Mineracão de padrões entre doenças relacionadas ao saneamento ambiental inadequado

Rafael P. Scholant[1], Sandro da Silva Camargo[2]

[1]Especialização em gestão estratégica de tecnologia da informação – Instituto de Desenvolvimento do Alto Uruguai (IDEAU) Av. Santa Tecla, 4200 – Bagé – RS – Brasil

[2]Orientador – Universidade Federal do Pampa (UNIPAMPA), Campus Bagé.
Rua Travessa 45, 1650, gabinete 3139 – Bagé – RS – Brasil

rafael.scholant@gmail.com, sandro.camargo@unipampa.edu.br

Abstract. *This study aims to gain insight into the technical data mining and use practices on cases of diseases related to inadequate sanitation. In the first stage, and will address briefly some data mining techniques and subsequently the technique that should be used to obtain the expected results will be set. The time horizon selected for this work will be a period of seventeen based on data obtained from the Brazilian portal of open data, which are provided by the fe- deral government. Finally, the knowledge gained throughout this work, which involves the interaction between two areas of expertise - health and data mining- will be summarized at the end of the work the results of data mining exercise.*

Resumo. *Este trabalho tem por objetivo obter conhecimento sobre as técnicas de mineração de dados e a sua utilização pratica sobre os casos de doenças relacionadas ao saneamento ambiental inadequado. Em uma primeira etapa, e de forma breve serão abordadas algumas técnicas de mineração de dados e, posteriormente, será definida a técnica que deverá ser utilizada para a obtenção dos resultados esperados. O horizonte temporal selecionado para este trabalhoseráde um período de dezessete anos tendo por base dados obtidos junto aoportal brasileiro de dados abertos, que são fornecidos pelo governo federal. Por fim, os conhecimentos obtidos ao longo deste trabalho, que pressupõe a interação entre duas áreas de conhecimento - saúde e mineração de dados -, serão sintetizados ao final do trabalho os resultados do exercício de mineração de dados.*

1. Introdução

Não restam dúvidas que o Brasil tem avançado muito ao longo dos últimos anos, passando por transições governamentais de uma forma madura, contudo, para um país que realmente quer ser protagonista, o Brasil ainda precisa avançar muito em certos aspectos básicos, e a questão do saneamento ambiental talvez seja uma das principais barreiras a serem superadas.

Em 2009, a organização mundial de saúde (OMS) apontou a falta de saneamento ambiental como o decimo primeiro fator de risco para mortes no mundo [OMS 2009].

Neste contexto, em 28 de julho de 2010, a organização das nações unidas

(ONU) reconheceu o acesso ao saneamento básico como um direito de todo ser humano, sendo um fator primordial para a prevenção de problemas de saúde.

O saneamento ambiental no Brasil encontra-se muito abaixo do esperado, principalmente no que se refere aos serviços relacionados à coleta e tratamento de esgotos [Neri 2011].

O presente artigo tem por objetivo analisar os dados sobre doenças relacionadas ao saneamento ambiental inadequado, obtidos junto ao portal de dados disponibilizado pelo governo federal, para que se possa criar situações e estatísticas sobre os mesmos, desvendando assim como o saneamento ambiental vem evoluindo durante os últimos anos.

2. Mineração de dados

Diariamente as organizações acumulam um grande volume de dados em seus *softwares* operacionais. Os dados que são obtidos através destes *softwares* são dados brutos, que não expressam nada da forma na qual são salvos, demonstram somente o que aconteceu naquele exato momento. Para que estes dados possam ser realmente uteis para os gestores, estes devem ser tratados e analisados, neste ponto que surge a necessidade da mineração de dados [O'brien 2011].

A mineração de dados é uma técnica que tem por objetivo explorar grandes con- juntos de dados, para que possa se estabelecer relações, associações e padrões que seriam de difícil visualização [Laudon and Laudon 2011]. Para que possa ser feita esta exploração, são utilizados algoritmos de aprendizagem ou de classificação baseados em redes neurais e estatísticas. Os resultados gerados geralmente são expressos em forma de regras, hipóteses, arvores de decisão e grafos.

