encontro da computação 2014
E-COMP 14

Visualização de Dados de Informações Extraídas da Web - Um Estudo sobre Conteúdo Popular Japonês

Gabriel Fontenelle Senno Silva[1]
Centro Universitário Senac - Campus Santo Amaro

Resumo

Este trabalho apresenta um estudo sobre conteúdo popular japonês por meio de visualizações geradas com dados extraídos automaticamente de websites com grande quantidade de dados.

Introdução

Visualização de dados é a comunicação visual de informações, com base em um conjunto de dados. Crawler é um programa que navega automaticamente em websites e extraí dados de páginas determinadas. Este trabalho apresenta visualizações de dados, desenvolvidos a partir de dados extraídos de websites com o uso de sistema de *crawling*. Foram escolhidos websites sobre conteúdo popular japonês. A cultura popular japonesa é conhecida pelo desenvolvimento de animações, revistas em quadrinhos e gêneros literários influenciados por um estilo de desenho único focado nas expressões de suas personagens.

Objetivo

Este trabalho apresenta um estudo sobre conteúdo popular japonês, exibindo informações com dois tipos de visualizações de dados: nuvem de palavras e nuvem de bolhas. As informações são obtidas a partir de dados extraidos com *crawling* de websites definidos.

Websites

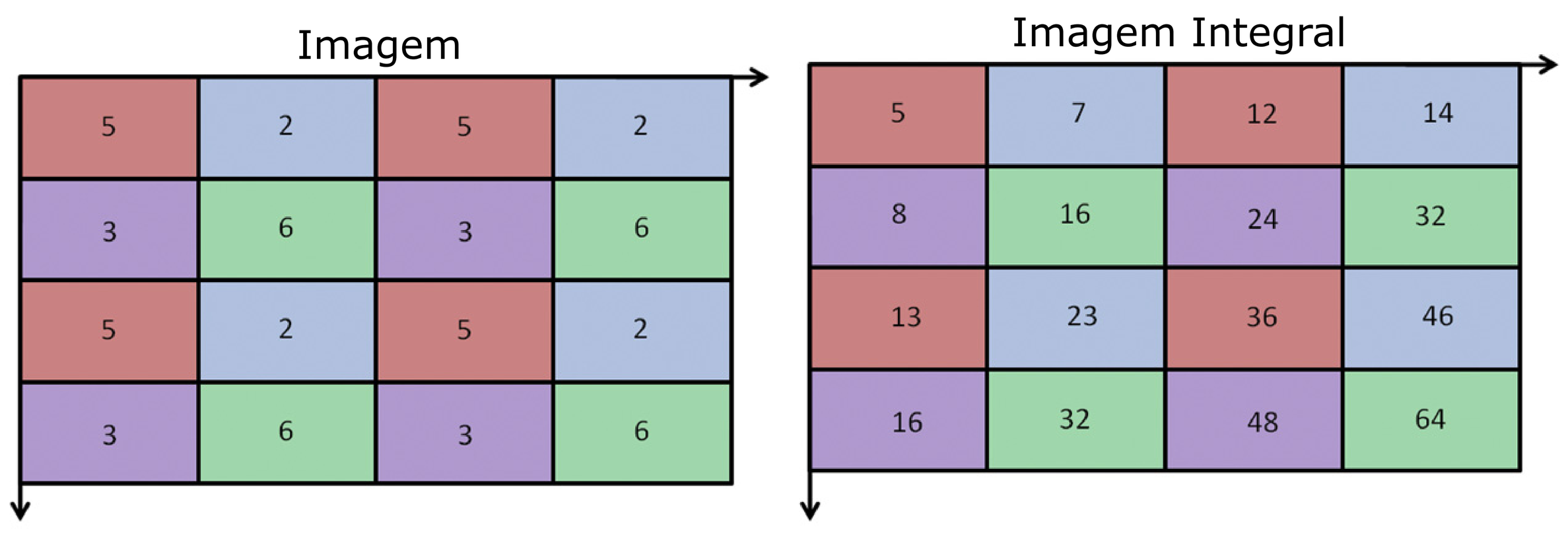
Foram escolhidos pela popularidade e pela quantidade de itens que possuem os websites abaixo:

- http://www.mangaupdates.com/
- http://myfigurecollection.net/
- http://animecharactersdatabase.com/

Para armazenar os dados um extenso banco de dados foi modelado, antes que o sistema de *crawling*, com a biblioteca Scrapy, pudesse ser utilizado. Durante a execução do crawler os dados foram normalizados e inseridos no banco de dados PostgreSQL.

