Caracterização de Perfil de Uso de Aplicativos Móveis

Augusto C. S. A. Domingues[1], **Fabrício A. Silva**[1], **Thais Regina M. B. Silva**[1]

[1]Universidade Federal de Viçosa - *Campus* Florestal

{augusto.domingues,fabricio.asilva,thais.braga}@ufv.br

Resumo. *O grande volume de informação gerado pelos avanços das tecnologias móveis faz com que os provedores de serviço se interessem cada vez mais pela coleta e análise desses dados. Neste estudo, nós investigamos um conjunto de dados real e de larga escala relacionado ao uso de aplicativos móveis para identificar padrões de acesso, de tráfego de dados, de tempo de uso e de transição entre serviços. A caracterização realizada mostrou que diferentes serviços são consumidos de maneira diferente. Com base nessa investigação, foi proposto um algoritmo para identificar usuários com padrão de uso predominantemente de poucos aplicativos. Os resultados mostraram que foi possível alcançar uma precisão de aproximadamente 97% na identificação desse tipo de usuário.*

Abstract. *The large amount of information generated by the advances of mobile technology makes the service providers get more and more interested in collecting and analyzing such data. In this study, we investigate a real and large-scale dataset related to the use of mobile applications to identify patterns of access, data traffic, period of use, and transition among services. The characterization results reveal that different services are consumed differently by their users. Based on this investigation, we propose an algorithm to identify users with access patterns of predominantly few apps. The results point out that the proposed algorithm leads to good precision when identifying such users.*

1. Introdução

O avanço das tecnologias móveis e sem fio alcançado nos últimos anos fez surgir uma nova frente de pesquisa que foca na análise de grandes volumes de dados [Laurila et al. 2012]. Por um lado, a redução no custo dos dispositivos móveis, principalmente *smartphones*, possibilitou um aumento significativo no número de usuários desses equipamentos. Por outro lado, as empresas que oferecem serviços móveis perceberam a importância em se conhecer os seus usuários, e passaram então a coletar dados de seus clientes.

Um dos grandes interesses na análise de dados está relacionado ao conhecimento de padrões de uso dos serviços móveis, dada a sua importância em diferentes aspectos [Yang et al. 2015]. Por exemplo, operadoras de redes celulares podem tomar melhores decisões em investimento em infraestrutura, focando em áreas com maior demanda. Além disso, as operadoras podem oferecer planos de dados específicos para atender cada perfil de uso, aumentando a retenção de clientes. Empresas de propaganda podem direcionar suas divulgações para usuários com algum perfil específico, aumentando assim a efetividade da propaganda. Esse tipo de empresa pode ainda utilizar os estudos para identificar os melhores dias e horários para divulgações, diminuindo a repulsão dos usuários por

propagandas em momentos inadequados. Por fim, diferentes segmentos de empresas podem utilizar as análises para conhecer melhor o perfil de seus usuários e, assim, prover melhores serviços.

Alguns fatores são fundamentais para a qualidade da caracterização de dados de serviços móveis [Naboulsi et al. 2015]. Primeiramente, a fonte de dados deve ser confiável e condizer com a realidade de uso dos serviços. Outro fator diz respeito ao local da coleta, sendo que quanto mais próxima dos usuários for feita a coleta, isto é, diretamente de seus *smartphones*, maior o nível de detalhes existentes. Além disso, a quantidade de usuários monitorados deve ser significativa para aumentar a representatividade dos resultados. Por fim, os dados devem conter informações relevantes para o contexto da análise.

O principal objetivo deste trabalho é caracterizar o perfil de uso de serviços móveis - com ênfase nas aplicações - a partir de grandes volumes de dados reais coletados diretamente de *smartphones* de mais de 5000 usuários durante o período de Janeiro a Dezembro de 2014. Como resultados, espera-se responder às seguintes perguntas:

1. Qual a frequência e duração de uso de alguns dos principais aplicativos móveis?
2. Existe algum padrão de navegação entre os aplicativos?
3. Qual o impacto de cada aplicativo na rede em termos de dados trafegados?
4. Como o padrão de uso está relacionado com o dia da semana e hora do dia?
5. É possível classificar os usuários de acordo com o padrão de navegação entre aplicativos?

Este texto está organizado da seguinte maneira. A Seção 2 apresenta os principais trabalhos relacionados. Na Seção 3, são descritos os detalhes do conjunto de dados analisado. A Seção 4 faz a análise desses dados com base em métricas de uso de aplicações móveis. Com base nessa análise, a Seção 5 propõe um algoritmo para identificar usuários com padrão de uso predominantemente de poucos aplicativos. Finalmente, as conclusões do trabalho são apresentadas na Seção 6.

2. Trabalhos Relacionados

Alguns trabalhos da literatura abordam a caracterização de dados para identificar perfis de usuários. A seguir, serão apresentados os principais estudos e suas relações com este trabalho.

