Um Sistema Paralelo para Predizer Informações de Usuários em Redes Sociais

Pedro Garcia Freitas*, Márcio A. Silva Souza*, Aletéia P. F. de Araújo*,
Li Weigang*, Érico Marx P. Fonseca*, and Mylène C.Q. Farias+
*Department of Computer Science,
+Department of Electrical Engineering,
University of Brasília (UnB),
Campus Universitário Darcy Ribeiro, 70919-970 Brasília, DF - Brazil
Emails: sawp@sawp.com.br, mass@ime.usp.br, aleteia@cic.unb.br, weigang@unb.br, ericofis@gmail.com, mylene@ieee.org

***Resumo*—Neste trabalho é proposto um método para fins de classificação de informações obtidas em redes sociais por meio de um classificador de estágio múltiplo. Esse classificador, estruturado em dois níveis, utiliza dados obtidos em redes sociais para estimar informações de um usuário de acordo com um critério de classificação. No caso, o critério de informação escolhido e investigado foi a idade, embora o método possa ser facilmente adaptado para estimar outros tipos de informações. O classificador utiliza a distância de Bhattacharyya e a divergência de Kullback-Leiber para relacionar informações coletadas em redes sociais com as informações inseridas para um usuário que se deseja estimar a idade. Como esse tipo de aplicação envolve um grande volume de dados, neste trabalho também é apresentado a estratégia para distribuição e computação dos dados utilizando o método proposto.**

***Keywords*-Distributed systems; Information prediction; Bhattacharyya distance; Kullback–Leibler divergence; Classification model;**

I. Introdução

A mídia social é uma experiência interativa que permite que um usuário esteja altamente conectado a uma rede, para o intercâmbio de informações em larga escala e em tempo real. Há muitas razões que envolvem o atual sucesso das mídias sociais, entre quais podemos destacar: (i) a integração de muitos usuários em uma única rede compartilhada; (ii) a propagação da informação através de uma ampla variedade de tipos de conteúdo (texto, áudio, imagem, vídeo, etc.); (iii) permite a comunicação através vários tipos de plataformas como *smartphones* e notebooks; (iv) chamando a atenção de grandes investimentos empresariais que impulsionam o desenvolvimento de melhorias para o sistema. Devido a estas capacidades, as mídias sociais são capazes de disseminar as informações em maior quantidade e com maior rapidez [1].

Nesta linha de frente, serviços de microblogs como Sina Weibo e Twitter estão ficando cada vez mais populares [2], [3]. O Sina Weibo é um microblog chinês semelhante ao Twitter e Facebook e é um dos sites mais populares na China, em uso por mais de 30% dos usuários de Internet, com uma penetração de mercado semelhante ao que o Twitter estabeleceu nos EUA.

O Sina Weibo implementa vários recursos do Twitter. Os usuários podem postar usando mensagens de texto com até 140 caracteres, falar ou conversar com outras pessoas, adicionar *hashtags*, seguir outras pessoas, suas mensagens aparecem na própria linha do tempo dos usuários, encaminhamento de postagem (semelhante à função de *retweet* do Twitter), colocar uma postagem na lista de favoritos, verificar se a conta do usuário é uma celebridade, etc. Aplicativos oficiais e de terceiros tornam os usuários capazes de acessar o Sina Weibo de outros sites ou plataformas.

Com base neste cenário, a 14ª Conferência Internacional em Sistema de Informação Web Engenharia (WISE 2013) propôs desafio com base no Sina Weibo – Weibo Prediction Track (T2). Esse desafio visava a análise automática do conteúdo dos usuários com a finalidade de predizer informações não fornecidas por eles. No caso, o desafio sugeriu que a faixa etária de um usuário fosse estimada a partir de outros dados secundários do usuário, tais como escolaridade, empregos, conteúdos produzidos, etc. Sendo assim, os organizadores o WISE alvitraram um importante problema de inteligência artificial, que é a análise de características de conteúdo.

A análise de características conteúdo é uma metodologia de ciência social que concerne na "descrição objetiva, sistemática e quantitativa do conteúdo da comunicação"[4, p. 140]. Ela consiste na técnica de codificação de conteúdos simbólicos (textos, imagens, etc) usadas na comunicação a partir de características estruturais do conteúdo (largura da mensagem, distribuição de informações em um texto, componentes de imagens, etc) [5].

No contexto da Internet, a análise de característica pode ser usada para determinar informações que ajudem a aprimorar a experiência do usuário, sistemas de *marketing* direcionado, prever falhas de segurança, etc. No caso de redes sociais, mais especificamente, esse tipo de técnica pode favorecer ao usuário a busca por informações de interesse, sugerindo relacionamentos ou indicando conteúdos.

No contexto deste trabalho, a análise de conteúdo pode ser usada para estimar informações de idade e, por consequência, faixa etária do usuário. Um exemplo de aplicação dessa aplicação está na filtragem de conteúdo exposto para o usuário, permitindo que um sistema como o Twitter não exiba informações impróprias para uma dada faixa etária.

Este trabalho investiga o problema é como estimar a idade e a faixa etária de usuários de uma rede social. No

caso, o foco da pesquisa foi a rede Weibo. Contudo, como uma rede social desse porte possui um grande número de usuários e conexões, o que implica em um grande volume de dados produzidos que precisam ser analisados e tratados. Portanto, recursos de computação distribuída são fundamentais nesse contexto.

