1.03.99

APLICAÇÃO DA MINERAÇÃO DE DADOS EM REPOSITÓRIOS DINÂMICOS PARA A EXTRAÇÃO E ANÁLISE DE DADOS NA SAÚDE PÚBLICA BRASILEIRA

Mariana Esteves da Silva Pereira[1], Daniel Couto Gatti[2]
1. Estudante de Engenharia Biomédica da Faculdade de Ciências Exatas e Tecnologia da PUC-SP
2. Professor do Departamento de Computação e diretor do câmpus da FCET/PUC-SP

Resumo:

A manutenção e o desenvolvimento da saúde pública brasileira dependem da disponibilização de dados e informações. Hoje, essas informações são escassas, e não atendem aos padrões de dados abertos. A tendência mundial é a disponibilização em formato processável por máquina para que sejam analisados grandes conjuntos de informações por software.
A escassez de dados e informações impede que os recursos, que também são escassos, possam ser utilizados mais racionalmente.
O objetivo da pesquisa foi identificar e entender maneiras de transformar e padronizar os dados não estruturados contidos no DATASUS.
Pretende-se: (1) estudar as deficiências existentes no DATASUS e portal de dados abertos do governo brasileiro; (2) estudar um modelo de mineração de dados e disponibilização dos dados em formato processável por máquina; (3) criar uma solução técnica de mineração; e (4) pesquisar os conhecimentos que podem ser gerados com os dados.

Palavras-chave: mineração de dados, dados abertos, saúde pública.

Apoio financeiro: CNPq.

Trabalho selecionado para a JNIC pela instituição: PUC/SP.

Introdução:

Os dados abertos são dados publicados na Web, que qualquer pessoa pode utilizar, reutilizar e distribuir. Portanto, são importantes para difundir informações para a sociedade de forma transparente, contribuindo com o aprimoramento dos dados públicos de uma população, por exemplo.

Em se tratando de saúde, esses dados podem, por exemplo, mapear doenças em determinadas regiões ou de forma nacional, auxiliar em pesquisas, revelar taxas de natalidade, e como consequência deste último, refletir fatores sociais e fisiológicos, já que a fertilidade feminina e masculina não são os únicos determinantes para a quantidade de nascimentos anuais.

A manutenção da saúde de um país depende muito da análise da situação atual e passada, pois assim é possível planejar ações que visam melhorar a qualidade de vida da população.

No Brasil existem iniciativas de disponibilização de dados acerca da saúde, como o Departamento de Informática do SUS – DATASUS. Mas esses dados são escassos, assim como os recursos para extrai-los, além de não possuírem a padronização adotada internacionalmente.

A dificuldade encontrada é que geralmente esses dados são volumosos, e em muitas situações a extração ainda é feita de forma manual, o que torna o processo ineficiente e impede o processamento por máquina.

A pesquisa objetiva identificar e entender maneiras de transformar os dados não estruturados contidos no DATASUS em dados abertos disponíveis para consulta e análise de interessados.
Como avaliação, propõe-se desenvolver um modelo de implementação que minere automaticamente dados sobre a saúde pública brasileira, disponibilize-os em formatos processáveis por máquina, além de estudar conhecimentos que podem ser extraídos

desses dados.

Metodologia:

Após a realização da pesquisa bibliográfica, prosseguimos para o estudo da linguagem de programação Python, muito utilizada no contexto da mineração de dados. O estudo da linguagem incluiu o aprendizado da sintaxe da linguagem e de bibliotecas utilizadas para a mineração.
O próximo passo da pesquisa consistiu em escolher o domínio de onde os dados deveriam ser extraídos para a análise. Existem duas grandes fontes de dados públicos sobre a saúde no Brasil: o Portal Brasileiro de Dados Abertos e o DATASUS. Inicialmente, na proposta inicial desta pesquisa, havíamos planejado extrair dados relacionados à área pública hospitalar, no entanto, ambos os referidos portais não possuiam estes conjuntos de dados disponíveis para download e, consequentemente, análise. Por este motivo, realinhamos o foco da pesquisa para saúde pública brasileira, área em que encontramos dados disponíveis.
A próxima etapa consistiu de uma análise dos formatos de disponibilização dos dados desses portais. Constatamos que o Portal Brasileiro de Dados Abertos já atende aos padrões e princípios de disponibilização dos dados abertos. Portanto, nossas próximas análises concentraram-se no DATASUS.
Seguimos para o estudo de um modelo de implementação em Python, que fosse flexível e adaptável para a extração de dados de diversos portais, se necessário. Para atender a este requisito, estudamos dois paradigmas das linguagens orientadas a objeto: tipagem do pato (em inglês, duck-typing) e o tradicional polimorfismo das linguagens tipadas como Java e C#.
Após a realização da implementação inicial de uma solução em Python, fizemos um estudo para o entendimento dos dados referentes aos nascidos vivos na cidade de São Paulo do portal DATASUS. Seguimos então para o refinamento da solução e consequente obtenção dos dados para uma base de dados local.
Ao final desse processo, fizemos uma análise mais abrangente que as permitidas pela interface do DATASUS como prova de conceito de nossa solução.

Resultados e Discussão:

Para este trabalho foi feita uma extração de dados de nascidos vivos de 1994 e 2014, no estado de São Paulo, segundo os municípios onde as mães moram, e a idade das mesmas. O gráfico mostra que a quantidade total de nascimentos diminuiu. Essa redução pode ser explicada por alguns fatores como: o maior custo para criar filhos, o maior número de mulheres que trabalham fora de casa e desenvolvem carreiras profissionais, maior acesso a tratamento médico, saneamento básico e programa de vacinação, métodos contraceptivos mais difundidos, entre outros.
Outra informação que podemos extrair do gráfico é que atualmente o número de mulheres que têm filhos até uma idade mais avançada aumentou cerca de 32%.
Alguns fatores que contribuíram para esta estatística são o avanço da medicina, que permite tratamentos sofisticados para fertilidade em mulheres mais velhas, e o crescente número de mulheres que trabalham fora e têm filhos mais tarde, fatos que não aconteciam com a mesma intensidade em 1994.