As redes sociais são fontes interessantes para o estudo de perfil de usuário, e são exploradas em alguns trabalhos. [Jin et al. 2013] abordam as características dos comportamentos dos usuários e as atividades de tráfego de dados em redes sociais como *Facebook* e *Twitter*, e como essas informações podem ser usadas para otimização da infraestrutura de rede. Similarmente, [Fiadino et al. 2015] fazem uma análise quantitativa das redes sociais *Facebook* e *WhatsApp*, do ponto de vista da infraestrutura de rede, com o objetivo de entender como as redes sociais online oferecem seus serviços. [Farseev et al. 2015] utilizam dados coletados de diferentes redes sociais para identificar as características dos usuários com base na localização e no conteúdo criado pelos mesmos. [Benevenuto et al. 2012] produzem uma análise da navegação dos usuários em redes sociais, para entender a frequência e elaborar um grafo de probabilidade de transições entre os conteúdos. Com base nas atividades e interações de usuários de comunidades online, [Fernandez et al. 2014] produzem uma ontologia para modelar o perfil desses

usuários, levando em consideração características extraídas de métodos de análises de redes sociais. Por fim, [Xu et al. 2015, Ferdous et al. 2015] analisam a relação das redes sociais com as características pessoais de um indivíduo e como as mesmas podem afetar a saúde dos usuários.

Diferentemente dos estudos citados, o presente trabalho visa caracterizar e identificar perfis de usuários considerando não somente as redes sociais, mas uma lista mais ampla de aplicações móveis. Além disso, nosso trabalho utiliza dados de duração de um ano coletados por um agente instalado no *smartphone* do usuário, gerando assim informações mais precisas do que as coletadas por meio de monitoração de tráfego de rede.

Outros estudos exploram não somente as redes sociais para identificar perfis de usuários. [Yang et al. 2015] analisam o padrão de uso de dados de cada usuário nas redes móveis através do tráfego HTTP coletado e o relaciona com o uso das aplicações. Com um foco em troca de mensagens SMS, [de Almeida Oliveira et al. 2015] estudam o comportamento de usuários de um serviço de chat móvel para elaborar perfis de usuário. [LeRouge e Ma 2010] desenvolvem perfis de usuário através de métodos qualitativos e apontam como esses podem ser usados para a elaboração de ferramentas de TI na área da saúde. [Chittaranjan et al. 2011, Chittaranjan et al. 2013] estudam a relação entre características comportamentais derivadas de dados provindos de *smartphones* e dados fornecidos pelos usuários; Para isso, desenvolvem um método de inferência de tipo de personalidade do usuário com base no uso de seu celular. [Wang et al. 2015] modelam os padrões de tráfego de torres celulares em ambientes urbanos, classificando o tráfego quanto ao horário e local de acontecimento, permitindo-os agrupar as torres em diferentes regiões como Residencial, Transporte e Escritório. Buscando obter padrões de mobilidade de usuários móveis na presença de grandes eventos, [Xavier et al. 2012] estudam um conjunto de dados que contém ligações feitas por usuários de uma companhia telefônica nos arredores de um estádio de futebol e debatem quem são e como é o comportamento destes usuários.

Existem também trabalhos que tentam identificar perfis de usuários de acordo com os aplicativos instalados em seus dispositivos. [Malmi e Weber 2016] fazem um estudo demográfico dos usuários com base em seus aplicativos instalados. Para isso, utilizam um conjunto de dados que contém atributos como gênero, idade, e estado civil. Similarmente, [Seneviratne et al. 2014] analisam como características de usuários podem ser inferidas através da observação dos aplicativos instalados pelo mesmo, utilizando técnicas de aprendizado supervisionado para prever informações como religião, estado civil e idiomas falados. [Li et al. 2015] questionam os motivos que tornam um aplicativo popular entre milhões de usuários, evidenciando os fatores que levam a padrões de uso e como as aplicações se comportam quanto ao tráfego de rede. Por fim, [Do et al. 2011] analisam quais aplicações são mais comuns dada uma localidade e a quantidade de pessoas na mesma, fator denominado por eles como contexto social; O estudo mostra fortes dependências entre o uso das aplicações e essa variação contextual.

Diferentemente, o presente trabalho não analisa somente a existência ou não de aplicativos, mas também detalhes da interação do usuário com os mesmos, como frequência de uso, duração de cada uso e a transição entre os aplicativos, dentre outros detalhes.

Além dos trabalhos descritos acima, existem outros na literatura que visam o es-

Tabela 1. Características dos principais trabalhos relacionados

Autor	Local da coleta	Critério de análise	Escala	Duração
Yang et al. 2015	Provedor	Tráfego de dados	$\approx$ 4.500.000	Uma semana
Oliveira et al. 2015	Provedor	Padrões de uso de SMS	$\approx$ 20.000	Uma semana
Wang et al. 2015	Provedor	Tráfego de dados	$\approx$ 150.000	Um mês
Gonçalves et al. 2016	Provedor	Tráfego de dados	$\approx$ 150.000	Um ano
Malmi e Weber 2016	Provedor	Uso de aplicações	$\approx$ 3.500	Um mês
Ferdous et al. 2015	*Smartphone*	Uso de aplicações	$\approx$ 25	Seis semanas
Chittaranjan et al. 2011	*Smartphone*	Características pessoais	$\approx$ 120	Dezessete meses
Xavier et al. 2012	Provedor	Padrões de mobilidade	$\approx$ 30.000	Três dias

