Adaptive Task Replication Strategy for Human Computation

Lesandro Ponciano, Francisco Brasileiro, Guilherme Gadelha, Adabriand Furtado
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande (UFCG)
Campina Grande – PB – Brazil
lesandrop@lsd.ufcg.edu.br, fubica@dsc.ufcg.edu.br, {guilherme,adabriand}@lsd.ufcg.edu.br

***Abstract*—Human computation systems are distributed systems in which the processors are human beings, called workers. In such systems, task replication has been used as a way to obtain results redundancy and quality. The level of replication is usually defined before the tasks start executing. This approach, however, generates the problem of defining the suitable task replication level. If the level of replication is overestimated, it is used an excessive amount of workers and, therefore, there is an increase in the cost of executing all tasks. On the other hand, if the level of replication is underestimated, a desired level of quality cannot be achieved. This work proposes an adaptive replication strategy that defines the level of replication for each task during execution time. The strategy is based on estimations of the degree of difficulty of tasks and the degree of credibility of workers. Results from simulations using data from two real human computation applications show that, compared to non-adaptive task replication, the proposed strategy reduces the number of replicas substantially, without compromising the accuracy of the obtained answers.**

***Keywords*-Human Computation; Task Redundancy; Task Replication; Task Difficulty; Inter-rater agreement**

I. Introdução

Sistemas de computação por humanos[1] são sistemas distribuídos que utilizam a cognição e a capacidade de raciocínio de seres humanos para resolver problemas que os computadores de silício ainda não podem resolver de forma satisfatória, mas que seres humanos são capazes de resolver de forma rápida e precisa [1], [2]. Esses problemas incluem compreensão de linguagem natural, recuperação de informação em imagens e tarefas ligadas à criatividade. Aplicações de computação por humanos podem ser modeladas como uma aplicação distribuída composta por diversas tarefas que podem ser executadas por seres humanos diferentes. O uso de computação por humanos se popularizou com as tarefas reCAPTCHA [1], mas atualmente o universo de aplicações baseadas em computação por humanos é amplo e diverso [2], [3].

Com o propósito de dar suporte à execução desse tipo de aplicação em larga escala, sistemas dedicados à computação por humanos têm sido desenvolvidos. Um sistema de computação por humanos é um sistema computacional distribuído onde os processadores são seres humanos, chamados de *trabalhadores*. Tal sistema orquestra o poder cognitivo de um grupo de trabalhadores conectados à Internet e gerência o poder computacional provido por eles de forma a executar as tarefas das aplicações. Dois tipos de sistemas de computação por humanos bastante difundidos atualmente são os sistemas de *trabalho online* [4] e os sistemas de *pensamento voluntário* [5]. Sistemas de trabalho online agregam trabalhadores que possuem uma motivação financeira. Um dos principais exemplos é a plataforma Amazon Mechanical Turk (Mturk)[2]. Sistemas de pensamento voluntário, por sua vez, agregam trabalhadores que executam tarefas como um trabalho voluntário. Um dos principais exemplos é o Zooniverse[3]. Sistemas como o Mturk e Zooniverse agregam trabalhadores não especialistas. Dessa forma, tem-se uma preocupação constante com a qualidade das respostas obtidas desses trabalhadores.

A forma como se avalia a qualidade das respostas geradas pelos trabalhadores em sistemas de computação por humanos depende do tipo de tarefa. Em tarefas que envolvem subjetividade e/ou criatividade, a qualidade das respostas não é avaliada em termos de corretude, i.e., não se define que uma resposta está correta ou incorreta. Entretanto, em tarefas ditas factuais as respostas podem ser avaliadas em termos de corretude. Por exemplo, define-se como factual uma tarefa que exibe uma fotografia de uma paisagem e pergunta ao trabalhador se existe ou não uma árvore na paisagem retratada na fotografia. Em tarefas desse tipo, busca-se a resposta correta. No entanto, diversos fatores não intencionais podem levar os trabalhadores a proverem respostas incorretas. Para se identificar e eliminar respostas incorretas, geralmente se utiliza redundância de respostas, que são providas por diversos trabalhadores diferentes. A ideia é que, sobre as respostas redundantes, pode-se aplicar uma estratégia de tratamento de incertezas e se obter a resposta correta. A estratégia mais simples e mais utilizada na prática é o voto majoritário [6], [2], em que se considera correta a resposta provida pela maioria dos trabalhadores.

Para se obter respostas redundantes em sistemas de computação por humanos, geralmente se utiliza *replicação de tarefas* realizada de forma ativa com nível de replicação definido de forma não adaptativa. Replicação ativa de tarefas, neste contexto, significa que para cada tarefa são geradas diversas réplicas e que cada réplica é executada integralmente de forma independente por um trabalhador diferente. O nível de replicação definido de forma não

[1]Do inglês *Human Computation*. Também tem sido traduzido como "computação humana".

[2]mturk.com, visitado pela última vez em 25/11/2013.

[3]zooniverse.org, visitado pela última vez em 25/11/2013.

978-1-4799-5612-8/14 $31.00 © 2014 IEEE
DOI 10.1109/SBRC.2014.30

adaptativa indica que a quantidade de trabalhadores que executarão diferentes réplicas das tarefas é definida antes delas começarem a serem executadas [7]. Essa abordagem de replicação, entretanto, gera o problema de se definir o nível de replicação adequado para as tarefas. Se o nível de replicação for subestimado, pode-se não obter a resposta correta. Por outro lado, se o nível de replicação for superestimado, aumenta-se desnecessariamente o custo de se executar todas as tarefas. Em mercados de trabalho *online*, esse custo significa maior gasto monetário em termos do valor pago aos trabalhadores para executar as diversas réplicas da tarefa. Em sistemas de pensamento voluntário, esse custo se manifesta pelo custo de recrutar e engajar voluntários para que executem as réplicas, ou de outra forma, em uma menor vazão do sistema.

