Um Sistema Distribuído para Predizer Informações de Usuários em Redes Sociais

Pedro Garcia

21 de outubro de 2013

Objetivo (problema)

- Prever informações de usuários.
- Objeto de estudo: estimar idade do usuário com base em informações secundárias (formação, empregos, etc).

Solução Proposta

- Utilizamos um estimador da informação desejada utilizando dois níveis.
- O primeiro utiliza uma estimação inicial, buscando em todo domínio de dados, minimizando-se a distância de Bhattacharyya.
- Após a distância inicial, busca-se refinar a estimativa, minimizando-se a divergência de Kullback–Leibler.

Solução Proposta

- Inicialmente, computamos a frequência de cada palavra que o usuário de uma rede social produza.
- Em seguida, agrupamos cada frequência de acordo com o tipo da informação do usuário (emprego, educação, etc), gerando um conjunto de distribuições distintas para cada informação que desejamos estimar.
- Então, computamos a distância de Bhattacharyya para cada uma das distribuições computadas.
- Finalmente, teremos um conjunto de estimativas para cada um dos tipos de informações estimadas, buscando qual distribuição corresponda com a distribuição do usuário, cuja idade deseja-se estimar.

Solução Proposta

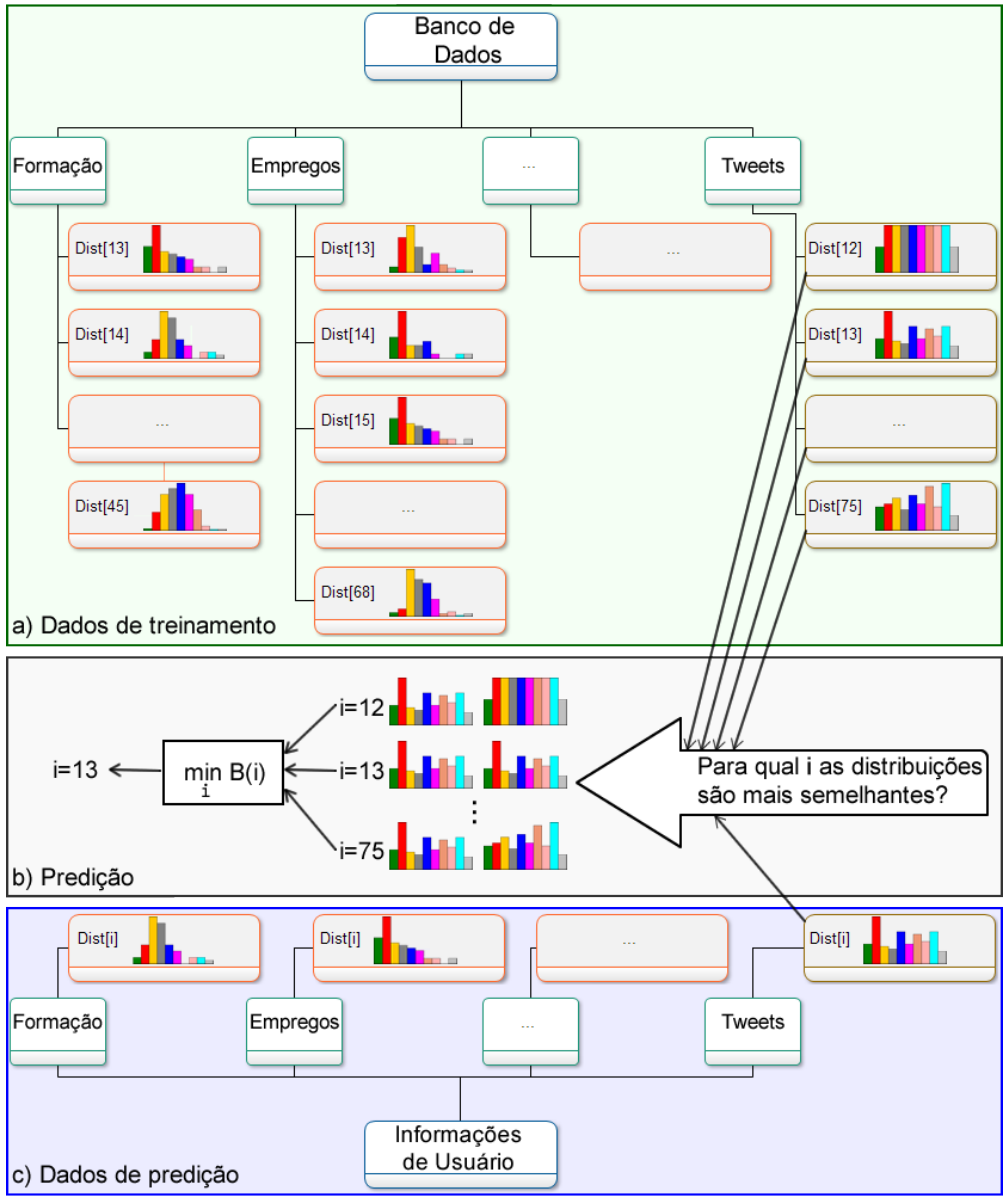

Figura : Correspondência de Distribuições.

Solução Proposta

- Dentro desse conjunto de Informações de Usuários estimadas, buscamos quais dessas informações são mais "confiáveis".
- O critério que usamos para isso é a mínima Divergência de Kullback–Leibler.

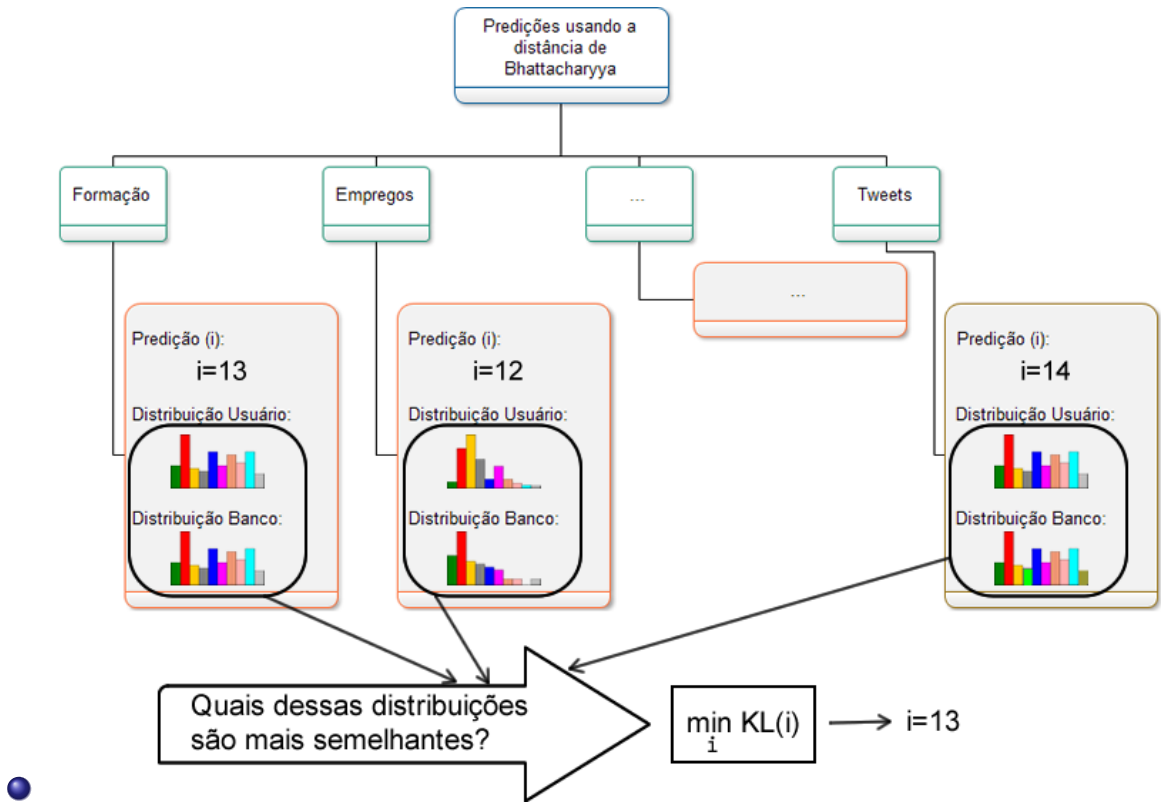

Figura : Seleção da Predição.