Conclusões:

Os dados abertos contribuem de forma significativa em diversos contextos da sociedade promovendo informações transparentes aos cidadãos, mas a área da saúde no Brasil ainda não utiliza esse recurso de forma eficaz.
Na área da saúde, dados abertos podem auxiliar no estudo de doenças por região ou por estação do ano mais frequente, gerar dados importantes para o avanço de pesquisas, revelar taxas de natalidade e mortalidade, auxiliando o censo nacional e consequentemente o planejamento social.
O Departamento de Informática do SUS, que deveria disponibilizar esses dados de forma que todo cidadão possa acessar e compartilhar, não os disponibiliza de forma processável por máquina.
O cadastro e a extração de dados de forma manual trazem desvantagens por serem processos lentos, que permitem maior margem de erro por parte das equipes interessadas, e dificuldade de armazenamento e consulta posterior.
A automatização desses processos traz benefícios como maior velocidade, facilidade de armazenamento e consulta, maior disponibilização de dados para a população, e consequente avanço na melhora da qualidade da saúde do país.
O estudo da automatização foi feito em linguagem Python, a fim que o entendimento acerca de softwares baseados em Mineração de Dados fosse satisfeito. Assim, o DATASUS foi escolhido como domínio de onde os dados

seriam extraídos, e foi feito um estudo sobre os paradigmas de duck-typing e polimorfismo com a finalidade de deixar o software flexível e adaptável para a extração de dados de diversos portais.
Com a implementação do software em Python foi feito um estudo sobre os dados extraídos da quantidade de nascidos vivos do estado de São Paulo de 1994 e de 2014, em função da idade das mães.
Com esses dados foi possível gerar um gráfico que revelou informações sobre a fecundidade e a natalidade brasileiras, como por exemplo, é possível constatar que o número total de nascimentos diminuiu aproximadamente 10% (de 685705, em 1994, para 616587 em 2014) em 20 anos, além de que hoje em dia as mulheres têm filhos com idades mais avançadas que antigamente.

Referências bibliográficas

APTE, Chidanand; GROSSMAN, Edna. Probabilistic Estimation-based Data Mining for Discovering Insurance Risks IEEE Intelligent Systems & their applications, Los Alamitos/CA, p. 49-58, nov./dec. 1999.

BERRY, Michael J. A.; LINOFF, Gordon. Data Mining Techniques 2nd Edition : Wiley, 2004.

CARLANTONIO, Lando Mendonça Di. Novas Metodologias para Clusterização de Dados. 2001. 157 f. Dissertação (Mestrado em Ciências em Engenharia Civil) – Universidade Federal do Rio de Janeiro: COPPE/UFRJ, Rio de Janeiro, 2001.

CHAN, Philip K.; FAN, Wei et al. Distributed Data Mining in Credit Card Fraud detection. IEEE Intelligent Systems & their applications, Los Alamitos/CA, p.67-74, nov./dec. 1999.

CORTÊS, Sérgio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. Mineração de Dados – Funcionalidades, Técnicas e Abordagens. Mai/2012. Disponível em: < ftp://ftp.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf> Acesso em: 17 Mai 2016

DATASUS. Nascidos Vivos. Disponível em: <http://tabnet.datasus.gov.br/cgi/sinasc/Nascidos_Vivos_1994_2012.pdf> Acesso em: 01 ago 2016

DECRETO Nº 7.508, DE 28 DE JUNHO DE 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/decreto/D7508.htm> Acesso em 25 ago 2016

DRUMMOND, Isabela Neves. Implementação do método de classificação contínua fuzzy k-médias no ambiente TerraLib. São José dos Campos, 2003. Disponível em: <http://www.dpi.inpe.br/cursos/ser300/Trabalhos/isabela.pdf> Acesso em 17 mai 2016

ELMASRI, Ramez; NAVATHE, Shankant B. Fundamentals of Database Systems: Addison Wesley, 2000.

FAYYAD, Usama M. Advances in knowledge Discovery and Data Mining. Menlo Park /CA: AAAI Press, 1996. 131

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. Data Mining: um guia prático. Rio de Janeiro/RJ: Campus, 2005.

OPEN KNOWLEDGE INTERNATIONAL. What is open? Disponível em <https://okfn.org/opendata/ >. Acesso em 15 mai 2016

PINTO, Eriane Nascimento. ESCALA DE APGAR. Disponível em: <http://www.uff.br/disicamep/escala_de_apgar.htm > Acesso em 22 jun 2016

RESENDE, Solange Oliveira. Sistemas Inteligentes: Fundamentos e Aplicações. Barueri/SP: Manole, 2003.

SMITH, Elmi; ELOFF, Jan. Cognitive Fuzzy Modeling for Enhanced Risk Assessment in Health Care Institution. IEEE Intelligent Systems & their applications, Los Alamitos/CA, p. 69-75, mar./abr. 2000.Souto, M. C. P.

Weka: Aprendizado de Máquina. Site da faculdade UFRN. Disponível em <http://www.dimap.ufrn.br/~marcilio/AM/course-AM.htm>. Acesso em 31 mar 2016.

WITTEN, Ian H.; Frank, Eibe. Data Mining : practical machine learning tools and techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers, 1999.

WITTEN, Ian H.; FRANK, Eibe. Data Mining: practical machine learning tools and techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers, 2000.