Nesse contexto, o problema tratado neste trabalho é a *definição do nível de replicação adequado para tarefas de computação por humanos*. Definir o nível de replicação envolve considerar fatores adversos que podem levar os trabalhadores a gerarem respostas incorretas e, que em muitos casos, são difíceis de serem previstos antes da tarefa ser submetida para execução. Por exemplo, a acurácia dos trabalhadores tende a variar com o grau de dificuldade da tarefa. A percepção de dificuldade, por sua vez, pode variar de um trabalhador para outro. Além disso, existem tarefas para as quais aumentar o grau de replicação não aumenta necessariamente a probabilidade de se obter a resposta correta no conjunto de respostas redundantes [8]. Esse é o caso, por exemplo, de tarefas que possuem algum complicador que leva trabalhadores não especialistas a gerarem respostas divergentes, tornando difícil identificar qual é a resposta representativa de escolha da maioria com alta confiança.

Este trabalho apresenta uma estratégia de *replicação de tarefas adaptativa* que visa otimizar o número de réplicas necessárias para se obter uma resposta que represente a escolha da maioria dos trabalhadores em sistemas de computação por humanos. A principal característica da estratégia proposta é usar o histórico de tarefas executadas pelos trabalhadores para estimar, em tarefas futuras, a probabilidade das respostas providas por alguns trabalhadores representarem uma resposta de escolha da maioria dos trabalhadores no sistema. Ao realizar essa estimativa, a estratégia leva em conta a ocorrência de respostas incorretas geradas de forma não intencional e sua relação com o grau de dificuldade das tarefas. A estratégia proposta é avaliada em simulações que usam dados de duas aplicações típicas de computação por humanos: análise de sentimentos e julgamento de fatos. Os resultados obtidos sugerem que a estratégia proposta é capaz de identificar a resposta da escolha da maioria com bem menos réplicas que uma replicação não adaptativa, sem prejuízo da acurácia obtida. Outro benefício da estratégia é destacar tarefas cujas respostas obtidas não são conclusivas. Isso permite, por exemplo, que o dono da aplicação conheça quais são as tarefas nas quais trabalhadores tendem a divergir. São tarefas que, por exemplo, podem requerer a execução por trabalhadores mais especializados.

O restante deste trabalho é dividido da seguinte forma. A Seção II apresenta o referencial teórico e os trabalhos relacionados. A Seção III apresenta a estratégia de replicação adaptativa. Em seguida, a Seção IV apresenta os materiais e métodos utilizados para avaliar a estratégia proposta. Essa seção é seguida pela apresentação e discussão dos resultados. Finalmente, a Seção VI apresenta as conclusões e sugestões para trabalhos futuros.

II. Referencial Teórico e Trabalhos Relacionados

Esta seção apresenta duas classes de trabalhos relevantes a este estudo: os que tratam da acurácia dos trabalhadores e os que tratam de replicação de tarefas.

A. Acurácia dos Trabalhadores

A acurácia de respostas obtidas em sistemas de computação por humanos está diretamente relacionada à acurácia dos trabalhadores, ou seja, à probabilidade do trabalhador prover uma resposta correta. Naturalmente, trabalhadores podem prover respostas incorretas de forma intencional. Por exemplo, planejarem um conluio de modo a todos proverem uma mesma resposta errada para a tarefa. Entretanto, esse tipo de comportamento não tem sido reportado como algo comum nesses sistemas [2], [9]. Geralmente, quando ele ocorre, se dá como uma reação dos trabalhadores a um comportamento inadequado do dono da aplicação, como submeter tarefas mal projetadas ou alta recusa dos trabalhos realizados pelos trabalhadores [9]. Este trabalho foca em *fatores não intencionais*.

A Teoria do Erro Humano [10] mostra que existem três causas de seres humanos proverem respostas incorretas de modo não intencional: ignorância (*mistake*), esquecimento (*lapse*) e deslize (*slip*). Ignorância ocorre em situações em que o ser humano não possui todo conhecimento necessário para executar a tarefa, assim ele não define corretamente os passos que precisam ser seguidos para se chegar à resposta correta. Em deslizes e esquecimentos o ser humano possui os conhecimentos necessários para executar a tarefa e ele define corretamente os passos que precisam ser seguidos para se chegar à resposta correta. Entretanto, no caso do deslize ele executa incorretamente algum dos passos e no esquecimento ele simplesmente se esquece de executar algum passo.

Para prevenir respostas incorretas causadas pelo fator ignorância, alguns estudos focam em tentar estimar a priori se o trabalhador possui os conhecimentos necessários para executar a tarefa. Isso é geralmente feito usando *qualificações* [11]. Qualificações são tarefas cujas respostas são conhecidas e são usadas para se avaliar a acurácia dos trabalhadores. Trabalhadores só são autorizados a executarem outras tarefas se obtiverem alta acurácia nas tarefas de qualificação. Essa estratégia requer que os itens das tarefas sejam muito parecidos com os itens da qualificação. Além disso, ela não previne esquecimentos e deslizes. Geralmente, a ocorrência de esquecimentos e deslizes é relacionada à atenção do trabalhador durante a execução da tarefa e à dificuldade que ele percebe na tarefa que está sendo executada [10].

Existem diversos fatores que podem tornar uma tarefa de computação por humanos difícil de ser executada. Eles podem ser fatores inerentes aos dados a que a tarefa se refere (e.g., quando é requerida muita atenção para encontrar um determinado item em uma imagem) ou ao projeto da tarefa, como a carga cognitiva (e.g., quando é preciso seguir uma longa sequência de passos ou comparar muitos itens para se chegar à resposta correta).

Neste trabalho, trata-se de cenários em que as tarefas são geradas a partir de uma grande quantidade de itens desconhecidos e que apresentam grande variação entre eles de modo que nem é possível preprocessá-los para estimar os graus de dificuldade e nem se conhece a priori as respostas corretas. Por exemplo, muitas das tarefas de computação por humanos consistem basicamente de um item a ser julgado, de uma pergunta referente a esse item e de um conjunto de opções de resposta. As tarefas diferem umas das outras em termos apenas do item a ser julgado. Considere uma tarefa projetada para detectar ironia em mensagens de texto. Ela pode consistir da pergunta "A frase abaixo apresenta uma ironia?", e as opções de resposta podem ser "Sim" ou "Não". Grande quantidade dessas tarefas pode ser gerada coletando-se automaticamente frases em comentários feitos em notícias em páginas Web. Neste caso, não se tem qualquer controle do conteúdo de cada frase que será avaliada em cada tarefa. Algumas frases podem ser fáceis de serem avaliadas e outras serem muito difíceis.