Para tanto, na Seção II é apresentado uma breve revisão sobre trabalhos relacionados, na Seção III, é descrito o conjunto de dados, e é apresentada as características do problema. Na Seção IV, é feita a modelagem do problema, através de uma proposta de solução, que na Seção V será tratada através de um sistema distribuído. Já na Seção VI são apresentados os experimentos e resultados que servirão de base para a conclusão apresentada na Seção VII.

II. Trabalhos Relacionados

Devido a semelhança entre Weibo e Twitter, uma pesquisa anterior com foco no Twitter, pode ser usada para analisar o perfil dos usuário no Weibo. Por exemplo, Lau e Li [6] propõem um novo algoritmo de extração de características, a fim de obter informações a partir do conteúdo produzido pelos usuários. Da mesma forma, Hue et al. [7] fornecem um conjunto de técnicas e métodos que podem ser utilizados para extrair a informação de *microblog*.

Uma ferramenta usada para extrair informações de *microblog* é o Twitalyzer [8], que fornece uma análise das atividades de qualquer usuário do Twitter, com base no sucesso do usuário nas mídias sociais. Outra ferramenta, TwitterStats [9], é muito popular e útil para revelar o comportamento de todos os usuários do Twitter, que reúne os dados de atividade e os apresenta em gráficos coloridos. As informações analisadas por esta ferramenta incluem os *tweets* por período (dia, semana e hora), a quantidade de *tweets* de repostas e a quantidade de retweets, as pessoas que mais interagiram com o usuário, etc. Além disso, infere as palavras que mais contribuiram para a popularidade do conteúdo. Finalmente, outra ferramenta, Brandtweet [10] é muito útil.

Além dos artigos mencionados na seção anterior, o trabalho feito para a extração de informações de outras fontes também pode ser usado e adaptado para resolver o problema da estimativa da idade em Weibo. Na obra intitulada *Mining the Blogosphere*: idade, gênero e as variedades de auto-expressão, Argamon et al. [11] analisam os principais fatores que contribuem para prever informações linguísticas, tais como sexo e idade a partir do conteúdo dos utilizadores. Também com o objetivo de estimar a idade e o sexo de um usuário a partir do conteúdo produzido, Heerden et al. [12], propôs um método que combina métodos de regressão com métodos de classificação. Contudo,embora este trabalho seja focado na inferência de informações a partir de dados coletados em uma conversa, é possível adaptá-lo para prever alguns serviços no *microblog*.

Outro trabalho com o objetivo de estimar a idade em redes sociais foi desenvolvido por Dey et al. [13]. Neste caso, os autores desenvolveram um método para calcular o ano de nascimento dos usuários do Facebook. Burger et al. [14] explorou vários métodos de classificação para estimar o sexo dos usuários do Twitter em seu trabalho intitulado Discriminação de Gênero no Twitter, eles demonstraram que o grau de precisão da estimativa varia de acordo com a quantidade de informação produzida pelo usuário. Embora o foco principal deste trabalho tenha sido de prever o sexo do utilizador, ele pode ser adaptado para inferir outras informações, uma vez que os resultados mostram uma significativa de desempenho, com elevada taxa de sucesso.

III. Conjunto de Dados

Para descrever e analisar os dados deste trabalho, primeiramente, foi necessário obter os dados a partir do Weibo. Após a obtenção desses dados, os IDs dos usuários foram anonimizados, os conteúdos de relação entre usuários e *retweets* foram removidos, preservando apenas os conteúdos originais. Esses conteúdos são escritos em língua chinesa, onde cada palavra possui um caractere único para representá-la. Assim, cada palavra distinta é mapeada em um número único.

O conjunto de dados em questão consiste em um identificador de usuário (*user ID*), um conjunto de rótulos definidos pelo usuário (*labels*), uma lista de empregos, descrição pessoal, data de nascimento, gênero, educação e conteúdo produzido (*tweets*). Além disso, o conteúdo produzido agrega informações sobre data e hora de publicação. Tais informações são importantes, pois servem como parâmetros de treinamento e estimativa de informações.

A. Caracterização do Problema

O objetivo do problema é construir um sistema que permita estimar a idade e, por consequência, a faixa etária de usuários de uma rede social. No caso, o objeto de estudo foi a rede Weibo. Assim, define-se quatro faixas etárias distintas (R1, R2, R3 e R4). Cada uma dessas faixas etárias representam um intervalo de idade, que são $[0, 18]$, $]18, 24]$, $[24, 35]$ e $[35, \infty]$, conforme ilustrado na Figura 1.

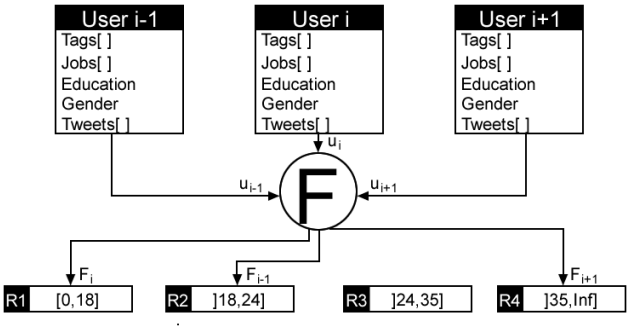

Figura 1. Esquema do problema de classificação para a faixa etária.