No entanto, a união de três principais recursos é o que torna a mineração de dados possível, o banco de dados no qual serão obtidos os dados para serem analisados,a estatística que seráutilizada para descobrir o quanto cada dado éimportante para ainformação final e por ultimo mas não menos importante a inteligencia artificial que fará combinações entre os dados e as estatísticas para a descoberta de padrões, conforme esquematizado na Figura 1.

2.1. Tarefas e técnicas de mineração de dados

É importante saber diferenciar o que é uma tarefa e o que é uma técnica de mineração
de dados. A tarefa consiste na especificação das informações que deverão ser obtidos dos dados, que tipo de regularidades ou categorias de padrões terão relevância para a pesquisa. Já a técnica de mineração consiste na especificação de métodos que possam garantir que os padrões poderão gerar alguma informação com relevância. A integração entre estes elementos é esquematizada na Figura 2.

Dentre as principais técnicas utilizadas em mineração de dados, exitem técnicas estatísticas e de aprendizado de máquina. A seguir, será feita uma breve descrição das principais técnicas de mineração.

- Associação: São ocorrências ligadas a um único evento, por exemplo: um estudo sobre modelos de compras em supermercados pode se descobrir que, quando houver uma compra de pão, o mesmo comprador em 70% das vezes também compra

Figura 1. Principais recursos que consistem a minerao de dados

manteiga, porem quando há uma promoção a manteiga é comprada em 90% das n vezes. Com estas informações, os gestores da organização têm decisões mais fáceis de se tomar, pois os mesmos podem ver mais sobre o assunto.

- Classificação: Reconhece modelos que podem descrever o grupo ao qual um item em especifico pertence por meio de um exame dos itens que já foram anteriormente classificados e também pela interferência de um conjunto especifico de regras. Como por exemplo empresas operadoras de cartões de credito, que podem descobrir regularidades entre clientes e assim poderá prever quando estes poderão abandona-la e assim oferecer vantagens para que isto não aconteça.
- Sequencias: Na técnica de mineração por padrões sequenciais os eventos estão ligados ao longo do tempo. Assim podendo descobrir que quando uma pessoa compra um carro dentro de um período de tempo curto a mesma, efetuara compra de pneus e também de um rádio para o seu automóvel.
- Aglomeração (*clustering*): A técnica de mineração por aglomeração funciona de forma semelhante a classificação, porem quando ainda não estiverem sido definidos os grupos. O algoritmo de aglomeração terá o trabalho de descobrir diferentes grupos dentro de uma grande quantidade de dados, como por exemplo encontrar grupos dentre usuários de cartões de credito com base na demografia e em inves- timentos pessoais.

2.2. Localizando padrões

Existem várias medidas objetivas para que se possa avaliar o grau de interesse que um padrão pode apresentar para o usuário. Estas medidas são baseadas na estrutura do padrão descoberto em estatísticas apropriadas. Por exemplo, uma medida objetiva para avaliaro interesse de uma regra de associação é o suporte, representando a porcentagem dastransações em um banco de dados de transações onde a regra se verifica.

3. Obtendo e preparando os dados

Os dados utilizados no presente estudo, foram obtidos juntamente ao portal brasileiro de dados abertos, que é mantido pelo governo federal [Brasil 2015].