tudo de perfil de usuários para fins específicos. [Pavan et al. 2015] usam dados provindos de *smartphones* para identificar pontos de interesse dos usuários, com base na frequência e na intensidade das visitas. Adicionalmente, demonstram como locais cotidianos - como congestionamentos - podem ser interpretados erroneamente como locais importantes para um usuário. [Iwata et al. 2013] desenvolvem um sistema de pesquisa que visa facilitar o uso dos dispositivos móveis por parte dos usuários, considerando fatores como a localização, hora do dia e dados fornecidos pelo usuário. Por sua vez, [Gonçalves et al. 2016] estudam especificamente a ferramenta *Dropbox* para identificar perfis de colaboração; Com isso, podem avaliar o impacto de trabalhos colaborativos na utilização de serviços de armazenamento na nuvem em grandes redes.

A caracterização de dados para identificação de perfis de usuários é uma área que vem crescendo nos últimos anos, dada a importância deste conhecimento para as tomadas de decisão. Para isso, o quanto mais próxima do usuário a coleta dos dados é feita, maior será a acurácia das análises. Porém, a caracterização com base nas aplicações móveis (*Apps*) específicas em uso ainda é pouco explorada. Além disso, existem poucos trabalhos que coletam dados diretamente dos dispositivos móveis e em larga escala, como pode ser visto na Tabela 1, que apresenta as características dos principais trabalhos relacionados. Neste trabalho, é feita uma caracterização de perfil de uso de aplicativos móveis com base em dados coletados diretamente de milhares de dispositivos móveis durante um ano.

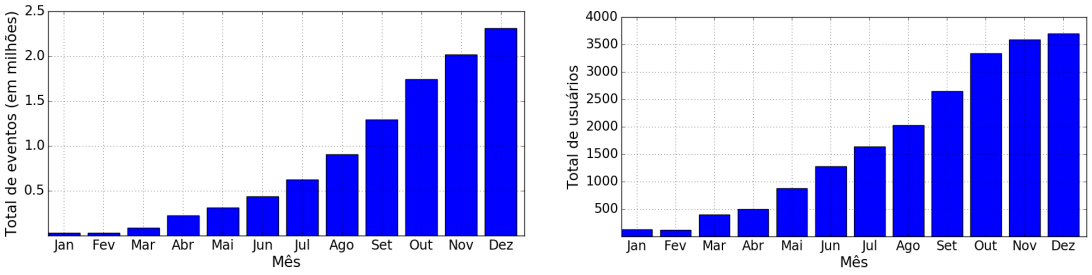
3. Conjunto de dados

3.1. Coleta

O conjunto de dados usado contém informações anônimas relacionadas ao uso de aplicativos móveis em *smartphones*. A coleta dos dados foi feita através de um agente instalado em dispositivos *Android* de usuários voluntários. Essa coleta ocorreu durante o período de um ano, de Janeiro a Dezembro de 2014, em algumas regiões do Brasil (principalmente nos estados de São Paulo e Rio de Janeiro). O conjunto de dados possui 12,060,344 de registros vindos de 5,643 usuários de planos pós-pago. O número de usuários monitorados foi crescendo ao longo do ano, e a distribuição desses e seus acessos ao longo do período avaliado pode ser observada nas Figuras 1(a) e 1(b). Cada registro do conjunto de dados representa um evento realizado pelo usuário e contém os seguintes campos:

- *Subscriber*: um identificador único e anônimo para cada usuário;
- *Aplicação*: nome da aplicação usada;

Figura 1. Características do conjunto de dados



(a) Eventos por mês

(b) Usuários distintos por mês

- *Tecnologia da rede*: tipo de rede usada durante a ocorrência do evento - Wifi ou Móvel;
- *Bytes Recebidos*: total de *bytes* recebidos ao longo da duração do evento;
- *Bytes Transmitidos*: total de *bytes* transmitidos ao longo da duração do evento;
- *Data de início*: Data e hora de início do evento;
- *Data de término*: Data e hora de término do evento.

Visando um melhor entendimento das aplicações utilizadas pelos participantes, as mesmas foram divididas em 5 categorias, de acordo com suas funcionalidades. As aplicações presentes no conjunto de dados e as categorias elaboradas estão listadas abaixo:

- Social: *Facebook*, *Instagram* e *WhatsApp*
- Escritório: *IBM Notes*
- Mapas: *Waze social GPS*
- Entretenimento: *Youtube* e *Netflix*
- Navegação: *Chrome*, *Firefox* e *Opera*

Durante o período de monitoração, foi gerado um evento com os campos descritos acima todas as vezes que o usuário abriu um dos aplicativos monitorados. Com isso, é possível conhecer detalhes do uso dos aplicativos pelos usuários.