B. Replicação de Tarefas

Replicação é um mecanismo bastante utilizado para se obter redundância em sistemas computacionais. Em sistemas distribuídos compostos por máquinas, muitas das primeiras abordagens que empregaram esse mecanismo são baseadas em máquina de estado (*machine state*) [12], [13]. Nessas abordagens, replicação é dita ativa quando cada réplica de uma tarefa é executada integralmente por máquinas diferentes, cada uma partindo de um mesmo estado inicial até o mesmo estado final da tarefa. De outro modo, a replicação é dita passiva quando é mantido um *backup* do estado da tarefa durante sua execução por uma máquina. Esse *backup* mantém as computações já realizadas e ele pode ser utilizado para que a tarefa não precise ser re-executada integralmente caso uma falha ocorra durante sua execução.

Nos últimos anos, replicação tem sido utilizada com diversos propósitos e em diversos tipos de sistemas. Por exemplo, em grades computacionais oportunistas como OurGrid, replicação é utilizada com o propósito de reduzir o tempo de resposta das aplicações [14]. Já em sistemas de computação voluntária como BOINC[4], replicação é utilizada como forma de identificar sabotadores e isolar as respostas providas por eles [15]. Em sistemas compostos por seres humanos, entretanto, replicação de tarefas é realizada para se obter respostas redundantes de modo que se possa identificar uma resposta que represente uma escolha social.

Em Escolha Social (*"Social Choice Theory"*) [16], parte-se de respostas providas por diversos indivíduos e, a partir da agregação dessas respostas, obtém-se uma resposta de preferência coletiva. A base dos algoritmos atuais de tratamento de incertezas em respostas obtidas em sistemas de computação por humanos é identificar uma resposta de preferência social, quando eliminadas respostas erradas [6]. Em tarefas em que há o conceito de corretude, essa resposta pode não ser necessariamente a resposta que seria julgada correta por um especialista. Geralmente quando isso ocorre o problema é atribuído ao nível de replicação utilizado, a algum fator dificultador no item que está sendo avaliado na tarefa ou à forma como a tarefa foi projetada [17].

O propósito deste trabalho é identificar uma resposta de escolha da maioria dos trabalhadores de modo que um número menor de trabalhadores tenham que ser consultados, i.e., menos réplicas sejam geradas. A estratégia de replicação proposta combina conceitos de escolha social com replicação de tarefas em sistemas com máquinas. A principal inspiração em escolha social é a *concordância* de um trabalhador com outros trabalhadores. Neste trabalho, utiliza-se informação da concordância entre trabalhadores em tarefas executadas no passado para estimar concordâncias em tarefas futuras, levando-se em conta também estimativas do grau de dificuldade dessas tarefas. De replicação em sistemas compostos por máquinas, utiliza-se um arcabouço de medidas de credibilidade proposto para sistemas de computação voluntária [15].

A ideia de otimizar a redundância de respostas em sistemas de computação por humanos por meio de replicação de tarefas que leva em conta medidas de credibilidade foi publicada recentemente como um resumo de trabalho em andamento [7]. O presente artigo apresenta o estudo desenvolvido. Além de apresentar todos os fundamentos da estratégia, incorpora-se um novo método para estimar a acurácia dos trabalhadores, considerando o histórico de concordância com outros trabalhadores e estimativas de dificuldade das tarefas. Além disso, estende-se a avaliação da estratégia para medir a acurácia, economia de réplicas e proporção de tarefas cujas respostas obtidas dos trabalhadores não são conclusivas.

III. Estratégia de Replicação

Nesta seção, antes de apresentar nossa estratégia de replicação, são definidas a medida de dificuldade e as medidas de credibilidade que ela utiliza.

A. Medindo o Grau de Dificuldade das Tarefas

O grau de dificuldade de uma tarefa é medido pelo grau de divergência dos trabalhadores entre as opções de resposta disponíveis. A ideia principal é que quanto mais os trabalhadores se dividirem em diferentes opções de resposta, maior a chance da tarefa apresentar *(i)* alguma característica atípica, levando os trabalhadores a gerarem uma resposta incorreta por ignorância ou *(ii)* algum fator

[4]Do inglês *Berkeley Open Infrastructure for Network Computing* (http://boinc.berkeley.edu/).

dificultador, gerando esquecimentos e deslizes. Dado N o multiconjunto de todas as respostas recebidas dos trabalhadores para uma mesma tarefa e f a frequência da resposta mais frequente em N, a Equação 1 apresenta o cálculo do grau de dificuldade d.

$$d = \frac{|N| - f}{|N|} \tag{1}$$

Esse grau de dificuldade é um valor real $d \in [0, 1]$. Ele assume o valor 0 quando todos os trabalhadores apresentam a mesma resposta para a tarefa, i.e., $f = |N|$. Por outro lado, ele tende ao valor 1 na proporção em que o número de opções de respostas aumenta e que os trabalhadores se dividem entre a opção mais frequente e as demais opções. Por simplicidade, nas análises realizadas neste trabalho, arredonda-se o grau de dificuldade da tarefa para 1 casa decimal, de modo a trabalhar com apenas 11 graus de dificuldade, de $0{,}0$ a $1{,}0$.

B. Medindo Concordância e Credibilidade

O grau de concordância de um trabalhador é a probabilidade dele, ao executar uma nova tarefa, prover uma resposta igual à resposta que a maioria dos trabalhadores proviria considerando seu histórico de tarefas. Um trabalhador pode concordar mais em tarefas com determinado grau de dificuldade e discordar mais em outros. Dessa forma, o grau de concordância do trabalhador varia com o grau de dificuldade da tarefa. Assim, define-se como $c_{t,d}$ o grau de concordância do trabalhador t em tarefas que possuem o grau de dificuldade d, $c_{t,d} \in [0, 1]$. Antes de executar qualquer tarefa, o grau de concordância do trabalhador t é $c_{t,d} = 0{,}5$ para todos os valores de d. Ou seja, ele possui 50% de chance de prover uma resposta igual à da maioria.