O problema é caracterizado da seguinte forma: 1) Há uma entrada (domínio de F) com diversos parâmetros e 2) Há uma saída (imagem de F) com poucas opções. Isso é claramente um problema de classificação [15]. Ou seja, o problema é como classificar um conjunto de n objetos

de acordo com um conjunto de m classes, decidindo a qual classe pertence cada objeto. Em outras palavras, o problema consiste em encontrar a função de classificação $F : P \rightarrow G$, onde P é o conjunto de n objetos e G é o conjunto de m classes. Assim, a melhor escolha para a classificação de um objeto leva em conta dois custos: o de alocação e o de separação.

Usando inteligência artificial, o problema pode ser resolvido em dois grandes passos os quais são: a fase de treinamento e a fase de predição. Para isso, é necessário cumprir três objetivos, que são a regularização dos dados a serem processados, a extração das características dos dados regularizados, e a utilização de um modelo de classificação para treinamento e predição da idade a partir dos dados obtidos.

1) Passo 1 – Regularização dos dados: como os dados coletados são mal-formatados, é necessário regularizá-los em uma representação robusta, eficientemente tratável e compatível com consultas. Portanto, um sistema gerenciador de banco de dados é altamente recomendável.

2) Passo 2 – Extração de características: as características extraídas do conjunto de treinamento deve ser em função de duas quantidades, a informação fornecida e o conteúdo produzido pelo usuário. Assim, a informação fornecida é caracterizada pelo conjunto de palavras fornecidas para cada categoria (trabalho, educação, etc), e o conteúdo é caracterizado pela frequência de cada palavra, normalizada pela largura do conteúdo.

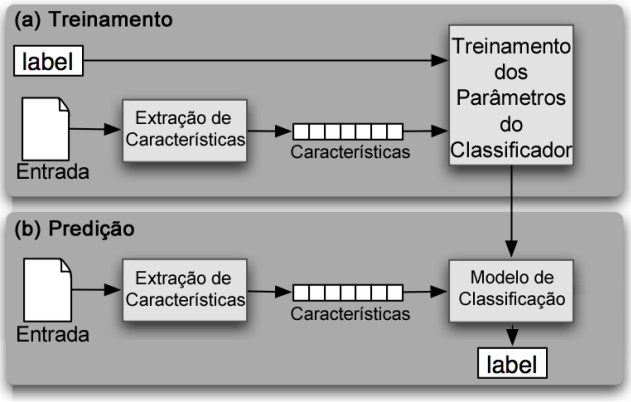

Figura 2. Etapas do problema.

3) Passo 3 – Predição de idade: a predição é feita utilizando-se um classificador na forma $F : P \rightarrow G$, onde o conjunto G é formado pelas idades e o conjunto P pelas características extraídas. A função de classificação F é ajustada na fase de treinamento, utilizando-se um conjunto G_t e outro P_t, onde as informações são conhecidas. Após o ajuste desta função, ela pode ser utilizada para estimar a idade a partir dos demais parâmetros, conforme ilustrado na Figura 2. No contexto deste trabalho, a função de classificação é feita minizando-se a distância de Bhattacharyya [16], conforme apresentado na próxima seção.

IV. Solução Proposta

Neste trabalho, implementou-se um método de classificação em duas etapa. A primeira etapa para estimar as idades mais prováveis com base em cada tipo de informação disponível pelo usuário. A segunda etapa consiste em analisar as idades mais prováveis, estimadas na primeira classificação, escolhendo-se apenas uma idade de acordo com as informações do usuário. Nesta abordagem, em cada etapa é selecionado um conjunto de classes (idades) que irá definir os parâmetros associados na etapa seguinte.

As classes selecionados são aquelas que possuem menor separabilidade com as classes criadas na etapa de treinamento. Em outras palavras, utiliza-se a distribuição das informações fornecidas pelo usuário com a distribuição das informações coletadas dos outros usuários e procura-se, dentre essas, quais são as que mais se assemelham. A classe é definida pela maior semelhança entre as distribuições.

O critério adotado para estimar a separabilidade entre as classes é a distância de Bhattacharyya [16]. A forma geral da distância de Bhattacharyya entre duas classes c_1 e c_2 é dada por

$$D(c_1, c_2) = -\ln\left(\int_{-\infty}^{\infty} \sqrt{P(x|c_1)P(x|c_2)}\,\mathrm{d}x\right),$$

onde $P(x|c_k)$ é a probabilidade *a priori* da informação x pertencer à classe c_k. Supondo que as distribuições das informações fornecidas pelos usuários serão, normalmente, distribuídas, pode-se utilizar a seguinte expressão:

$$D(c_1, c_2) = \frac{1}{4}\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2} + \frac{1}{2}\ln\left(\frac{\left|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}\right),$$

onde μ_i e Σ_i são os vetores de médias e a matriz de covariância da classe c_i. Além disso, o primeito termo da equação acima fornece a separabilidade das classes pelas médias, e o segundo termo fornece a separabilidade pela covariância entre as distribuições.

A distância de Bhattacharyya é uma métrica conveniente para a estimar a separação de pares de classes, pois para dados, normalmente distribuídos, ela fornece um limite superior para o erro de Bayes [16]. Portanto, este trabalho consiste em agrupar as informações em equipes rotuladas conforme a idade, buscando-se a distância de Bhattacharyya mínima entre as distribuições geradas na fase de treinamento e a distribuição de teste. Quando essa distância for ótima (mínima), a idade associada à distribuição de treinamento é, provavelmente, a idade predita.

