Um Sistema Distribuído para Análise de Recurso de Conteúdo para Prever Informações de Usuários em Mídias Sociais

Érico Marx P. Fonseca*, Pedro Garcia Freitas*, Aletéia P. F. de Araújo*,
Li Weigang*, 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: ericofis@gmail.com, sawp@sawp.com.br, aleteia@cic.unb.br, weigang@unb.br, 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 grande volumes de dados, neste trabalho também é apresentado estratégias 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 as 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 por meio de 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 com maior quantidade, maior rapidez e efetividade do que as tradicionais mídias.

Com base neste cenário, a 14ª Conferência Internacional em Sistema de Informação Web Engenharia (WISE 2013) propôs desafio com base no Sina Weibo – T2 – Weibo Previsão Track. Esse desafio é baseado em dados reais. No entanto, uma rede desse porte tem 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.

O objetivo deste trabalho é a construção de um sistema distribuído para avaliar a idade de usuários de uma rede social. No caso, o foco da pesquisa foi a rede Weibo. Para tanto, na Seção II, é descrito o conjunto de dados, e é apresentada as características do problema. Na Seção III, é feita a modelagem do problema, através de uma proposta de solução, que na Seção IV será tratada através de um sistema distribuído. Já na Seção V são apresentados os experimentos e resultados que servirão de base para a conclusão apresentada na Seção VI.

II. Conjunto de Dados

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 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 $[35, \infty]$.

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.

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. No contexto deste trabalho, a função de classificação é feita minizando-se a distância de Bhattacharyya [5], conforme apresentado na próxima seção.

III. Solução Proposta

Neste trabalho, implementou-se um método de classificação em duas etapas. 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 selecionadas 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 [5]. 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. Agrupamento dos Dados

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 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

O primeiro passo para a predição consiste em extrair as características a partir dos dados do usuário cuja idade deseja-se estimar. 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 1-(a), como pode-se notar na Figura 1-(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.

A Figura 1 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 III-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.

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 forma, após a primeira etapa de classificação, descrita na Seção III-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

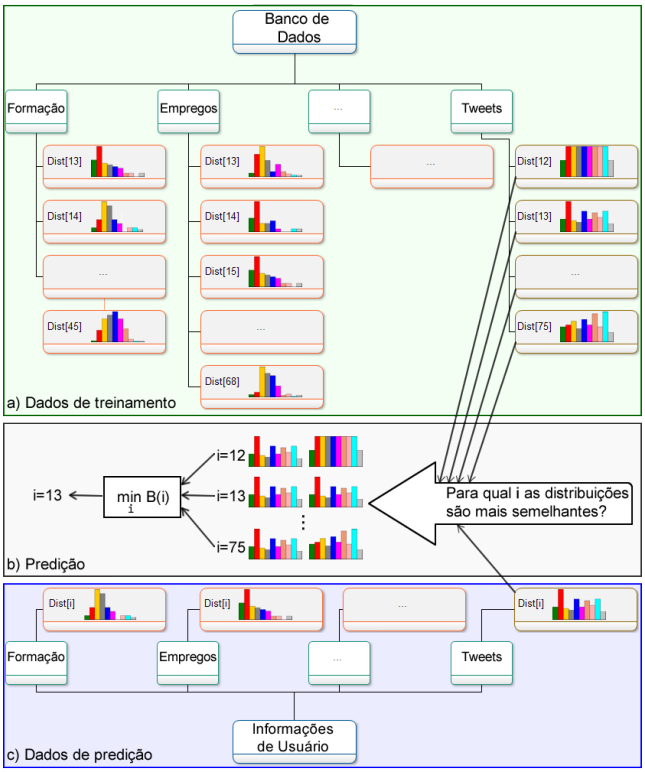

Figura 1. Correspondência de Distribuições.

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 [6]. 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 2 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 distirbuição 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).

