportunidade e uma inquietação, visto que os dados gerados excedem largamente a capacidade de armazenamento instalada.

Se não se pode alterar de imediato a complexidade dos algoritmos, a resposta pode estar na redução da dimensionalidade dos dados. Em Investigação Operacional existe uma larga experiência neste campo. A análise das componentes principais e a análise fatorial são técnicas conhecidas na estatística e em *machine learning* para redução do número de variáveis.

A redução da dimensão pode ainda ser realizada pela transformação do problema e pela sumarização dos casos (ou linhas). Os dados são condensados com vista a encontrar padrões de grandes subconjuntos de dados, utilizando portanto a abordagem dos macro-padrões. De seguida apresentamos vários exemplos de transformações em redes e grafos.

4.1 Análise Topológica de Dados

A Análise Topológica de Dados [6] representa os dados utilizando redes. A rede agrupa dados semelhantes em nós e cria arcos se existe partilha de dados entre dois nós diferentes. Visto que cada nó representa vários pontos, a rede permite comprimir os dados com uma alta dimensionalidade para uma representação de mais baixa dimensionalidade.

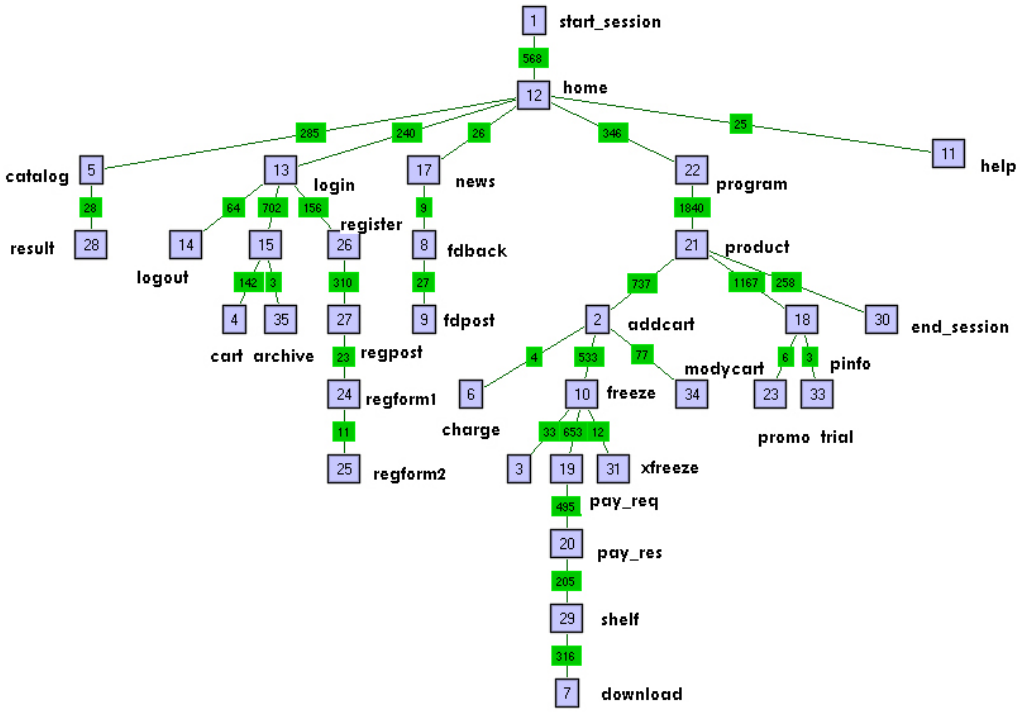

Figura 1: Resultado do algoritmo Ramex num problema de *web mining*.

A topologia é o estudo da forma, em que se distinguem três propriedades que relaxam o conceito de forma: a invariância das coordenadas, a invariância da deformação e a compressão das representações. A invariância da deformação é particularmente interessante, já que se adapta perfeitamente à capacidade dos humanos de compreenderem as formas. Por exemplo, a letra "A" é compreendida pelos humanos qualquer que seja a fonte tipográfica utilizada. Estas características fazem com que a Análise Topológica de Dados se esteja a tornar numa nova área de estudo em *data mining* e na visualização de dados.

4.2 *Process Mining*

Uma percentagem apreciável do *Big Data* corresponde aos históricos de eventos (*eventlogs*) que são gerados em cada instante, durante 7 x 24 horas, por milhares de milhões de dispositivos fixos e móveis.