A. Agrupamento dos Dados

Conforme exposto na Seção III, os dados coletados são divididos entre diferentes tipos (*tags*, educação, empregos, *tweets*, etc). Cada um desses tipos gera um conjunto de distribuições de dados independente. Além disso, para cada um desses tipos, a distribuição dos dados de acordo com as classes (idades) é distinta. Em outras palavras, os dados são hierarquizados em uma árvore de três níveis,

onde a raiz representa todo o conjunto de dados (banco de dados), o segundo nível representa o conjunto dos tipos de informação, e o terceiro nível contém os conjuntos das distribuições das informações rotuladas pela idade. Essa hierarquia está ilustrada no diagrama da Figura 3-(a).

Para gerar o agrupamento, filtra-se no banco de dados de treinamento todos os dados correspondentes à cada um dos tipos de informação. Neste ponto, tem-se que todos os dados de treinamento estão agrupados em conjuntos de acordo com os tipos de informação. Em seguida, percorre-se cada um desses conjuntos, separando os dados de acordo com a idade. Então, para estes dados filtrados por tipo e idade, computa-se a frequência de ocorrência de cada palavra, salvando-se a distribuição gerada de acordo com um rótulo único. Dessa maneira, as distribuições geradas podem ser rotuladas pelo tipo e pela idade. .

Em outras palavras, essa fase de treinamento utiliza os dados que contém as idades para gerar essa árvore de agrupamentos de forma que os grupos possam ser acessíveis fela função $F : X \rightarrow G$, onde $X \subset \{tipo, idade\}$ e G é a distribuição correspondente. Após esses grupos serem gerados nessa fase de aprendizagem, é necessário utilizar um modelo de classificação para predizer qual a idade a partir das informações do usuário. Esse modelo de classificação ocorre buscando-se a distância de Bhattacharyya ótima.

B. Minimização da Distância de Bhattacharyya

Após o aprendizado – que consiste na geração, agrupamento em conjuntos e indexação das distribuições dos dados de acordo com tipo e idade – entra em questão o problema da predição. Isto é, como utilizar as distribuições dos dados de treinamento para estimar informações de um usuário qualquer.

O primeiro passo para a predição consiste em extrair as características a partir dos dados do usuário cuja idade deseja-se estimar, conforme ilustrado na Figura 2-(b). No caso, essas características são extraídas da mesma forma que as características de treinamento. Assim, a partir dos dados do usuário, separam-se os diferentes tipos de informação (educação, empregos, conteúdos dos *tweets*, etc).

As características extraídas das informações de usuários são, por sua vez, agrupadas de forma semelhante à hierarquia ilustrada na Figura 3-(a), como pode-se notar na Figura 3-(c). A diferença neste caso é que a idade é desconhecida.

Assim, o problema é encontrar qual deve ser o parâmetro $idade$ que permite encontrar a distribuição que melhor relacione os dados do usuário com os dados treinados (correspondência de distribuições). Em outras palavras, seja $F_t(t, i) \rightarrow G_t$ a função que mapeia a distribuição G_t nos dados de treinamento, onde t é o tipo da informação (educação, trabalhos, *tweets*, etc) e i é a idade correspondente nos dados de treinamento. Supondo uma função correspondente para mapear os dados do usuário, tal que $F_u(t, i) \rightarrow G_u$, onde t é o tipo da informação e i é a idade que mapeia a distribuição G_u, então o problema consiste em encontrar i. Ou seja,

$$\min_i \quad B(i),$$

onde

$$B(i) = D(F_t(t, i), F_u(t, i)) = D(G_t, G_u)$$

é a distância de Bhattacharyya das distribuições G_t e G_u.

A Figura 3 ilustra esse estágio de resolução do problema para as informações dos *tweets*. Na parte (a) dessa figura, tem-se os agrupamentos gerados na fase de treinamento, como explicado na Seção IV-A. Na parte (c) dessa figura, tem-se as características (distribuições) extraídas dos dados do usuário. A predição, ilustrada na parte (b), é feita calculando-se todas as distâncias de Bhattacharyya entre distribuições treinadas e a distribuição do usuário. Entre essas distâncias, a que for menor estará associada à idade predita.

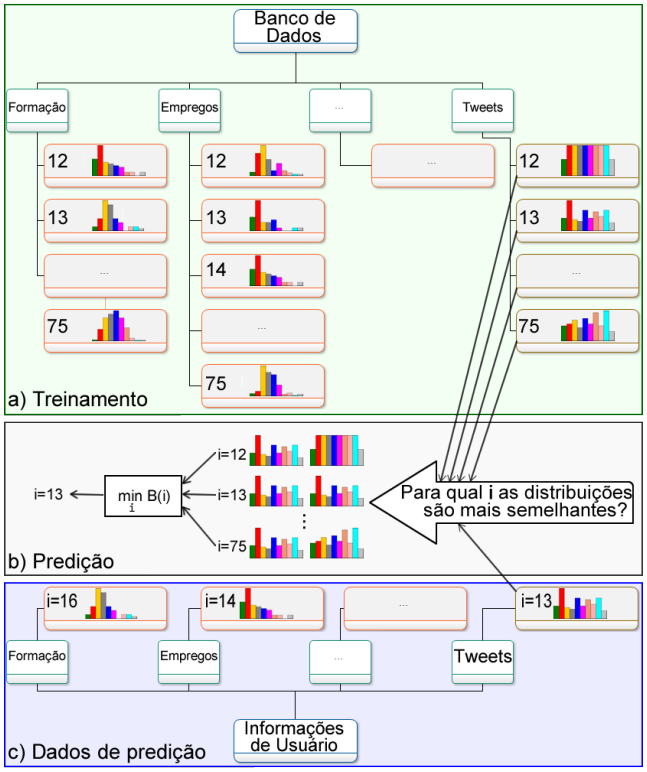

Figura 3. Correspondência de Distribuições.