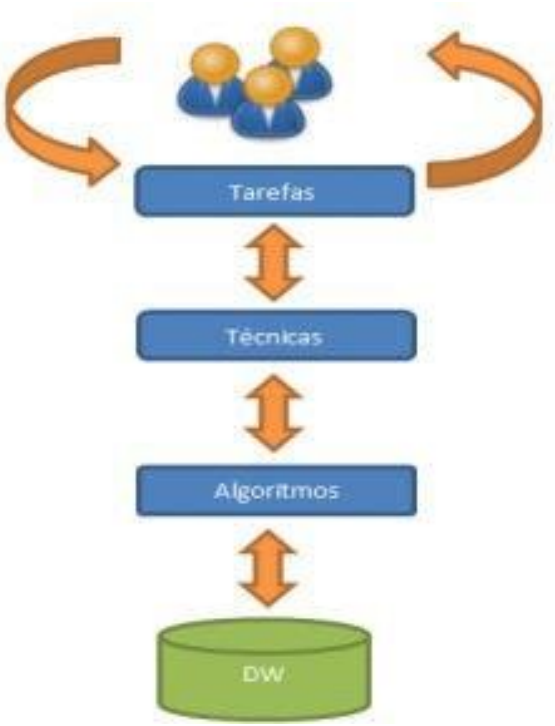

Figura 2. Interação entre os elementos da mineração de dados.

A população do estudo na presente pesquisa totalizou 7165 casos de internação hospitalar referente a doenças relacionadas ao saneamento ambiental inadequado a cada 100 mil habitantes [Brasil 2014]. As variáveis que foram utilizadas, são os números totais de internações por categoria de doença e ano de referência. Estes elementos podem ser observados através da Figura 3.

Doenças relacionadas ao saneamento ambiental inadequado (DRSAI) são doenças que podem estar associadas ao abastecimento de água deficiente, esgotamento sanitário inadequado, contaminação por resíduos sólidos ou condições precárias de moradia [Brasil 2014].

O período de tempo foi escolhido a partir de 1993, que foi quando o sistema único de saúde (SUS) passou a registrar as internações hospitalares de forma mais sistemática, até 2010, quando foram disponibilizados para o público através do portal brasileiro de dados abertos, que é mantido pelo governo federal.

A partir da obtenção dos dados, foi feito um pré-processamento das informações contidas neles e foram categorizados, isto é, definidas classes para determinados atributos ou variáveis em algumas informações como, doenças de transmissão feco-oral, doenças transmitidas por inseto vetor, entre outras. Esta preparação foi necessária para que as informações resultantes da análise sejam de melhor utilização.

4. Métodos

A escolha da técnica de mineração de dados foi feita após o pré-processamento dos dados, quando notou-se uma grande necessidade de se aglomerar os dados em grupos, para que assim possa ser feita uma melhor analise das informações. Assim então escolhendo a técnica de Aglomeração (*clustering*).

Figura 3. Gráfico de totais de internações.

Para a realização deste estudo, optou-se pelo uso do algoritmo K-Means do WEKA (*Waikato Environment for Knowledge Analysis*), que permite a descoberta destes padrões na forma de *clusters*.

Trabalhou-se com o programa WEKA, para a obtenção dos *clusters*, em virtude da sua adequação ao estudo e por três importantes razões:

- É uma ferramenta amistosa ao uso por profissionais de saúde, muitas vezes nãoafeitos à informática.
- Por ser *software* livre, pode ser utilizado sem custo e com grande facilidade nas secretarias de saúde.

Analise de *clusters* é o processo de aglomerar um conjunto de dados em classes de objetos similares. Um *cluster* é uma coleção de objetivo que são similares uns aos outros (de acordo com algum critério de similaridade pré-fixado) e diferentes a objetos pertencentes a outros *clusters*. Analise de *clusters* é uma tarefa de aprendizado não super- visionado, pelo fato de que os *clusters* representam classes que não puderam ser definidas no início do processo de aprendizagem, como é o caso das tarefas de classificação, onde o banco de dados de treinamento é composto de tuplas classificadas.

Na etapa seguinte do processo, que consistia na mineração de dados propriamente dita, foram extraídas regras que permitem relacionar as diversas variáveis sob estudo.

5. Aplicação do modelo escolhido

Para que pudesse ser obtido informações mais consistentes, foram feitas diversas analises dos dados, simulando universos com 2, 3, 4 e 5 *clusters*.

Os resultados gerados com base nestes parâmetros mostraram-se muito eficientes para o que era necessário se mostrar, assim foi feito uma analise sobre os resultados gerados pelo algoritmo, para que pudesse se escolher o que mais poderia mostrar ao usuário as informações desejadas. Assim foi observado que os universos com 3 e 4 *clusters* iriam suprir a necessidade deste estudo, porém para uma melhor e mais simples visualização dos dados, foi escolhido o universo com 3 *clusters* para que fossem feitas as análises mais a fundo. O universo escolhido pode ser observado através da Figura 4.

Cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(18)	(8)	(5)	(5)
Ano	2001.5	2006.25	2000.4	1995
Feco-oral	360.5889	284	312.6	531.12
Inseto vetor	30.7778	38.0625	21.38	28.52
Contato com a agua	2.5444	1.8875	2.6	3.54
Higiene	2.9556	1.875	0.68	6.96
Geo-helmintos e teniases	1.2333	0.6125	0.98	2.48

Figura 4. Centroids.

Através destas informações geradas, podemos observar que algumas das internações decorrente de doenças relacionadas com o saneamento ambiental inadequado esteve a diminuir durante o período analisado, tais como, doenças de transmissão feco- oral, transmitidas através do contato com a água, Geo-helmintos e teníases. Porem como pode ser ver nas figuras 5 e 6, as doenças transmitidas por inseto vetor e relacionadas com a higiene mostraram-se instáveis durante o período analisado, podendo ver em um primeiro momento entre os *clusters* 2 e 1 uma diminuição de ocorrências significativa, podem entre os *clusters* 1 e 0 ocorreu um aumento nestas ocorrências, assim podendo demonstrar que pode ter ocorrido algum fator externo ou que o saneamento ambiental referente a estes problemas não esta sendo tratado de forma correta pelas entidades cor- respondentes.

Também pode se notar a diminuição de ocorrências relacionadas a transmissão feco-oral teve uma grande queda.

Também vale ressaltar que as doenças relacionadas a transmissão feco-oral

tiveram uma redução substancial em suas ocorrências, mostrando assim que a entidade responsável para tratar do saneamento ambiental referente a esta área teve uma grande preocupação com a quantidade de casos que estavam acontecendo e assim puderam me- lhorar o seu serviço.

Figura 5. Doenças transmitidas por inseto vetor.

Figura 6. Doenças relacionadas com a higiene.

3. Conclusão e trabalhos futuros

Conforme as experiências realizadas nos dados obtidos, podemos notar que com o passar do tempo as internações decorrentes a doenças relacionadas com o saneamento ambiental inadequado, vem diminuindo em sua maioria, mostrando assim, que o País vem melhorando as suas práticas em relação a este problema que pode representar um grande risco para a população caso não tratado. Porem pode notar-se também que em relação a algumas doenças houve uma diminuição na sua ocorrência e após algum

período de tempo houve uma elevação dos casos, este ponto é de suma importância para a pesquisa, porque ela demonstra um problema que pode estar acontecendo em certas áreas, que o governo federal estava tratando, mas que por algum motivo não está dando tanta importância quando deveria.

Desta forma para que se possa ter mais detalhes sobre o problema em questão, devesse avaliar não somente a quantidade de internações relativas a este tipo de doença,mas também em que circunstancias as mesmas ocorreram, como período do ano, se a época que teve aumento na ocorrência foi um período chuvoso ou não e entre outras variáveis que seriam de grande ajuda em uma futura analise.

Os próximos trabalhos a serem executados devem levar em consideração mais variáveis a serem tratadas, para que assim possa se ter um universo mais amplo e assim poder gerar melhores informações e análises.

Referências

Brasil (2014). Visão geral da prestação de serviços de água e esgoto.

Brasil (2015). Portal brasileiro de dados abertos.

Laudon, K. and Laudon, J. (2011). *Sistemas de Informações Gerenciais: Fundamentos da inteligência de negócios: gestão da informação e de banco de dados*.

Neri, M. C. (2011). *Os emergentes dos emergentes : Reflexões globais e ações locais para a nova classe média brasileira*.

O'brien, J. A. (2011). *Sistemas de Informação e As Decisões Gerenciais Na Era da Internet*.

OMS (2009). *Global health risks - Mortality and burden of disease attributable to selec- ted major risks*.