Sempre que a execução de uma tarefa é concluída atingindo um valor aceitável de credibilidade, calcula-se o seu grau de dificuldade pela Equação 1 e verifica-se os trabalhadores que proveram respostas iguais à resposta da maioria dos trabalhadores. Com isso, pode-se recalcular a proporção de convergência de cada trabalhador considerando todas as tarefas que ele executou no passado. Esse cálculo é dado por $p_{t,d} = m_{t,d}/n_{t,d}$, sendo $m_{t,d}$ o total de tarefas com grau de dificuldade d que a resposta provida pelo trabalhador t foi igual à resposta provida pela maioria e $n_{t,d}$ o total de tarefas com dificuldade d que o trabalhador t executou. O novo grau de concordância do trabalhador é computado pela Equação 2. Essa equação é uma média harmônica ponderada entre $p_{t,d}$ e $0{,}5$. Essa média visa impedir valores de concordância muito altos ou muito baixos quando poucas tarefas tiverem sido executadas. Por ser ponderada, essa média faz com que o valor inicial $0{,}5$ tenha menor peso no grau de concordância do trabalhador, quanto maior for o número de tarefas que ele tiver executado ($n_{t,d}$).

$$c_{t,d} = \frac{n_{t,d} + 1}{\frac{n_{t,d}}{p_{t,d}} + \frac{1}{0{,}5}} \tag{2}$$

A resposta gerada por um trabalhador t tem probabilidade $c_{t,d}$ de ser a resposta de escolha da maioria dos trabalhadores. Respostas iguais geradas por diferentes trabalhadores para diferentes réplicas de uma mesma tarefa são agregadas em grupos. O número de grupos é definido como g e ele varia com a diversidade de respostas geradas, sendo o grupo G_a o multiconjunto das probabilidades das respostas a recebidas de diferentes trabalhadores. Pode-se computar a credibilidade de cada grupo de respostas. Essa credibilidade é a probabilidade condicional da resposta do grupo ser a resposta de escolha da maioria dos trabalhadores no sistema e das outras possíveis respostas para a mesma tarefa não serem a escolha desses trabalhadores. O cálculo dessa probabilidade condicional é formalizado na Equação 3. Essa equação apresenta o cálculo da probabilidade condicional da resposta a do grupo de respostas G_a representar a escolha da maioria dos trabalhadores e das demais respostas dos outros grupos não representarem. Esse cálculo é inspirado na credibilidade de grupos de respostas em sistemas de computação voluntária [15].

$$C(G_a) = \frac{P(G_a \text{ good}) \prod_{i \neq a} P(G_i \text{ bad})}{\prod_{j=1}^{g} P(G_j \text{ bad}) + \sum_{j=1}^{g} P(G_j \text{ good}) \prod_{k \neq j} P(G_k \text{ bad})} \tag{3}$$

O grupo G_a com maior valor de $C(G_a)$ é o grupo que representa a resposta candidata da tarefa. Por sua vez, a credibilidade requerida r, definida pelo dono da aplicação, indica um limiar mínimo para $C(G_a)$ de modo que essa resposta candidata possa ser aceita como a escolha da maioria. Dessa forma, uma resposta a é dita ser de escolha da maioria dos trabalhadores se, e somente se, a condição na Equação 4 for satisfeita. Essa condição indica que $C(G_a)$ precisa ser maior ou igual a r e que $C(G_a)$ também precisa ser maior que a credibilidade dos demais grupos. Assim, se a é dita ser a resposta de escolha da maioria, nenhuma outra resposta para a mesma tarefa pode atingir essa condição.

$$C(G_a) \geq r \wedge C(G_a) > C(G_i), 1 \leq i \leq g, i \neq a \tag{4}$$

C. Estratégia de Replicação

A principal ideia dessa estratégia é que novas réplicas de uma tarefa precisam ser geradas apenas se as respostas já obtidas não atingirem um nível requerido de credibilidade (r). Isso é realizado utilizando os conceitos de credibilidade apresentados na seção anterior.

Para toda tarefa, sempre gera-se uma réplica inicial. Após a execução de cada réplica de uma tarefa, as respostas iguais obtidas de diferentes trabalhadores são agrupadas e o grau de credibilidade de cada grupo de respostas é recalculado pela Equação 3. Novas réplicas da tarefa são geradas até que pelo menos uma das condições seguintes seja satisfeita:

- Critério de credibilidade é atingido. Isso ocorre quando a condição descrita na Equação 4 é satisfeita, o que indica que uma resposta final para a tarefa foi

obtida.
- Limite máximo de réplicas é atingido. Isso ocorre quando um limite máximo de réplicas definido pelo dono da aplicação é atingido.

Quando a replicação termina com o critério de credibilidade satisfeito, calcula-se o grau de dificuldade da tarefa (d) usando a Equação 1 e identifica-se os trabalhadores que proveram respostas que coincidiram com a maioria e os que proveram respostas que não coincidiram com a maioria. Esses dados são utilizados para atualizar os graus de concordância dos trabalhadores que participaram da execução. Como já discutido anteriormente, esse cálculo é realizado pela Equação 2.

Quando a replicação termina porque o número máximo de réplicas foi atingido, mas sem satisfazer o critério de credibilidade, duas situações são possíveis, dependendo de como a aplicação que usa replicação adaptativa é configurada. Em uma *configuração conservadora*, as tarefas que não atingem o limiar de credibilidade requerida são marcadas como "sem conclusão" e não têm uma resposta associada. Por outro lado, em uma *configuração não conservadora*, a resposta para a tarefa é aquela que obteve o maior valor de credibilidade, mesmo que abaixo do limiar requerido.

Considerando a literatura de replicação de tarefas discutida na Seção II-B, essa estratégia de replicação pode ser definida como ativa e adaptativa. A estratégia de replicação é ativa porque cada réplica de uma tarefa é executada integralmente por cada trabalhador, não há *backup* do estado da tarefa durante sua execução. A estratégia é dita adaptativa no sentido de que o número de réplicas geradas não é pré-definido. Ele é calculado em tempo de execução e é dependente das respostas recebidas dos trabalhadores.

IV. Materiais e Métodos

Esta seção apresenta as bases de dados, métricas e os cenários avaliados.