Como é possível notar, essa etapa de classificação é feita de forma individual para cada um dos tipos de informação do usuário. Assim, há um conjunto de estimativas individuais para cada um dos tipos, ou seja, há um conjunto de predição para cada um dos tipos. Assim, deve-se utilizar um segundo nível de classificação para se escolher qual a melhor estimativa entre essas predições.

C. Escolha da Predição

Após o processamento das distâncias de Bhattacharyya para cada um dos tipos de informação, haverá um conjunto contendo diversas estimações, cada uma associada à um tipo. Dessa formar, após a primeira etapa de classificação,

descrita na Seção IV-B, as informações filtradas serão os tipos, associados às idades preditas; as distribuições do usuário; e a distribuição correspondente no banco de treinamento.

A partir desse conjunto de informações, utiliza-se as distribuições dos dados do usuário para verificar qual delas possui mais informações em comum com o banco de treinamento. A seleção da melhor escolha é feita tomando-se a predição que possui a menor divergência de Kullback-Leiber [17]. Em outras palavras, toma-se

$$\min_i \quad KL(i),$$

onde

$$KL(i) = \frac{1}{n}\left[-\sum_x G_t(x)\log G_u(x) + \sum_x G_t(x)\log G_t(x)\right],$$

sendo n o número de elementos da distribuição dos dados de usuário G_u para um dado tipo. Por essa métrica, G_t representa a "verdadeira" distribuição das informações, enquanto G_u representa a aproximação de G_t. Note que o fator da largura da distribuição n é inserido como um peso. Isso é feito para favorecer os campos em que o usuário inseriu mais informação. A ideia é que quanto mais informações o usuário inserir, mais classificável em um grupo ele pode ser.

A Figura 4 ilustra o processo de seleção. Conforme pode ser notado a partir dela, após encontrada a distância de Bhattacharyya para cada um dos tipos (formação, empregos, *tweets* etc), há uma predição e um par de distribuições associados a cada um dos tipos (distribuição das informações do usuário e distribuição correspondente àquela idade, salva no banco treinado). Para cada um desses pares, computa-se qual a divergência de Kullback-Leiber de cada um deles. A que tiver a menor divergência, é escolhida como a predição mais correta (segunda correspondência de distribuições).

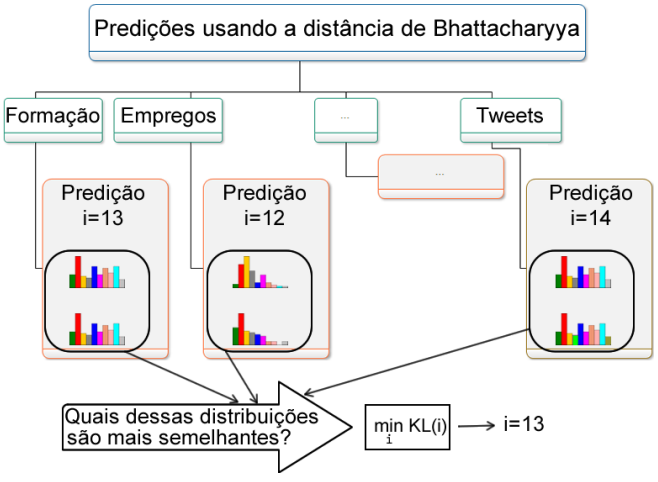

Figura 4. Seleção da Predição.

Os cálculos envolvidos nesta solução são simples e não são muito custosos. Contudo, as predições são mais precisas quanto mais dados puderem ser utilizados no modelo de predição. Portanto, para tratar com essa grande quantidade de dados, é conveniente se trabalhar com processamento paralelo a fim de reduzir o tempo de processamento e utilizar de forma mais eficiente recursos de memória e armazenamento.

V. Distribuição da Solução Proposta

O algoritmo de predição da idade a partir de diferentes dados coletados em uma rede social descrito na Seção IV possui propriedades que favorecem a computação paralela. Isso fica claro observando-se a Figura 3, uma vez que os dados utilizados são, inicialmente, segregados por informação e a geração das distribuições são computadas de forma independente para cada um dos tipos. Além disso, na primeira fase de classificação, o cálculo da distância de Bhattacharyya é realizado para todos os grupos de idade, também de forma independente uma da outra.

Como trata-se de uma aplicação em que os cálculos são simples, mas que envolvem grande quantidades de dados, a estratégia principal consiste em distribuí-los. A estratégia mais simples de distribuição da informação consiste em construir um banco de dados hierárquico. Essa hierarquia pode obedecer àquela ilustrada na Figura 3. Nesse caso, há um nó responsável por indexar cada tipo de informação, podendo manter outros nós filhos, cada um armazenando os dados correspondentes a cada uma das idades.