Distribuição da Solução Proposta

- Redes sociais envolvem, geralmente, um grande volume de dados.
- Isso implica em uma grande quantidade de dados a ser analisada.
- Logo, uma abordagem distribuída para esse sistema é desejável.

Distribuição da Solução Proposta

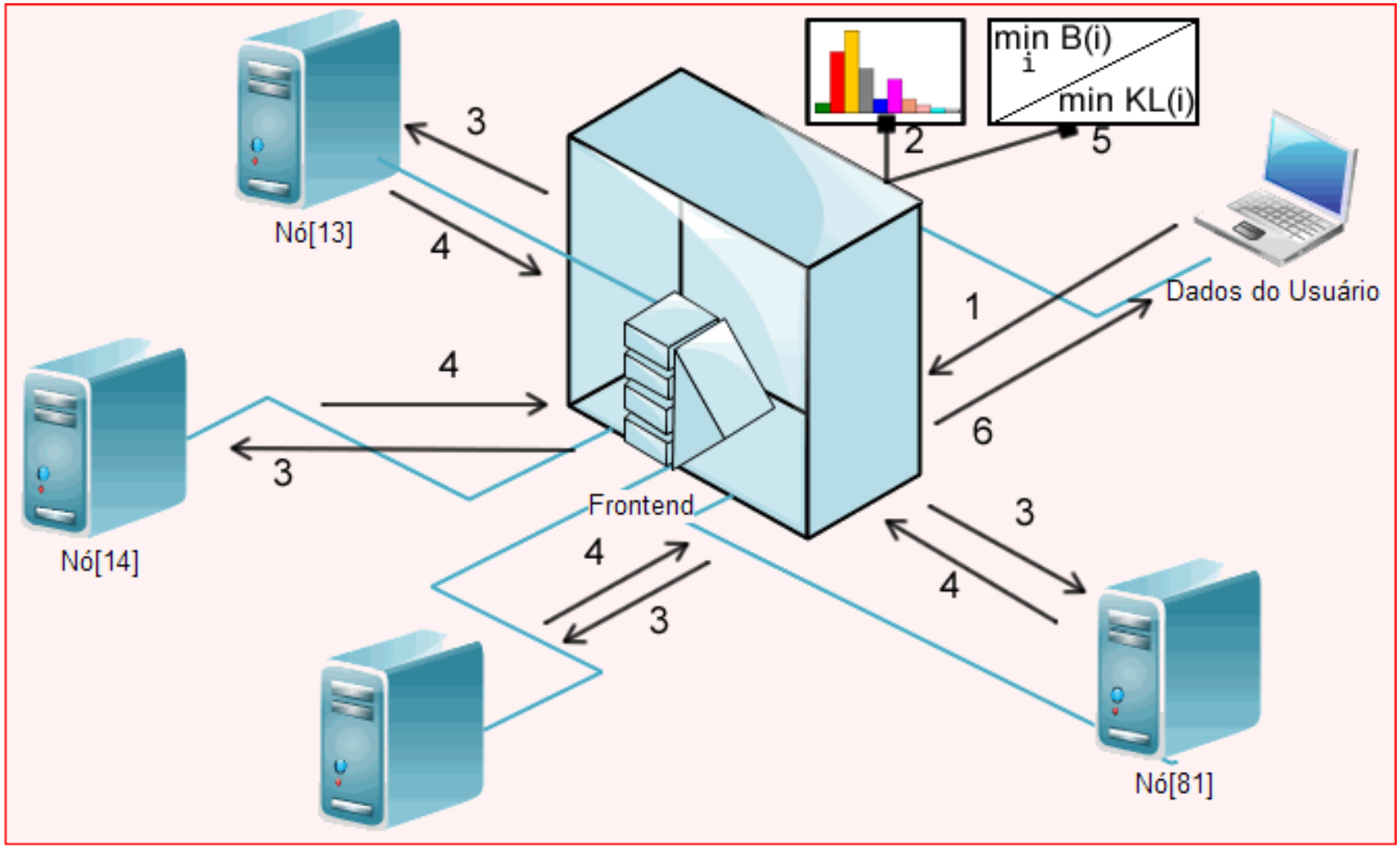

Figura : Distribuição das Tarefas ao Longo dos Nós.