3.2. Filtro

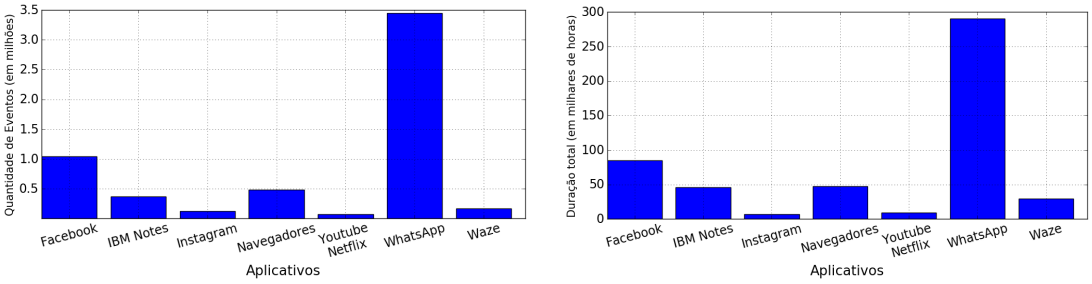
Para evitar o impacto de alguns usuários com perfil de geração de evento diferenciado na análise (gerando um viés de acordo com o perfil de uso dos mesmos), primeiramente foi feito um filtro para descartar alguns usuários das análises. Para isso, fizemos uma análise estatística da quantidade de eventos gerados por cada usuário, e do período de tempo que cada usuário foi monitorado. Com essa análise, vimos que 10,44% dos usuários geraram menos que 20 eventos, e 6,76% mais que 10000 eventos. Além disso, alguns usuários foram monitorados por menos que 4 semanas. Resolvemos então desconsiderar os usuários que se encaixam nas seguintes características:

- $20 \leq$ Número de eventos ≤ 10000
- Número de dias monitorados ≤ 30

3.3. Limitações

Os dados utilizados neste trabalho foram coletados de usuários de planos pós-pagos de uma operadora de celular. Com isso, a representatividade das análises não engloba os

Figura 2. Análise dos aplicativos



(a) Acessos por aplicativo

(b) Tempo total de uso por aplicativo

usuários de planos pré-pagos, que podem apresentar um perfil de uso diferente. No Brasil, os usuários de planos pós-pagos representam aproximadamente 30% dos contratos das operadoras [TELECO 2016]. Porém, em termos de valores de receita, representam um percentual significativo para as operadoras. Isso mostra que, apesar de englobar um percentual menor da população, os dados analisados são relevantes para as operadoras e provedores de serviços móveis, e portanto, importantes de serem analisados.

Em relação à distribuição espacial, a maioria dos usuários monitorados se concentram na região Sudeste do Brasil, principalmente nas cidades de São Paulo e Rio de Janeiro. Apesar de representarem uma pequena parte do país, essa região é responsável por 60% do PIB, sendo que as duas cidades de maior abrangência nos dados são responsáveis por 14,45% do PIB e 8,99% da população do país.

Em resumo, os resultados apresentados neste trabalho não podem ser considerados corretos para toda a população brasileira em todas as suas regiões. Por se tratar de um território extenso, sabe-se das particularidades de cada região, estado e até mesmo cidade. Assim sendo, as análises representam uma parcela da população que utiliza planos pós-pagos na região Sudeste do país, parcela essa que tem uma significativa importância para a economia do mesmo.

4. Caracterização do Uso de Aplicativos Móveis

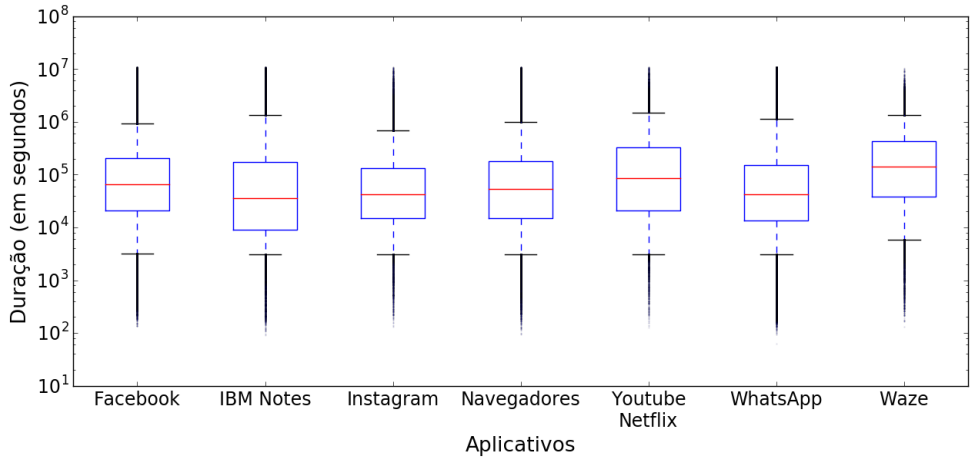
Nesta seção analisamos como as diferentes métricas de uso de aplicações móveis podem definir comportamentos dos usuários. Para isso, estudamos quatro características: os acessos aos aplicativos, com o intuito de obter a frequência e duração de uso dos mesmos; os padrões de navegação entre os aplicativos, que representam os fluxos mais comuns de uso; os dados trafegados, isto é, o tráfego de rede (*Download* e *Upload*) em cada evento; e por fim, a perspectiva temporal dos dados, verificando o comportamento dos usuários em relação aos aplicativos quanto aos dias da semana e às horas do dia.