Essa estratégia é simples de se implementar e possui a vantagem de tornar a aplicação mais escalável. Em outras palavras, a aplicação pode permitir que a inserção de novos tipos de informação sem que haja necessidade de se re-estruturar o banco nos nós. Contudo, como em redes sociais existem menos classes de tipos de informação do que classes de idade, é conveniente que o banco divida as informações conforme essa última.

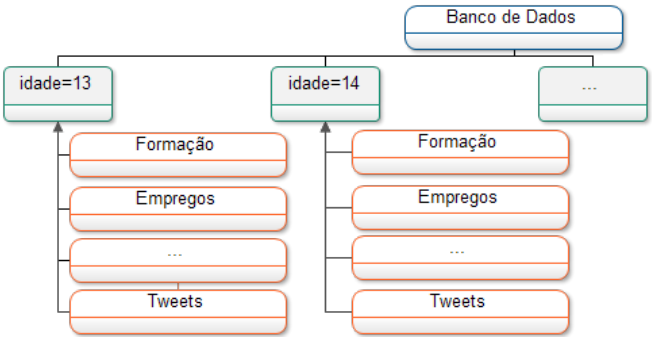

Figura 5. Organização do Banco ao Longo dos Nós.

Portanto, é conveniente que o banco de dados seja distribuído conforme ilustrado na Figura 5. Dessa forma, reduz-se os registros em cada nó, pois cada nó possuirá um subconjunto do total de tuplas envolvidas no processamento, o quer permite que também o cálculo das distâncias entre as distribuições seja feito mais paralelamente, uma vez que cada nó precisa computar apenas m distâncias, onde m é o número de tipos de informações distintas. O esquema do banco de dados (que define a quantidade de tabelas e sua estrutura) é fixo e replicado em cada nó. Essa solução caracteriza um banco de dados distribuído de fato com algo semelhante a uma "fragmentação horizontal".

Para gerar esse banco de dados distribuído é necessário replicar a estrutura das tabelas correspondentes aos tipos em cada um dos nós. Assim, ao coletar os dados antes do treinamento, envia-se ao nó correspondente a idade. Após isso, computa-se as distribuições em cada nó, mantendo-se os dados treinados organizados ao longo dos nós, os quais são organizados conforme a idade, assim como mostrado na Figura 5.

Após a distribuição dos dados e o treinamento calculado de forma independente em cada nó, há o problema da predição. Como a solução proposta possui duas fases de predição, onde a primeira consiste em gerar um conjunto de possibilidades de acordo com os tipos de informação e a segunda consiste em utilizar esse conjunto para verificar qual é a informação mais confiável, há uma clara dependência de dados.

Essa dependência de dados ocorre porque a seleção da segunda fase requer informações da fase anterior. Contudo, essa dependência não é muito forte, uma vez que os cálculos para computar a divergência de Kullback-Leiber não dependem do resultado da distância de Bhattacharyya. Sendo assim, uma abordagem a ser tomada é processar todas as distâncias de Bhattacharyya e as divergências de Kullback-Leiber entre as distribuições de todos os tipos de dados para todas as idades (em todos os nós).

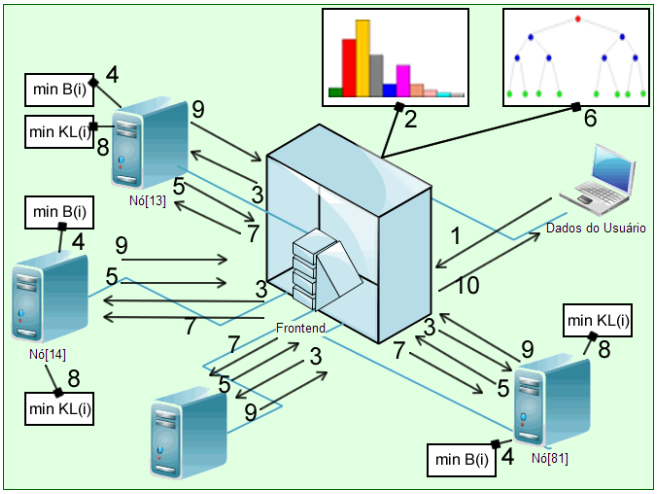

Figura 6. Distribuição das Tarefas ao Longo dos Nós.

A Figura 6 ilustra como isso pode ser feito em dez passos. Nesse caso, (1) o usuário envia os dados ao *frontend*, (2) que fica responsável por extrair as características para, então, (3) enviar aos nós as distribuições geradas. Em seguida, (4) os nós computam as distâncias de Bhattacharyya; (5) retornando para o *frontend* as informações; (6) o *frontend* coleta essas distâncias de todos os nós, decidindo quais delas são ótimas. Em seguida, (7 e 8) ele requisita apenas aos nós correspondentes para que eles calculem as divergências de Kullback-Leiber somente para o tipo de informação necessário. Os nós, por sua vez, (9) respondem com os valores das divergências solicitadas. Por fim, (10) o *frontend* verifica qual estimativa corresponde à menor divergência e retorna a predição ao usuário da aplicação.

VI. Experimentos e Resultados

Com o intuito de testar o modelo proposto, foi implementado o esquema de distribuição ilustrado na Figura 6. Nesse caso, os algoritmos foram implementados na linguagem Groovy, que compila e executa em ambiente Java. Além disso, a execução dos testes foi executado em dois ambientes. Para computar as predições sem distribuição de carga, utilizou-se uma máquina Dell PowerEdge R710 16xQuad-Core Intel Xeon E5630 (2.53GHz) com 141GB RAM. Para computar as predições utilizando a abordagem distribuída, foi utilizado um conjunto de 10 computadores HP Compac 8200 com processadores Intel Core i5-2500S (2.70 GHz) com 8GB RAM.