Resultados

- Erro quadrático médio: 1.4545

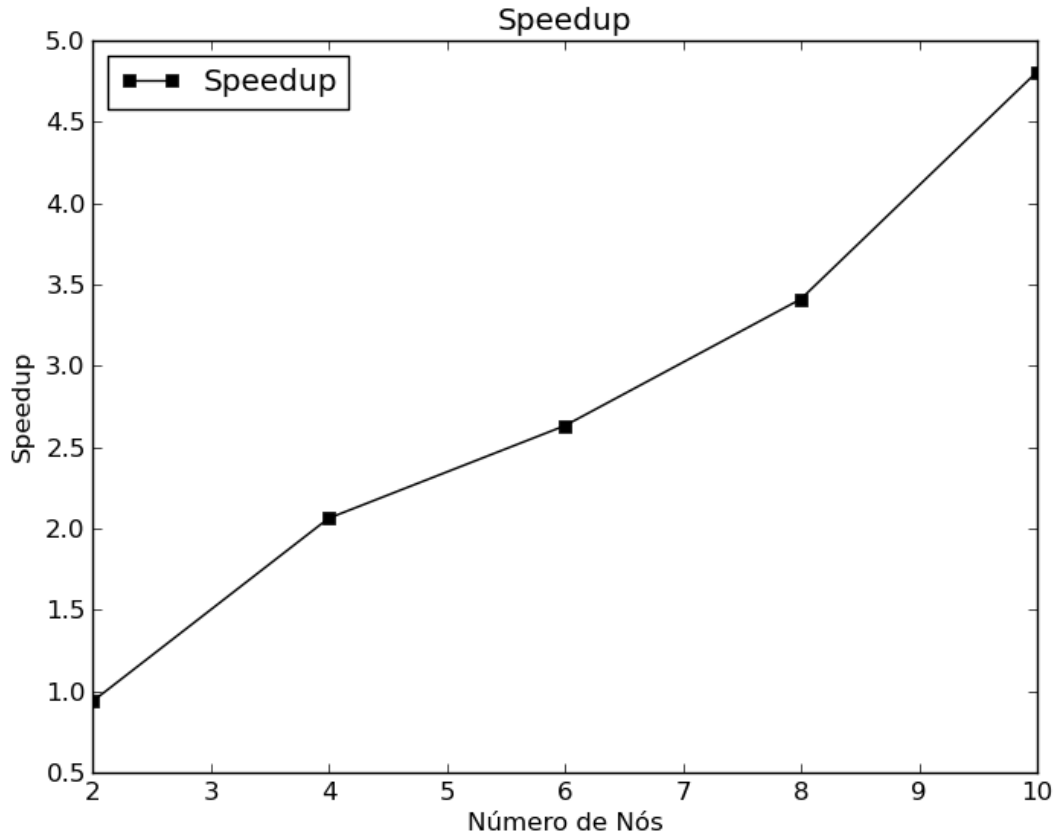

Figura : Speedup.

Conclusão

- O sistema utiliza uma estratégia de *distribution matching*, utilizando somente dados de frequência das palavras.
- A solução não utiliza informações semânticas na estimação, o que torna o algoritmo independente da linguagem.
- Pela Figura 4 é possível notar como o speedup cresce conforme o aumento do número dos nós, mostrando que o tempo total de solução do problema diminui, conforme o esperado.
- Melhorias desse trabalho consistirão em
  - combinar os resultados estimados,
  - analisar a capacidade de predição para outros tipos de informações e
  - comparar com outros trabalhos semelhantes.