Process Mining [1] é uma técnica que permite extrair informação de históricos de eventos. Segundo o autor, o *Process Mining* cria pontes entre o *Data Mining* e o *Business Process Modeling*. A técnica considera a acumulação de eventos tendo como objetivo melhorar a representação dos dados. A abordagem utiliza redes de Petri, onde cada processo é representado por um nó, e as sequências de eventos podem ser condensadas na rede. Esta abordagem cria novos padrões para os problemas antigos de *sequence mining*, tendo em vista criar equilíbrios entre a simplicidade e a exatidão dos resultados.

4.3 Similis

As regras associativas ficaram célebres, ao encontrar um padrão nos supermercados onde jovens casais com filhos às sextas-feiras e sábados, ao comprar fraldas também compravam cerveja. A regra fraldas => cervejas, tendo uma métrica de suporte de algumas centésimas, estava associada a uma confiança (probabilidade condicionada) relevante.

O algoritmo Apriori [2] foi o primeiro algoritmo para o *Market Basket Analysis*. O Apriori gera para um pequeno número de produtos um enorme conjunto de regras associativas, i.e. micro-padrões, que devem ser criteriosamente escolhidas pelo utilizador final. O trabalho de Cavique [7] com o algoritmo Similis, resolve o *Market Basket Analysis* para um elevado número de produtos e evita a escolha entre milhares de micro-padrões, devolvendo padrões baseados na acumulação, i.e. macro-padrões que representam a totalidade dos dados. O algoritmo Similis está dividido em duas partes. Na primeira transforma o problema num grafo ponderado e na segunda encontra subgrafos completos que correspondem aos cabazes de compras mais comprados.

4.4 Ramex

A descoberta de padrões sequenciais é um tema muito importante em *data mining*, dado o grande número de aplicações que incluem a análise de compras, *web mining*, sequência de ADN, entre outros.

O algoritmo inicial, o AprioriAll [15], para além de ter uma elevada complexidade temporal, encontra milhares de micro-padrões de difícil seleção e que requerem um trabalho exaustivo na atribuição de utilidade ou interesse.

As cadeias de Markov representam um conjunto de estados associados com um conjunto de transições entre estados. No caso da análise do cabaz de compras, cada estado corresponde a um item e no caso da navegação da *web*, cada estado é uma página. Os modelos de Markov foram usados para representar e analisar os utilizadores que navegam na *web* em [4].

No problema de *sequence mining* tratado por Cavique [8], é apresentado o algoritmo Ramex que gera árvores que envolvem todos os elementos numa perspetiva de macro-padrões. Ramex tem origem no latim e significa ramos de uma árvore. Na abordagem pretendemos manter a visão global dos itens e evitar tempos computacionais não-polinomiais. Usando heurísticas baseadas no algoritmo da árvore geradora, podem ser encontradas em redes cíclicas as estruturas de árvores com maior peso, que correspondem aos padrões sequenciais mais frequentes.