Os serviços descritos para os nós foram disponibilizados para o *frontend* por meio de uma interface de Transferência de Estado Representativo (REST) [18]. Isto é, após o *frontend* extrair as características das informações do usuário, ele envia as informações aos nós, que rodam um serviço para calcular as distâncias de Bhattacharyya. As distâncias computadas são devolvidas em formato JSON [19] pelos nós. Após a decisão ser feita pelo *frontend*, ele verifica em quais nós deve computar as divergências de Kullback-Leiber. Então, os nós apropriados recebem essas requisições e respondem por meio de um segundo serviço.

O banco de dados utilizado foi o PostgreSQL. Nesse caso, a estrutura das tabelas dos nós foram replicadas em todos os nós. Em um estágio anterior ao processamento das simulações, a coleta feita dos dados é feita de forma que distribua as informações para os nós de acordo com a idade. Portanto, a fase de treinamento já é executada com os dados devidamente organizados. Além disso, para simplificar as simulações, os dados foram filtrados para que houvesse a mesma quantidade de classes de idade e de nós disponíveis.

Para analisar o desempenho do método proposto, foram utilizadas tanto métricas para qualidade da predição quanto para a computação distribuída. Para medir o desempenho da predição, as métricas escolhidas foram a *média quadrática* [20, p. 82], *precisão* [20, p. 75], *recall* [20, p. 75] e *F-Measure* [20, p. 82]. O *speedup* foi utilizado para medir a redução do tempo na computação distribuída do algoritmo.

As métricas de desempenho das predições foram feitas escolhendo-se um conjunto de dados contendo todas informações do usuário, incluindo a idade. Esse conjunto de dados, por sua vez, não foi utilizado no estágio de treinamento, a fim de se evitar o sobre-ajuste de dados. Como os dados de teste contém a informação que se deseja predizer (a idade), é possível verificar tanto o acerto da predição (se ela foi ou não exata), quanto o tanto que a predição desviou do valor correto.

O acerto exato da predição, ou seja, se o algoritmo acertou a idade ou não, é medido pela *precisão* e pelo *recall*. Nesse caso, *recall* é a frequência do número de registros relevantes obtidos do número total de registros da base de dados, geralmente expresso como um fator

percentual. A *precisão* é o número de registros relevantes obtidos em relação ao número de registros irrelevantes obtidos, também expresso como uma percentagem. Em outras palavras,

$$P = \frac{vp}{vp + fp}$$

e

$$R = \frac{vp}{vp + fn},$$

onde P e R são, respectivamente, a precisão e o *recall*. As variáveis vp, fp, e fn representam os *verdadeiros positivos*, *falsos positivos* e *falsos negativos*.

Uma medida que sintetiza tanto a precisão quanto o *recall* é a *F-Measure*, que é a média harmônica ponderada entre elas:

$$F = \frac{1}{\frac{\alpha}{P} + \frac{(1-\alpha)}{R}},$$

onde $\alpha \in [0, 1]$ é o fator de balanço dos pesos das medidas. Geralmente, $\alpha = 0.5$.

Além das predições exatas (se o algoritmo acertou a idade ou não), foi medido o quanto a predição se afastou da idade real do usuário. Para isso, foi utilizado a média quadrática da diferença entre os valores preditos e os valores reais. Mais precisamente,

$$rmsd = \sqrt{\frac{\sum_{k=1}^{n}(x_p(k) - x_o(k))^2}{n}},$$

onde $x_p(k)$ e $x_o(k)$ são os valores preditos e originais do k-ésimo usuário.

Os valores obtidos para as métricas de foram:

- Erro quadrático médio: 1.4545,
- Recall: 0.3644,
- Precisão: 0.5,
- F-Measure: 0.4215

Esses valores indicam que não houve uma relevante taxa de valores preditos que não corresponderam à idade exata do usuário. Contudo, pelo valor do erro quadrático médio, as estimativas ficaram bem próximas da idade verdadeira. Para aplicações onde deseja-se classificar usuários dentro de uma faixa etária, o método proposto é aceitável.

O desempenho da computação dos dados distribuídos foi medido pela razão entre o tempo (em horas) necessário para resolução do problema em uma única máquina e o tempo necessário ao se dividir as tarefas ao longo dos nós. Essa medida é conhecida como *speedup*.

Pela Figura 7 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.

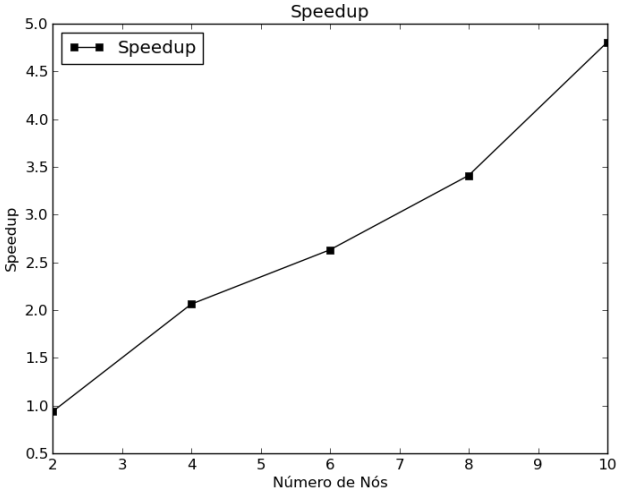

Figura 7. Speedup.