4.1. Perspectiva de Acessos

Identificar a aplicação que os usuários estão mais interessados pode nos ajudar a prever o comportamento dos mesmos [Yang et al. 2015]. Nesta subseção analisamos a quantidade de eventos e a duração dos eventos por aplicativo.

Observando a Figura 2(a), é possível notar que as aplicações com o maior número de acessos são *WhatsApp*, *Facebook*, e os Navegadores, nessa ordem. De todas as aplicações estudadas, os Navegadores são os únicos que se encaixam no perfil de qualquer

Figura 3. Gráfico *Boxplot* da duração dos eventos por aplicativo

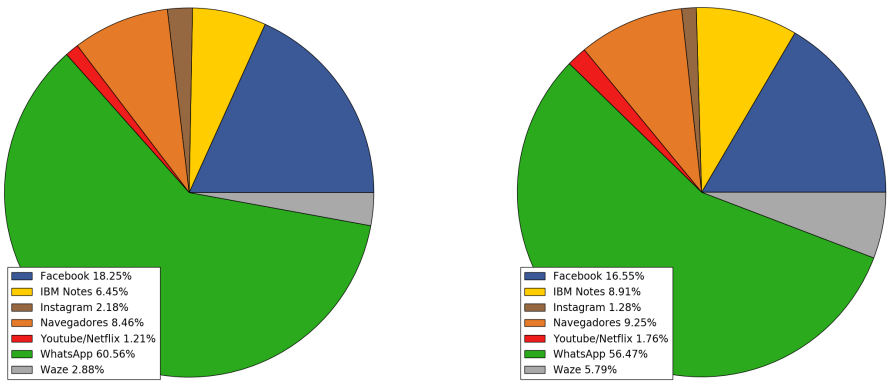


usuário, dada a sua natureza de uso geral, o que justifica o grande número de acessos. As aplicações da categoria *Social* também estão presentes nos dispositivos de grande parte dos usuários, que tendem a realizar inúmeros acessos por dia a esses aplicativos.

A análise quanto ao tempo de uso por aplicativo (Figura 2(b)) nos mostra a relação entre a duração dos acessos e a quantidade de acessos. Novamente, os aplicativos com maior tempo total de uso são *WhatsApp*, *Facebook* e os Navegadores.

Na Figura 3, temos um gráfico de *Boxplot* da duração dos eventos para cada aplicativo. Apesar de todas possuírem comportamento semelhante, podemos destacar algumas particularidades. *Youtube*, *Netflix* e *Waze* apresentam uma mediana de duração de acessos maior quando comparados aos outros aplicativos. Isso mostra que essas aplicações possuem um padrão de acesso prolongado. Usuários do *Waze* tendem a acessá-lo por longos períodos de tempo durante um trajeto. Já os usuários do *Youtube* e *Netflix* os utilizam para atividades de entretenimento, como assistir vídeos e filmes. Na Figura 4(a), podemos observar que estes aplicativos são responsáveis por somente 4.09% dos eventos analisados, porém a Figura 4(b) mostra que os mesmos possuem uma fatia maior (7.55%) quanto ao tempo total de uso dos eventos analisados.

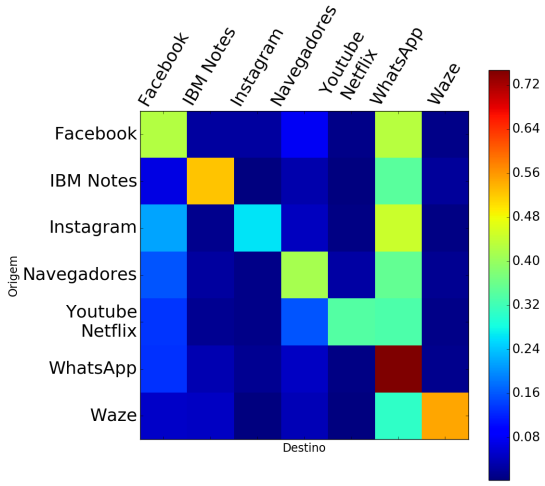
Figura 4. Análise percentual dos aplicativos



(a) Acessos por aplicativo

(b) Tempo total de uso por aplicativo

Figura 5. Mapa de nível das transições entre aplicativos



4.2. Perspectiva de Navegação entre Aplicativos

A análise da navegação entre aplicativos é importante para que se conheça quais são os fluxos de uso dos usuários. Segundo [Li et al. 2015], diferentes usuários possuem diferentes padrões de usos de aplicações. Para os provedores de serviços, tais informações podem ajudar a prover planos mais personalizados que favoreçam aplicativos correlacionados.

A Figura 5 mostra um mapa de nível gerado das transições entre aplicativos de todos os usuários em conjunto. Cada região do mapa representa a probabilidade de transição entre dois aplicativos (o eixo Y representa o último aplicativo usado e o eixo X o próximo a ser acessado). Nota-se que, ao sair do *WhatsApp*, o usuário tem cerca de 75% de probabilidade de acessá-lo novamente. Isso demonstra que a aplicação é de uso recorrente, ou seja, o usuário realiza múltiplos acessos seguidos (neste caso, para visualizar novas mensagens e respondê-las). Outras transições com alta recorrência são *IBM Notes* para *IBM Notes* (uso frequente no trabalho), e *Waze* para *Waze* (uso frequente no trânsito).