A. Bases de Dados

A estratégia proposta neste trabalho foi avaliada por meio de simulações que utilizam dados de duas aplicações de computação por humanos: Julgamento de Fatos[5] e Análise de Sentimentos[6]. A base de dados Julgamento de Fatos foi disponibilizada pelo Google. Ela consiste em um conjunto de julgamentos realizados por seres humanos sobre relações referentes a pessoas públicas na Wikipedia. As relações são do tipo "a pessoa X se graduou na universidade Y". Os trabalhadores eram solicitados a julgar se a relação é "verdadeira", "falsa" ou "não responder". A base de dados Análise de Sentimentos, por sua vez, foi disponibilizada pelo CrowdFlower (crowdflower.com). Ela consiste em um conjunto de julgamento de seres humanos sobre a condição climática relatada em *tweets*[7]. Cada tarefa apresenta aos trabalhadores um tweet e eles respondem se a informação sobre o clima constante no tweet é "negativa", "neutra (o autor do tweet está apenas compartilhando informação)", "positiva", "O tweet não é relacionado à condição do clima", ou "não sei responder". Em ambas as bases, 95% das tarefas foram executas por 5 trabalhadores e as restantes foram executadas por mais trabalhadores.

A Tabela I apresenta um sumário estatístico das bases. Duas importantes distinções entre essas bases no contexto deste trabalho são: *(i)* número de opções de resposta por tarefa e *(ii)* proporção de tarefas por trabalhador. A aplicação Análise de Sentimentos apresenta 5 opções de resposta e aplicação Julgamento de Fatos 3. Quanto mais opções de resposta, maior a chance de divergência entre os trabalhadores em tarefas difíceis. Quanto à proporção de tarefa por trabalhadores, ela é de 748 na aplicação Julgamento de Fatos e 51 na aplicação Análise de Sentimentos. Para o algoritmo proposto, quando mais tarefas o trabalhador executa, mais informação histórica se mantém e mais acurada pode se tornar a estimativa da concordância dele com outros trabalhadores. Essas bases também contêm um conjunto de tarefas *Ground truth*, i.e., tarefas para as quais as respostas corretas são conhecidas. Essas tarefas foram geradas de modo a serem representativas de toda a base de dados. Elas consistem em 576 tarefas na base de dados Julgamento de Fatos e em 300 tarefas na base de dados Análise de Sentimentos. Neste estudo, essas tarefas são utilizadas para avaliar a acurácia das respostas obtidas pelas estratégias de replicação.

A estratégia de replicação proposta foi simulada usando como entrada os dados das tarefas e as respostas providas pelos trabalhadores para cada réplica. A ordem em que as tarefas são executadas e as respostas que os trabalhadores retornam são conforme registados nas bases de dados. No entanto, a ordem em que os trabalhadores geram as respostas para as réplicas das tarefas não é conhecida. Como o algoritmo proposto pode terminar a replicação sem que as respostas de todas as réplicas sejam utilizadas, a ordem em que as respostas armazenadas nas bases de dados são utilizadas tem impacto nos resultados. Esse impacto foi medido por 5 simulações usando as respostas ordenadas de forma aleatória. Na análise dos resultados, sempre se apresenta a média dos resultados obtidos nessas simulações com barras de erros para um nível de confiança de 95%. Naturalmente, o número de réplicas que o algoritmo proposto pode gerar é limitado ao existente na base de dados. Um dos principais objetivos da avaliação é verificar em que situações a estratégia de replicação proposta é capaz de gerar menos réplicas.

B. Métricas

Três métricas são utilizadas na avaliação realizada.

- **Economia de réplicas:** Dado que na base de dados são geradas x réplicas e a estratégia proposta gerou y réplicas, a economia de réplicas é dada por $\frac{x-y}{x}$. Essa economia é calculada em duas situações: por tarefa da aplicação e em toda a aplicação. Quando é calculada por tarefa, x depende da tarefa, sendo que 95% das

[5]Disponível em https://code.google.com/p/relation-extraction-corpus/

[6]Disponível em https://minbox.com/gi/Uh0vixtBNw

[7]Tweets são mensagens de texto de até 140 caracteres compartilhadas na rede social twitter.com.

Table I
SUMÁRIO ESTATÍSTICO DAS BASES DE DADOS

Característica	Julgamento de Fatos	Análise de Sentimentos
#Trabalhadores únicos	57	1.960
#Tarefas diferentes	42.624	98.980
#Opções de resposta por tarefa	3	5
#Réplicas	220.000	500.000
#Tarefas *ground truth*	576	300

tarefas têm 5 réplicas e o restante tem mais que 5. Quando a economia é calculada em toda a aplicação, x é o total de réplicas na aplicação apresentada na Tabela I.

- **Acurácia:** É a taxa de acerto nas tarefas *ground truth*. Ela é a razão entre o total de tarefas cujas respostas obtidas coincidem com as existentes na base *ground truth* e o total de tarefas *ground truth*.
- **Tarefas sem conclusão:** É a proporção de tarefas que o algoritmo proposto não atingiu o limiar de credibilidade definido pelo dono da aplicação. É calculado como a razão entre o número de tarefas que não atingiram o limiar de credibilidade e o número total de tarefas.

C. Cenários

Os cenários de avaliação foram projetados de modo a permitir uma análise custo/benefício da estratégia proposta. São dois os cenários: um que assume uma configuração conservadora da aplicação, e outra que não. No cenário com configuração não conservadora, a estratégia sempre gera uma resposta para as tarefas, que é a resposta do grupo de respostas de maior credibilidade. Utiliza-se o valor de $0{,}95$ como condição de parada com conclusão. Neste cenário se analisa: acurácia por aplicação; economia de réplicas por tarefa e por aplicação e distribuição da credibilidade das respostas obtidas. No segundo cenário, a configuração é conservadora e uma resposta só é gerada se o grupo de respostas com maior credibilidade for maior ou igual à credibilidade requerida, que é variada entre $0{,}90$ a $0{,}99$. Neste cenário se analisa: acurácia das respostas obtidas com conclusão, proporção de tarefas sem conclusão e economia de réplicas.

V. APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

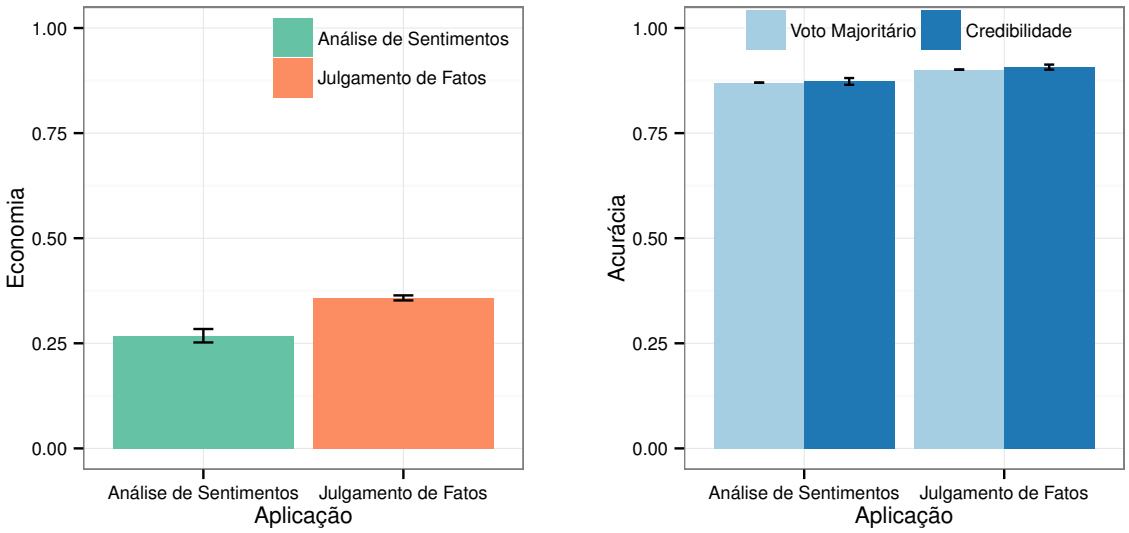
Esta seção apresenta os resultados obtidos nos cenários de avaliação.

A. Configuração não Conservadora da Aplicação

A Figura 1 apresenta o total de economia obtida e da acurácia das respostas em todas as tarefas das aplicações avaliadas. A Figura 1(a) mostra a economia de réplicas que a estratégia proposta gera em relação à replicação não adaptativa. Em média, a economia obtida é de 27% na aplicação Análise de Sentimentos e de 36% na aplicação Julgamento de Fatos. A Figura 1(b), por sua vez, mostra a acurácia que se obtém quando a resposta final para cada tarefa é escolhida pelo *voto majoritário* sobre todas as respostas obtidas na replicação não adaptativa e a acurácia que se obtém quando a resposta final para cada tarefa é escolhida pelo seu grau de *credibilidade* calculado pelo algoritmo adaptativo proposto neste trabalho. Essa figura mostra que as duas estratégias atingem acurácia similar, sendo em média 87% na aplicação Análise de Sentimentos e 90% na aplicação Julgamento de Fatos. Pelos resultados apresentados na Figura 1, pode-se concluir que, para as aplicações avaliadas, o algoritmo proposto permite reduzir substancialmente o número de réplicas das tarefas sem afetar negativamente a acurácia das respostas.

Além do resultado agregado apresentado na Figura 1, também é importante verificar como o algoritmo proposto se comporta em cada tarefa. A Figura 2 tem como objetivo permitir essa análise. Essa figura apresenta as funções de distribuição acumulada da economia de réplicas obtida em cada tarefa (Fig. 2(a)) e do nível de credibilidade obtido na resposta final de cada tarefa (Fig. 2(b)). Na Figura 2(a) a economia de réplicas obtida é indicada no eixo horizontal e no eixo vertical é indicada a proporção de tarefas que atingiram economia menor ou igual à correspondente no eixo horizontal. O máximo de economia obtida em uma mesma tarefa é $0{,}66$, na aplicação Julgamento de Fatos, e $0{,}98$ na aplicação Análise de Sentimentos. Nas duas aplicações existem tarefas para as quais a economia de réplicas foi nula. Porém, a estratégia proposta economiza réplicas em 81% das tarefas na aplicação Julgamento de Fatos e em 56% das tarefas na aplicação Análise de Sentimentos.

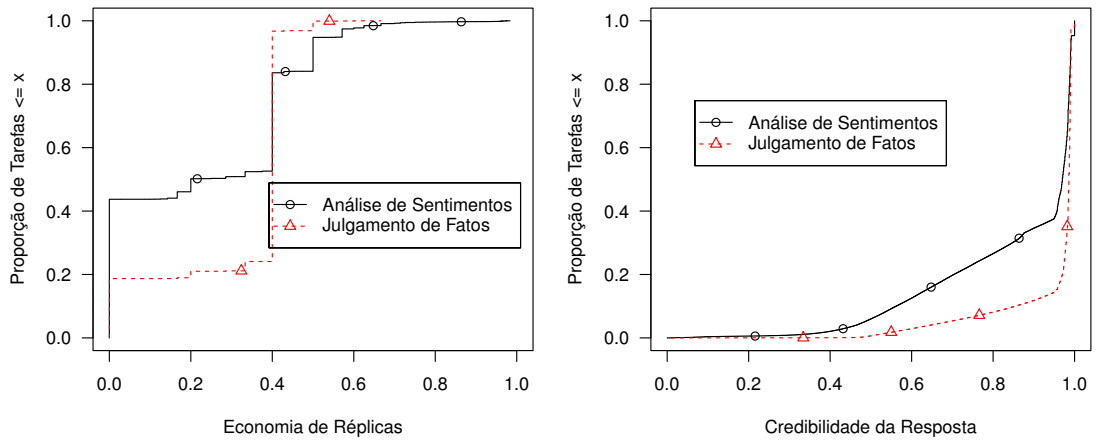
Na Figura 2(b), pode-se ver a distribuição da credibilidade da resposta final das tarefas obtida pelo algoritmo proposto. As tarefas cujas respostas não atingiram credibilidade de $0{,}95$ são uma proporção $0{,}10$ na aplicação Julgamento de Fatos e $0{,}45$ na aplicação Análise de Sentimentos. Ambos os resultados mostrados na Figura 2 indicam que a estratégia de replicação proposta apresenta pior desempenho em termos de economia de réplicas e da credibilidade final das respostas na aplicação Análise de Sentimentos. Provavelmente isso ocorre porque essa aplicação tem maior número de opções de resposta por tarefa e requer uma avaliação mais subjetiva do que a outra aplicação. Essas duas características contribuem para aumentar as chances dos trabalhadores divergirem e mais réplicas serem necessárias para tentar atingir maior credibilidade na resposta.