VII. Conclusão

Neste trabalho foi proposto um método para estimar informações a partir de dados de redes sociais. Tendo como foco na estimação da idade dos usuários a partir de informações secundárias, foi proposto um modelo que utiliza classificação estocástica.

Para isso, utilizou-se dois níveis de classificação. O primeiro deles busca a correspondência entre distribuições em um banco de dados treinado. A métrica de análise dessas correspondências foi a distância de Bhattacharyya. O segundo nível visa escolher uma única predição entre as melhores predições feitas pelo primeiro nível, verificando qual tipo de dados secundários (educação, empregos, etc) melhor descreve o perfil do usuário.

Além disso, considerando que as aplicações relacionadas à redes sociais envolvem grandes volumes de dados, neste trabalho foram apresentadas técnicas para se processar de forma paralela o algoritmo proposto. Para isso, foi apresentada uma estratégia de distribuição, que foi implementada para análise dos resultados.

Estudos visando aprimorar os critérios de seleção das predições são necessários, uma vez que a precisão exata da idade não é muito alta, conforme mostra os valores de *recall* e de precisão. Contudo, conforme os resultados apresentados na Seção VI, o método é satisfatório para aplicações que visam classificar e predizer a faixa etária de um dado usuário. Isso se justifica pelo baixo erro quadrático médio, indicando que as predições são próximas das idades reais dos usuários usados nos testes.

Trabalhos futuros envolvendo teoria da informação são recomendados. Assim, técnicas que permitam associar as informações das predições feitas no primeiro nível, ao invés de excluir parte dessas informações, devem ser incorporadas no segundo nível de predição a fim de aprimorar a acurácia média geral.

Pesquisas futuras também devem ser realizadas no sentido de aprimorar as estratégias de computação distribuída. A comparação da eficiência entre a primeira e a segunda estratégia de paralelização em diferentes contextos pode ser investigada. Além disso, a implementação de um *pipeline* é recomendável, caso necessita-se utilizar a aplicação proposta como serviço para o usuário final.

Agradecimentos

Este trabalho foi possível graças ao apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico

(CNPq), à Fundação de Empreendimentos Científicos e Tecnológicos (Finatec), e ao DPP - University of Brasília (UnB).

REFERÊNCIAS

[1] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07. New York, NY, USA: ACM, 2007, pp. 29–42. [Online]. Available: http://doi.acm.org/10.1145/1298306.1298311

[2] HubSpot. (2013, Jun.) State of the twittersphere. [Online]. Available: http://bit.ly/sotwitter/

[3] TheBritishChamber. (2013, Jun.) The growing influence of China's Weibo. [Online]. Available: http://britishchamber.cn/content/growing-influence-china\T1\textquoterights-weibo

[4] R. Campbell, C. Martin, and B. Fabos, *Media and Culture: An Introduction to Mass Communication*. Bedford/St. Martin's, 2011.

[5] M. W. Bauer, "Classical content analysis: A review," *Qualitative researching with text, image and sound*, pp. 131–151, 2000.

[6] C. H. Lau, Y. Li, and D. Tjondronegoro, "Microblog retrieval using topical features and query expansion," in *TREC*, E. M. Voorhees and L. P. Buckland, Eds. National Institute of Standards and Technology (NIST), 2011.

[7] W. Hua, T. D. Huynh, S. Hosseini, J. Lu, and X. Zhou, "Information extraction from microblogs: A survey," *International Journal of Software and Informatics*, vol. 6, no. 4, pp. 495–522, 2012.

[8] Twitalyze. (2013, May) Twitalyze. [Online]. Available: http://www.twitalyzer.com/

[9] Ttweetstats. (2013, May) Ttweetstats. [Online]. Available: http://www.tweetstats.com/

[10] B. statistics. (2013, May) Brandtweet statistics. [Online]. Available: http://stats.brandtweet.com/

[11] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Mining the blogosphere: Age, gender and the varieties of self-expression," *First Monday*, vol. 12, no. 9, 2007.

[12] C. J. van Heerden, E. Barnard, M. H. Davel, C. van der Walt, E. van Dyk, M. Feld, and C. A. Müller, "Combining regression and classification methods for improving automatic speaker age recognition," in *ICASSP*. IEEE, 2010, pp. 5174–5177.

[13] R. Dey, C. Tang, K. W. Ross, and N. Saxena, "Estimating age privacy leakage in online social networks," in *INFOCOM*, A. G. Greenberg and K. Sohraby, Eds. IEEE, 2012, pp. 2836–2840.

[14] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1301–1309. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145568

[15] G. Hjorth, "Classification problems in mathematics," *University of California, Berkeley*, 2010.

[16] E. Choi and C. Lee, "Feature extraction based on the bhattacharyya distance." *Pattern Recognition*, vol. 36, no. 8, pp. 1703–1709, 2003. [Online]. Available: http://dblp.uni-trier.de/db/journals/pr/pr36.html#ChoiL03

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[18] R. T. Fielding, "REST: architectural styles and the design of network-based software architectures," Doctoral dissertation, University of California, Irvine, 2000. [Online]. Available: http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

[19] D. Crockford, "Rfc4627: Javascript object notation," 2006.

[20] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.