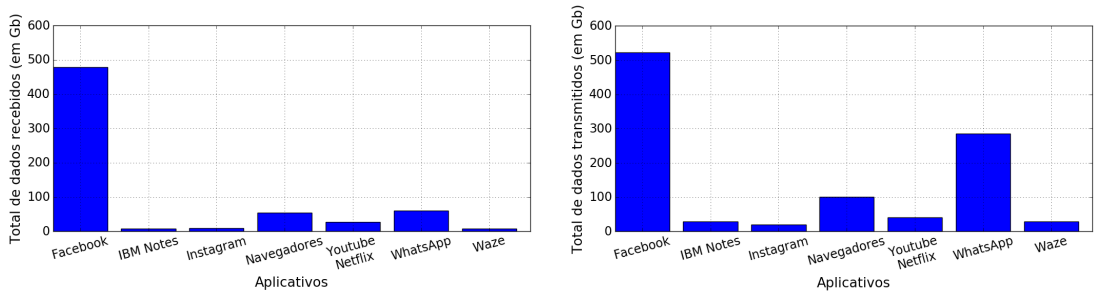
Outra análise que pode ser feita com base na Figura 5 são as transições que ocorrem para os aplicativos *WhatsApp* (vistas na coluna *WhatsApp*) e *Facebook* (vistas na coluna *Facebook*). Pode-se afirmar que todos os usuários tendem a realizar acessos alternados entre suas aplicações preferidas, o *Facebook* e o *WhatsApp*. Isto se explica pela popularidade das aplicações da categoria *Social*, que é a categoria dominante na Internet móvel [Yang et al. 2015].

Vale ressaltar que a análise feita se refere a todos os usuários em conjunto. Porém, cada usuário possui um perfil diferente de transição entre aplicativos, e essa análise individual é feita na Seção 5, em que é realizado um estudo para identificar o perfil de cada usuário.

4.3. Perspectiva de Dados Trafegados

Conhecer o comportamento dos usuários quanto ao uso de dados é uma necessidade urgente para os provedores de serviço [Yang et al. 2015]. Da mesma forma, [Wang et al. 2015] afirmam que entender os padrões de tráfego de torres celulares em centros urbanos de larga escala é extremamente valioso para os provedores de Internet e usuários móveis.

Figura 6. Análise dos dados trafegados



(a) Download total por aplicativo

(b) Upload total por aplicativo

Assim, esta subseção analisa os eventos em relação aos dados trafegados, para que se conheça as aplicações que causam um maior impacto nas redes móveis.

As Figuras 6(a) e 6(b) apresentam a quantidade total de dados enviados e recebidos para cada aplicativo, respectivamente. Com base nelas, é possível observar que as aplicações com a maior quantidade de *upload* e *download* são *Facebook* e *WhatsApp*. Esses aplicativos permitem que os usuários enviem e recebam arquivos multimídia, como fotos, vídeos e músicas, o que justifica o uso elevado de dados. Segundo [Fiadino et al. 2015] mais de 75% do tráfego de dados do WhatsApp corresponde ao compartilhamento desses arquivos.

Adicionalmente, nota-se que o aplicativo *WhatsApp* apresenta grande diferença de sua quantidade total de *upload* (aproximadamente 300Gb) para o seu total de *download* (menos de 100Gb). Nas Figuras 7(a) e 7(b), vemos que, enquanto essa aplicação é responsável por somente 9,36% do *download* total, os dados transmitidos por ela correspondem a 27,9% do *upload* total. Desta forma, conclui-se que os usuários do *WhatsApp* tendem a enviar mais conteúdo (mensagens de texto ou multimídia) do que receber.

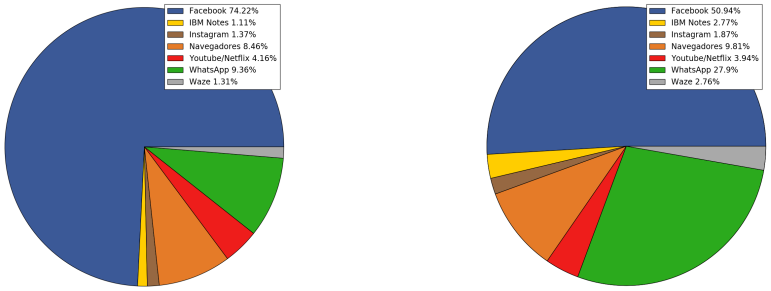
A identificação do contexto das aplicações mais dependentes de rede pode ajudar a caracterizar o comportamento de um grupo de usuários. Na análise feita, é possível observar que as aplicações pertencentes à categoria *Social* são responsáveis pela maior carga da rede, dadas as suas características de uso. Assim, usuários frequentes dessa categoria podem se interessar por planos de telefonia que focam no acesso a dados da rede móvel. Outra observação interessante é que, mesmo o *WhatsApp* sendo responsável por 60,56% do número de eventos e 56,47% da duração total (ver Figuras 3(a) e 3(b)), ele representa apenas 9,36% do *download* e 27,9% do *upload* total.