(a) Economia de réplicas

(b) Acurácia

Figure 1. Economia de réplicas e acurácia das respostas por aplicação.



(a) Economia de réplicas em cada tarefa

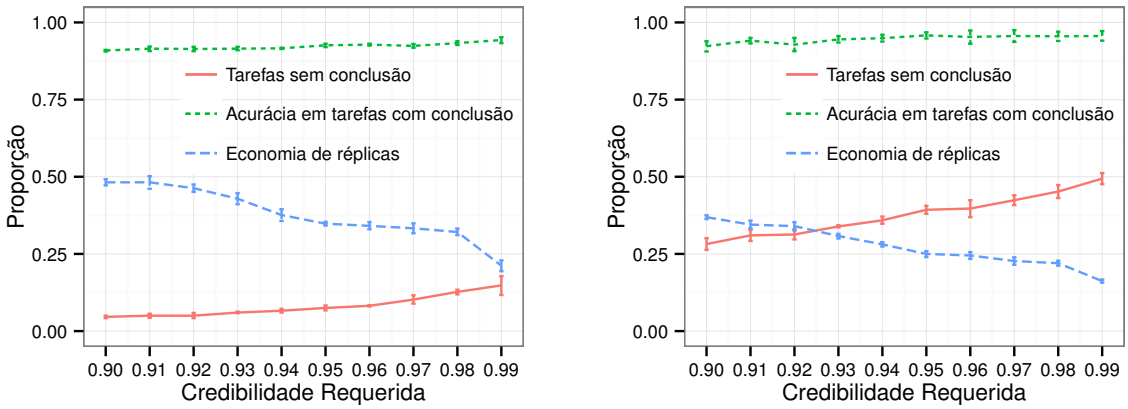
(b) Credibilidade da resposta final de cada tarefa

Figure 2. Funções de Distribuição Acumulada (FDAs) da (a) economia de réplicas obtida em cada tarefa e da (b) credibilidade da resposta final de cada tarefa.

B. Configuração Conservadora da Aplicação

Neste cenário, apenas respostas que atingem o limiar de credibilidade requerido são consideradas na análise da acurácia. Tarefas em que a resposta final não atingiu o limiar são marcadas como "sem conclusão", porque são tarefas que podem requer uma avaliação mais especializada. Na Figura 3 são apresentados os resultados obtidos, sendo que na Figura 3(a) estão os resultado na aplicação Julgamento de Fatos e na Figura 3(b) estão os resultados na aplicação Análise de Sentimentos. Em ambos os gráficos, o eixo horizontal é a credibilidade requerida pelo dono da aplicação. O eixo vertical, por sua vez, é o valor obtido para cada uma das métricas que são apresentadas como linhas nos gráficos.

As Figuras 3(a) e 3(b) mostram que na medida em que se aumenta a credibilidade requerida na resposta, ocorre uma redução na proporção de economia em relação à replicação não adaptativa e um aumento na proporção de tarefas sem conclusão. Na Figura 3(b) é possível ver o ponto em que o número de tarefas que não atingiram o limiar de credibilidade ultrapassa a economia obtida, isso ocorre entre a credibilidade requerida de $0{,}92$ e $0{,}93$. A proporção de acurácia, por sua vez, não sofre grandes variações quando a credibilidade requerida muda. Quando a credibilidade requerida varia de $0{,}90$ a $0{,}99$, a acurácia das respostas varia de $0{,}93$ para $0{,}97$ na aplicação Análise de Sentimentos e de $0{,}91$ para $0{,}93$ na aplicação de



(a) Julgamento de Fatos

(b) Análise de Sentimentos

Figure 3. Impacto da variação da credibilidade requerida na proporção de: i) economia de réplicas, ii) tarefas com conclusão iii) tarefas sem conclusão.

Julgamento de Fatos. Esses resultados indicam que, para os níveis de credibilidade requerida avaliados, quanto mais rigoroso o dono da aplicação é em termos da credibilidade requerida, menor é a economia de réplicas, maior é o número de tarefas sem conclusão e ocorre baixa variação na acurácia das tarefas com conclusão.

Além de obter economia de réplicas e manter a acurácia das respostas, outra vantagem da estratégia proposta é destacar como "sem conclusão" as tarefas em que nenhum grupo de resposta atingiu o limiar de credibilidade requerido. São tarefas cujas respostas não são conclusivas dado que os trabalhadores apresentam divergência entre as opções de resposta. Em computação por humanos, esse tipo de tarefa pode ter grande importância. Uma tarefa em que os trabalhadores divergem sobre a resposta correta pode revelar algo contraditório ou fora do padrão que o dono da aplicação pode ter interesse em investigar com maior cuidado. Dessa forma, a estratégia proposta também contribui para facilitar a análise dos dados obtidos nesses sistemas e também dá suporte a uma nova decisão como, por exemplo, usar trabalhadores mais especialistas para executar as tarefas sem conclusão.

C. Limitações e Implicações

A estratégia de replicação proposta neste trabalho possui algumas limitações que precisam ser destacadas. Ela foi projetada para ambientes em que os trabalhadores sempre executam os mesmos tipos de tarefas mudando apenas o item que está sendo avaliado. Esse é o caso, por exemplo, de sistemas de pensamento voluntário como Galaxy Zoo em que os voluntários sempre executam o mesmo tipo de tarefa, mudando-se apenas a imagem da galáxia a que a tarefa se refere. Esse também é o caso de grupos de tarefas que requerem inteligência humana (HITs, do inglês *Human Intelligence Tasks*) no Mturk, se eles admitirem que cada trabalhador execute mais de um HIT do grupo.