Metodologia

Área ocupada por pixels pode ser cálculada rapidamente utilizando a estrutura de dados Imagem Integrada. Se a área calculada em determinado local for maior que zero ela está ocupada. Cada valor na Imagem Integral equivale ao resultado da soma de todos os elementos das linhas e das colunas anteriores de matriz utilizada em sua criação.



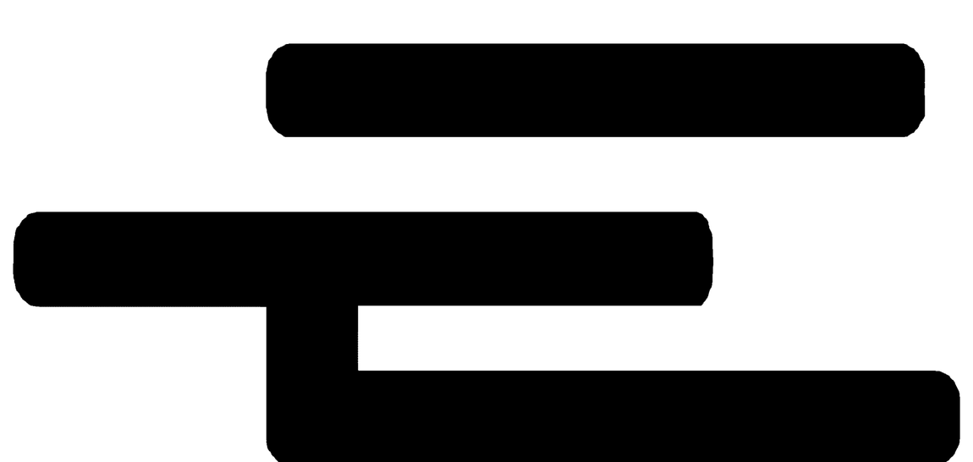
Para calcular a área com a Imagem Integral se utiliza a seguinte fórmula:

$$área = I(i,j) + I(i+1,j+1) - I(i+1,j) - I(i,j+1). \quad (1)$$

I representa a Imagem Integral e (i,j) representam as linhas e colunas.

Obtendo uma máscara

Como espaços disponívels são iguais a zero, utilizando uma imagem com ilustração na cor preta, por ter valor zero, pode ser obtida uma máscara que restringirá a posição de palavras ou círculos. A seguir exemplo de imagem máscara:



Metodologia

Definindo o tamanho dos textos e dos círculos

$$tamanho = \frac{c * ranking}{\ln(ranking + 10)} \quad (2)$$

Onde:

- c é a razão entre a área útil da máscara e a quantidade de objetos a serem inseridos.

Para evitar divisão por zero é somado o valor 10 ao número *ranking* antes do cálculo do Logaritmo Natural.

Com o tamanho definido dos objetos (textos ou círculos), verificamos na Imagem Integral da máscara os espaços disponíveis e determinamos uma posição entre eles. Adicionamos portanto o objeto a imagem máscara e geramos uma nova Imagem Integral. Repetimos esse procedimento para cada objeto retirado do banco de dados. E por fim salvamos uma imagem com todos os objetos posicionados.

Resultados: Nuvens de palavras e de bolhas

Figura: Franquias com tamanho definido pela quantidade de itens que possuem

Figura: A esqueda representação da quantidade de itens por franquia. A direita representação das profissões com maior número de envolvidos na produção de itens da cultura popular japonesa.

Conclusão

Pode-se supor com as informações obtidas, que entre a imensa quantidade de produtos produzidos no Japão a maioria é de caráter literário devido a um alto nível de educação no país.

Referências Bibliográficas

P. Viola; M. Jones.
Rapid object detection using a boosted cascade of simple features. In IEEE Computer Vision and Pattern Recognition (pp. I:511–518), 2001.

Konstantinos G. Derpanis.
Integral image-based representations, Department of Computer Science and Engineering, York University, New York, 2007.

Marco Lui; Timothy Baldwin.
langid.py: An Off-the-shelf Language Identification Tool, Department of Computing and Information Systems, University of Melbourne, VIC 3010, Australia.

Fredrik Lundh.
The Python Imaging Library Handbook, 2014, disponível em http://effbot.org/imagingbook/