Os cálculos envolvidos nesta solução são simples e não envolvem grande esforço computacional. Contudo, as

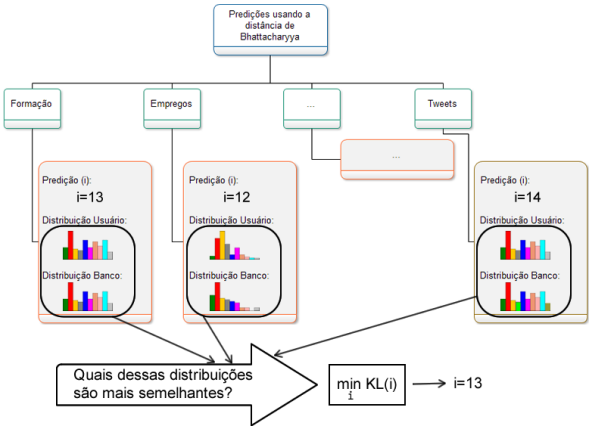

Figura 2. Seleção da Predição.

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.

IV. Distribuição da Solução Proposta

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. 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.

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).

A Figura 3 ilustra como isso pode ser feito em seis passos. Nesse caso, (1) o usuário envia os dados ao *frontend*, (2) que fica responsavel por extrair as características; e (3) enviar as distribuições para os nós. Em seguida, (4) cada nó processa tanto a distância quanto a divergência, (5) devolvendo ao *frontend*. Após a coleta de todas essas informações, o *frontend* realiza os dois passos de classificação, buscando entre todas as idades, qual fornece a menor distância para, em seguida, verificar

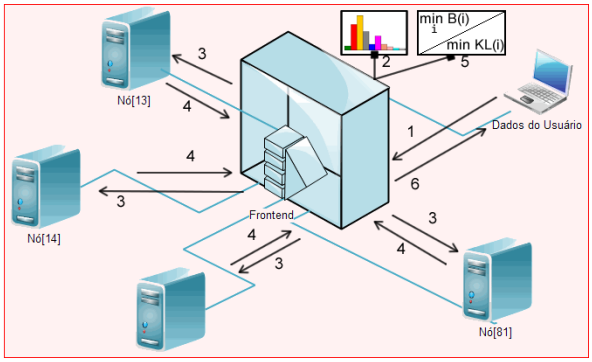

Figura 3. Distribuição das Tarefas ao Longo dos Nós.

qual tipo de dado fornece a predição mais provável (5). Finalmente, (6) essa predição é enviada de volta ao usuário da aplicação.

V. Experimentos e Resultados

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*, *precisão*, *recall* e *F-Measure*. O *speedup* foi utilizado para medir o ganho de eficiência da computação distribuída do algoritmo.

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. 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 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.

VI. Conclusão

Neste trabalho foi proposto um método para estimar informações a partir de dados de redes sociais. 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, trabalhos, etc) melhor descreve o perfil do usuário.

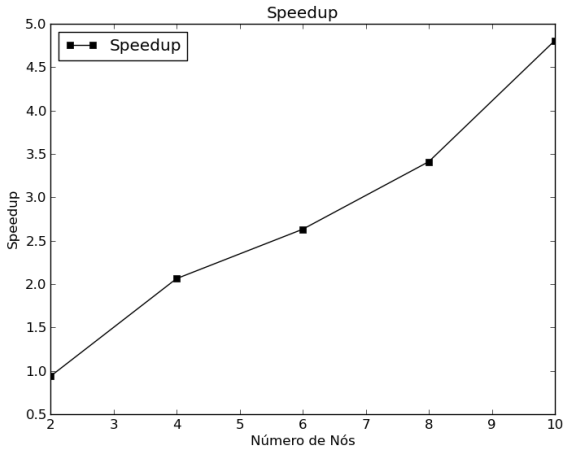

Figura 4. Speedup.

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 descrita uma estratégia de distribuição e os resultados de simulação correspondente.

Estudos visando aprimorar critérios de seleção das predições são necessários. 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.

Referências

[1] C. H. Lau, Y. Li, and D. Tjondronegoro, "Microblog retrieval using topical features and query expansion," in TREC'11, 2011.

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

[3] 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.

[4] R. Dey, C. Tang, K. W. Ross, and N. Saxena, "Estimating age privacy leakage in online social networks," in INFOCOM, 2012, pp. 2836–2840.

[5] Euisun Choi and Chulhee Lee. Feature extraction based on the bhattacharyya distance. Pattern Recognition, 36(8):1703–1709, August 2003.

[6] S. Kullback, and R. A. Leibler. Ann. Math. Statist. 22(1):79-86 (1951).