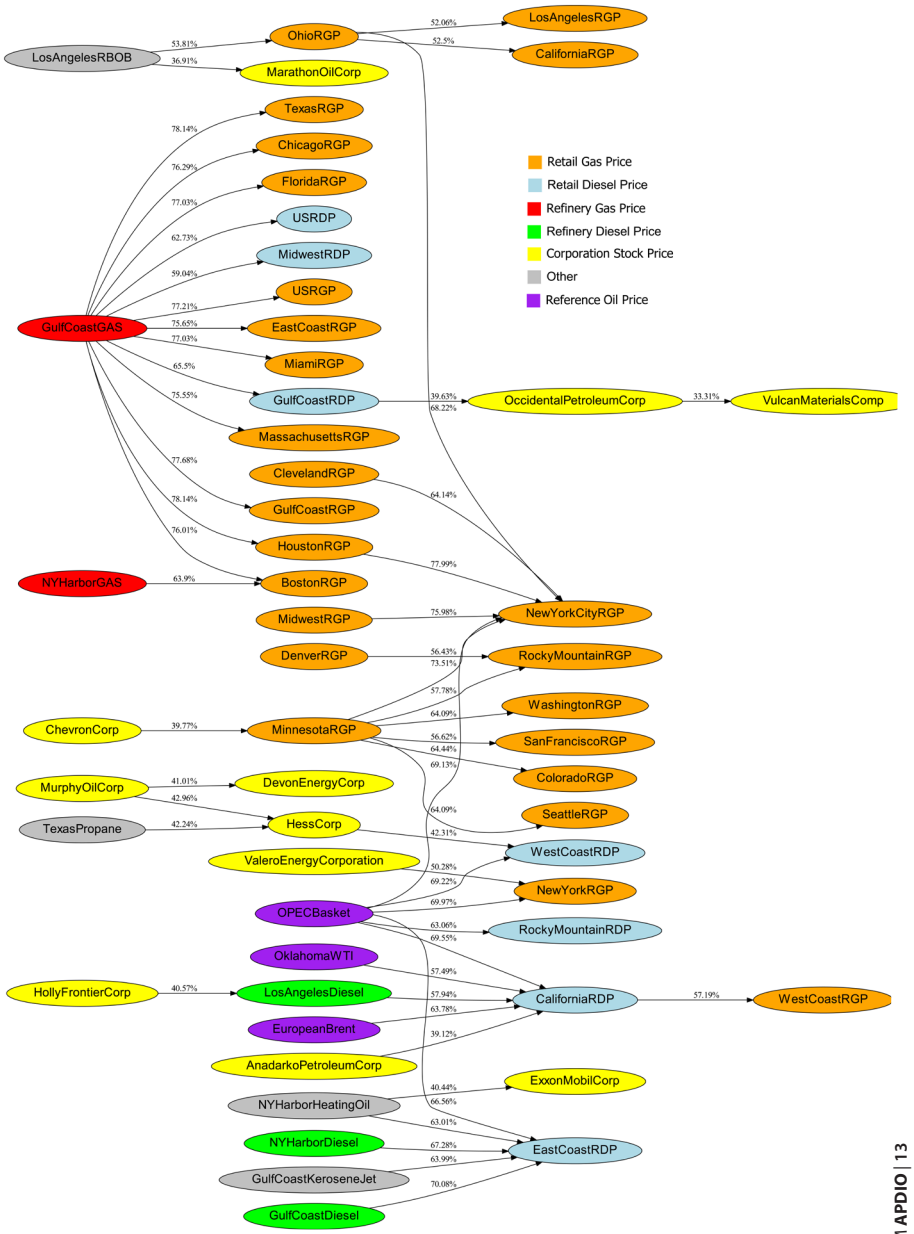

Figura 2: Preços dos petróleos e derivados.

O algoritmo Ramex divide-se em duas fases. Numa primeira fase o ficheiro de entrada é transformado num grafo orientado. Na segunda fase é aplicado o algoritmo *Maximum Weight Rooted Branching* definido por Fulkerson [12]. Neste algoritmo é dado um nó inicial a partir do qual se desenvolve uma árvore.

O algoritmo foi testado num problema de *web mining* tendo sido encontradas as sequências da Figura 1. Cada ramo da árvore corresponde a uma sequência de eventos com ramos idênticos à estrutura do *web site*.

4.5 Ramex com poli-árvores

Uma poli-árvore é um grafo orientado acíclico com um arco entre cada par de nós no máximo. O grau interno dos vértices de uma árvore é zero (a raiz) ou um. Por sua vez, o grau interno dos vértices de uma poli-árvore pode ser maior que um. Podemos ainda acrescentar que numa poli-árvore para cada par de nós só pode existir uma única sequência de nós.

No trabalho [9] é apresentada a versão do Ramex utilizando poli-árvores para a deteção de padrões sequenciais. Para testar a nova abordagem foram usados ficheiros de grandes dimensões. As experiências foram implementadas utilizando os ficheiros gerados pelo IBM Quest Synthetic. O algoritmo utiliza uma matriz inicial semelhante às das Cadeias de Markov, mas usa uma heurística polinomial baseada no algoritmo de Prim para determinar os padrões. Nesta abordagem encontramos as seguintes vantagens:

i) Incremental: Visto que os dados dos eventos são transformados em pesos no grafo, a atualização de novos eventos pode ser realizada de forma incremental.

ii) Inexistência de parâmetros: A maior parte dos algoritmos para deteção de sequências utilizam o suporte mínimo com o parâmetro para controlar a explosão combinatória. Para o algoritmo proposto não há necessidade de qualquer parâmetro.

iii) Escalabilidade: Em comparação com os demais algoritmos, a nossa abordagem não faz uma procura exaustiva. Contudo, utiliza os dados condensados numa rede. O procedimento que devolve o resultado da árvore tem uma complexidade polinomial e apresenta uma ótima escalabilidade.

iv) Visualização: Usualmente os pacotes de *software* mais conhecidos geram um grande número de regras, perdendo-se portanto a visão global. Na nossa abordagem todos os itens são tomados em consideração e a visualização das poli-árvores mais pesadas corresponde ao raio-X das sequências de eventos.