4.4. Perspectiva de Dia da Semana e Hora do Dia

A perspectiva temporal é extremamente importante para a análise do comportamento dos usuários. Com base nela, é possível identificar e modelar padrões de acesso quanto a hora do dia e o dia da semana, por exemplo. Esses padrões podem ser usados por provedores de serviço para otimizar os recursos da rede através de mecanismos de engenharia de tráfego baseados em tempo, ajustando-os dinamicamente com base nas predições de carga [Fiadino et al. 2015].

A Figura 8(a) nos mostra a distribuição temporal dos eventos em relação às horas do dia. É possível observar que eles acontecem em maior quantidade das 06:00h (horário

Figura 7. Análise percentual dos dados trafegados



(a) Download total por aplicativo

(b) Upload total por aplicativo

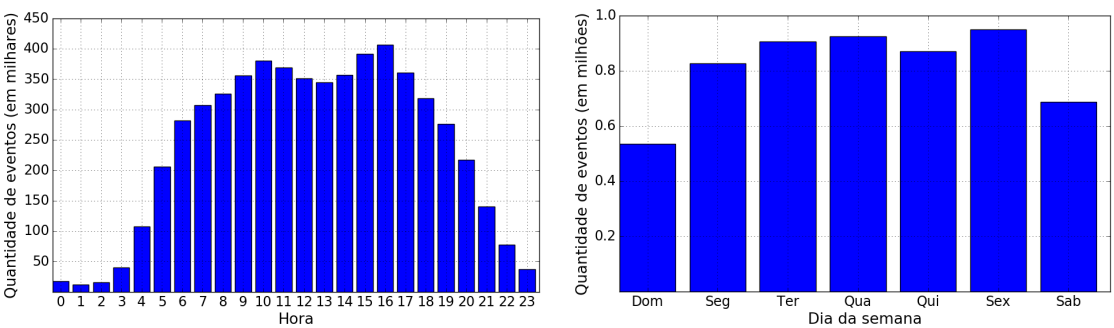
que muitos usuários saem para o trabalho) até às 19:00h (horário que os mesmos retornam para casa). Nesse intervalo, temos um pico às 10:00h , uma ligeira queda nos acessos de 11:00h até às 13:00h, e novamente um pico às 16:00h. Assim, podemos concluir que os usuários tendem a diminuir o fluxo de acessos durante o horário de almoço.

Semelhante, a Figura 8(b) apresenta a distribuição temporal dos eventos em relação aos dias da semana. Nela, nota-se que nos finais de semana (Sábado e Domingo) o número de acessos é menor do que nos dias úteis. Isso implica que os usuários diminuem o uso de aplicativos móveis em seu tempo livre (o que também pode ser visto na Figura 8(a).

A seguir, realizamos a análise temporal para cada aplicação, considerando ambas as métricas de acessos por hora do dia e por dia da semana.

Quanto às aplicações da categoria *Social*, podemos observar, na Figura 9(a), que os usuários do *Facebook* realizam muitos acessos durante todos os dias da semana, com um crescimento durante as tardes. Para o *WhatsApp* (Figura 9(b)), temos um padrão de acesso semelhante, porém pode-se observar uma queda nos acessos nos finais de semana. Inversamente, os usuários do *Instagram* tendem a acessar a aplicação com maior frequência nos finais de semana (Figura 9(c)). Portanto, não se pode chegar a um padrão para a categoria, visto que o contexto específico de cada aplicativo influencia diretamente nos horários de utilização.