Para ser utilizada em sistemas de pensamento voluntário, a estratégia proposta pode ser implementada no nível do sistema de *middleware*, como o PyBossa[8]. Usuários do Mturk podem implementar a estratégia no nível da aplicação por meio de interfaces como o TurkIt[9]. Usuários de diversos sistemas de computação por humanos também podem usar a estratégia proposta para combinar diferentes sistemas. Por exemplo, primeiro as tarefas podem ser submetidas para serem executadas em sistemas como Mturk, que agregam trabalhadores não especialistas e que geralmente aceitam receber menos que trabalhadores especialistas. As tarefas cujos resultados não atingirem o limiar de credibilidade desejado, podem então ser submetidas para serem executadas em sistemas de computação por humanos que agregam trabalhadores mais especialistas e que, geralmente, cobram mais para executá-las, como Mobile Works (mobileworks.com) e InnoCentive (innocentive.com).

Existem diversas otimizações que podem ser realizadas na estratégia proposta. As duas principais são: fase de aquecimento e escalonamento de réplicas. A fase de aquecimento consiste nas primeiras tarefas a serem submetidas para execução serem preferencialmente tarefas fáceis. Isso pode permitir que a calibragem do grau de concordância dos trabalhadores ocorra de forma mais rápida do que se as primeiras tarefas forem tarefas difíceis. No escalonamento de réplicas, por sua vez, pode-se fazer uma adequação entre a credibilidade requerida e o nível de concordância dos trabalhadores de modo a agilizar a obtenção de uma resposta com a credibilidade desejada. Assim, tarefas em que as respostas estão tendo maior divergência (possivelmente mais difíceis) podem ser escalonadas preferencialmente para trabalhadores que possuem a escolha mais representativa dos trabalhadores nesse tipo de tarefa. Isso pode permitir que trabalhadores mais experientes executem tarefas mais difíceis e trabalhadores iniciantes (que não possuem alto grau de convergência) executem preferencialmente tarefas mais simples. Ao se implementar essas estratégias, o impacto delas no comportamento dos

[8]https://github.com/PyBossa/pybossa, último acesso em 25/11/2013.
[9]https://code.google.com/p/turkit, último acesso em 25/11/2013.

trabalhadores precisa ser avaliado.

VI. Conclusões

Este trabalho apresenta uma estratégia de replicação adaptativa que visa reduzir o número de réplicas necessárias para se obter uma resposta confiável em sistemas de computação por humanos. A estratégia é construída a partir de conceitos da Teoria de Erro Humano, da Teoria de Escolha Social e de replicação de tarefas em sistemas distribuídos compostos por máquinas. Resultados obtidos em simulações usando dados de duas aplicações reais mostram que a estratégia proposta permite uma considerável economia de réplicas, com acurácia similar à obtida quando se realiza voto majoritário sobre respostas redundantes obtidas por replicação não adaptativa. Outro benefício da estratégia proposta é destacar tarefas sem conclusão. Conhecer essas tarefas permite ao dono da aplicação tomar novas ações como usar trabalhadores mais especialistas em outros sistemas para executá-las.

Trabalhos futuros podem investigar formas alternativas de medir o grau de dificuldade das tarefas e a convergência entre os trabalhadores, considerando medidas como entropia de Shannon e Cohen's kappa. Nesse estudo, pode-se identificar se variações na forma de medir o grau de dificuldade de tarefas e a convergência entre os trabalhadores permite melhorar o desempenho da estratégia de replicação proposta neste trabalho. Em futuras extensões da estratégia, além da economia de réplicas, acurácia das respostas e tarefas sem conclusão, outras métricas de desempenho como tempo de resposta das aplicações e vazão do sistema devem ser avaliadas.

References

[1] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "recaptcha: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[2] A. J. Quinn and B. B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *CHI*. ACM, 2011, pp. 1403–1412.

[3] J. Oliveira, L. Ponciano, N. Andrade, and F. Brasileiro, "Estratégias de obtencão de um item máximo em computacão por humanos," in *SBRC*. SBC, 2013, pp. 253–266.

[4] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Human Computation Workshop*. ACM, 2010, pp. 64–67.

[5] L. Ponciano, F. Brasileiro, R. Simpson, and A. Smith, "Volunteers' engagement in human computation astronomy projects," *Computing in Science and Engineering*, vol. 99, p. 1, 2014.

[6] A. Sheshadri and M. Lease, "Square: A benchmark for research on computing crowd consensus," in *Proc. First AAAI Conference on Human Computation and Crowdsourcing*. AAAI, 2013, pp. 156 – 164.

[7] L. Ponciano, F. Brasileiro, and G. Gadelha, "Task redundancy strategy based on volunteers' credibility for volunteer thinking projects," in *Proc. First AAAI Conference on Human Computation and Crowdsourcing*. AAAI, 2013, pp. 60–61.

[8] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? improving data quality and data mining using multiple, noisy labelers," in *SIGKDD*. ACM, 2008, pp. 614–622.

[9] A. Kulkarni, M. Can, and B. Hartmann, "Collaboratively crowdsourcing workflows with turkomatic," in *CSCW*. ACM, 2012, pp. 1003–1012.

[10] J. Reason, *Human error*. Cambridge University Press Cambridge, 1990.

[11] J. J. Chen, N. J. Menezes, A. D. Bradley, and T. North, "Opportunities for crowdsourcing research on amazon mechanical turk," *Interfaces*, vol. 5, no. 3, 2011.

[12] F. B. Schneider, "Implementing fault-tolerant services using the state machine approach: A tutorial," *ACM Comput. Surv.*, vol. 22, no. 4, pp. 299–319, Dec. 1990.

[13] P. Jalote, *Fault tolerance in distributed systems*. Prentice Hall, 1994.

[14] W. Cirne, F. Brasileiro, D. Paranhos, L. Góes, and W. Voorsluys, "On the efficacy, efficiency and emergent behavior of task replication in large distributed systems," *Parallel Computing*, vol. 33, no. 3, pp. 213–234, 2007.

[15] L. F. Sarmenta, "Sabotage-tolerance mechanisms for volunteer computing systems," *Future Generation Computer Systems*, vol. 18, no. 4, pp. 561–572, 2002.

[16] A. Taylor and A. Pacelli, *Mathematics and Politics: Strategy, Voting, Power, and Proof*. Springer, 2008.

[17] S. Kochhar, S. Mazzocchi, and P. Paritosh, "The anatomy of a large-scale human computation engine," in *Human Computation Workshop*. ACM, 2010, pp. 10–17.