A utilização do algoritmo Ramex aplicado aos mercados financeiros deu origem ao Ramex-Forum [13]. A Figura 2 apresenta os resultados do algoritmo para as influências dos preços dos petróleos e derivados, extraída do trabalho de Tiple [16].

5. Conclusões

Neste artigo foram apresentados os conceitos básicos de *Big Data* e a nova área a que deu origem, a *Data Science*. Em *Data Science* foi discutida e exemplificada a noção de redução da dimensionalidade dos dados.

Como conclusões para a IO em ação, podemos referir duas grandes oportunidades que o *Big Data* oferece:

i) A Investigação Operacional encontra-se numa situação privilegiada, ao combinar, desde sempre, a matemática e a informática, para lidar com o *Data Science* e para liderar a formação numa das profissões mais atraentes do século XXI. A necessidade de voltar a recorrer aos algoritmos de baixa complexidade da estatística, coloca as técnicas de IO na vanguarda.

ii) O ambiente *Big Data* exige aos programadores e investigadores um conjunto de novos algoritmos, tornando-se urgente a redução da complexidade temporal de quase todos os algoritmos, desde o simples cálculo da variância, em estatística, até ao mais complexo problema de *sequence mining*. Para responder a este desafio a redução da dimensionalidade é uma abordagem já demonstrada. Os exemplos apresentados utilizam duas fases distintas. A primeira fase acumula os dados em bruto numa estrutura de dados condensados: rede na Análise Topológica de Dados [6], rede de Petri [1], cadeia Markov [4] ou grafo [8]. Na segunda fase é possível procurar os macro-padrões na estrutura de dados condensados. Os algoritmos para as referidas estruturas de dados são igualmente conhecidos na Investigação Operacional, tornando este tipo de redução da dimensão dos problemas muito aliciante.

Referências

[1] Aalst, W. van der, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, Springer-Verlag Berlin Heidelberg, 2011.

[2] Agrawal, R., Srikant, R., Fast algorithms for mining association rules, *Proceedings of the 20th International Conference on Very Large Data Bases*, 487-499, 1994.

[3] Alexandre, J., Cavique, L., NoSQL no suporte à análise de grande volume de dados, *Revista de Ciências da Computação*, 8, 37-48, 2013.

[4] Borges, J., Levene, M., Evaluating variable-length Markov chain models for analysis of user web navigation sessions, *IEEE Transactions on Knowledge and Data Engineering*, 19, 441-452, 2007.

[5] Breiman, L., Statistical modeling: the two cultures, *Statistical Science*, 16, 199-231, 2001.

[6] Carlsson, G., Topology and data, *Bulletin of the American Mathematical Society*, 46, 255–308, 2009.

[7] Cavique, L., A scalable algorithm for the market basket analysis, *Journal of Retailing and Consumer Services*, Special Issue on Data Mining Applications in Retailing and Consumer Services, 14, 400-407, 2007.

[8] Cavique, L., A network algorithm to discover sequential patterns, *Progress in Artificial Intelligence, EPIA 2007, Lecture Notes in Computer Science*, 4874, J. Neves, M. Santos e J. Machado (eds.), Springer-Verlag Berlin Heidelberg, 406-414, 2007.

[9] Cavique, L., Coelho, J. S., Descoberta de padrões sequenciais utilizando árvores orientadas, *Revista de Ciências da Computação*, 3, 12-22, 2008.

[10] Davenport, T. H., *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*, Harvard Business School Publishing Corporation, 2014.

[11] Davenport, T. H., Patil, D. J., Data scientist: the sexiest job of the 21st century, *Harvard Business Review*, 90, 70–76, 2012.

[12] Fulkerson, D. R., Packing rooted directed cuts in a weighted directed graph, *Mathematical Programming*, 6, 1-13, 1974.

[13] Marques, N. C., Cavique, L., Sequential pattern mining of price interactions, *EPIA 2013, 16th Portuguese Conference, Advances in Artificial Intelligence, Local Proceedings*, Angra do Heroísmo, Açores, Portugal, 314-325, 2013.

[14] Moore, G. E., Cramming more components onto integrated circuits, *Electronics*, 114–117, 1965.

[15] Srikant, R., Agrawal, R., Mining sequential patterns: generalizations and performance improvements, *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology, EDBT, Lecture Notes in Computer Science*, 1057, 3–17, 1996.

[16] Tiple, P. S., *Tool for Discovering Sequential Patterns in Financial Markets*, Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Faculdade de Ciências e Tecnologia da Universidade Nova Lisboa, 2014.