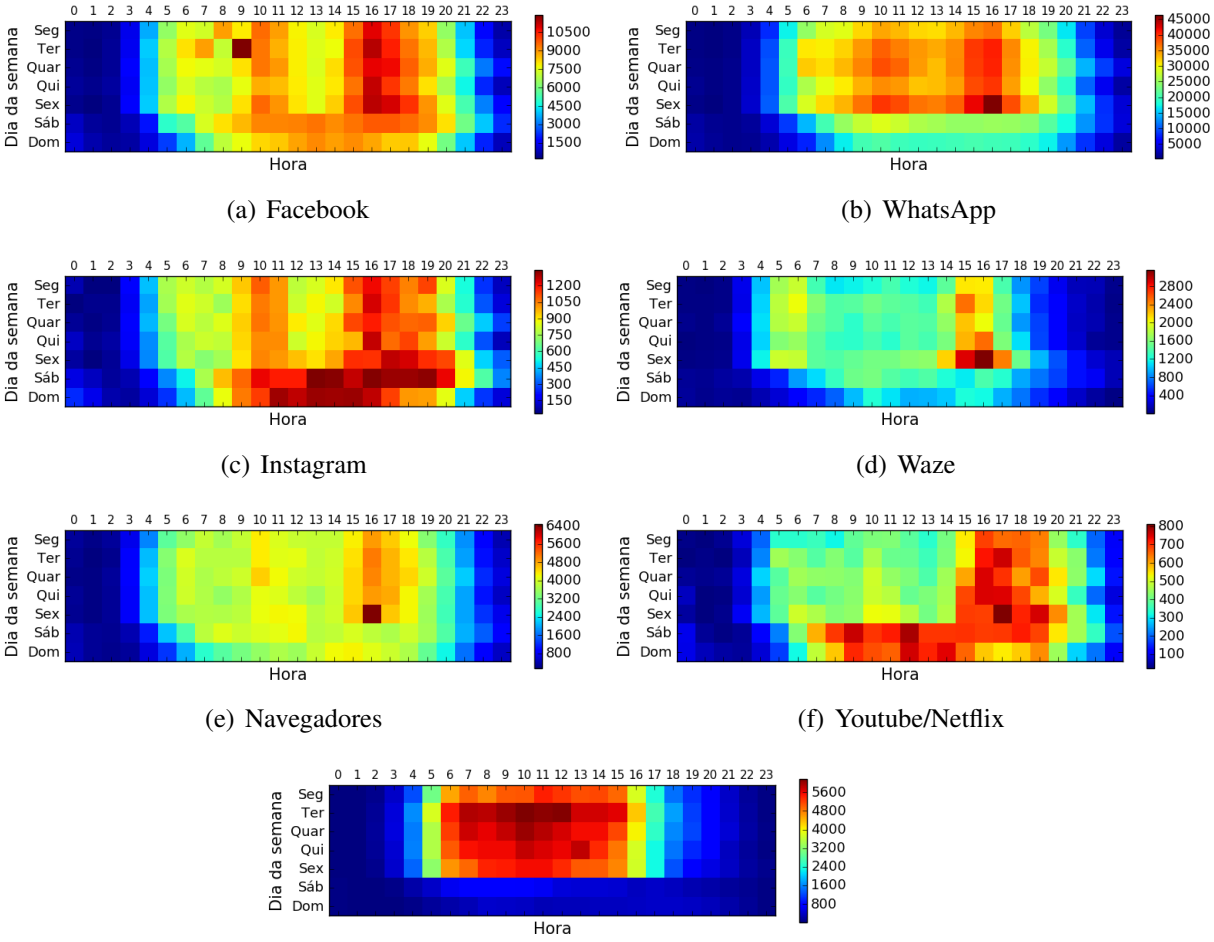
Figura 8. Análise temporal dos dados



(a) Eventos por hora do dia

(b) Eventos por dia da semana

Figura 9. Acessos agrupados por dia e hora



(a) Facebook

(b) WhatsApp

(c) Instagram

(d) Waze

(e) Navegadores

(f) Youtube/Netflix

(g) IBM Notes

Para o *Waze*, a Figura 9(d) mostra que seus acessos ocorrem com muita frequência no fim da tarde, principalmente às sextas-feiras. Esse comportamento incomum pode ser explicado pelo fato do aplicativo ser usado para a mobilidade no trânsito. Com isso, conclui-se que a frequência de acesso está relacionada à necessidade do usuário de saber informações sobre o trânsito. Mais além, os horários de pico de acesso podem revelar eventos atípicos, como congestionamentos e acidentes.

Os aplicativos de navegação na *Web* (Figura 9(e)) possuem uma distribuição homogênea de acessos, abrangendo grande parte do dia e da semana. Este padrão pode ser explicado pela generalidade desses aplicativos, isto é, a necessidade do uso pode ocorrer em diversos contextos.

Algumas aplicações apresentam padrões de uso que refletem diretamente as suas categorias. No caso dos aplicativos *Youtube* e *Netflix*, é visível o maior número de acessos durante o fim da tarde e principalmente durante o fim de semana. Esse padrão compreende o período em que os usuários não estão trabalhando e que dedicam a si próprios. Já para o *IBM Notes* (Figura 9(g)), que contém funcionalidades que são usadas durante o horário de trabalho, como visualização de *emails*, o padrão é ainda mais claro: muitos acessos de

05:00h até às 17:00h de Segunda à Sexta-Feira, sem nenhum outro pico.

5. Identificação do Perfil dos Usuários

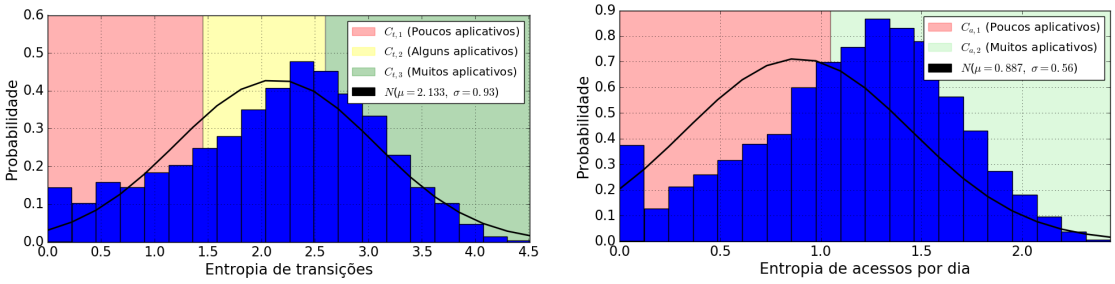
O objetivo desta seção é propor um algoritmo para identificar usuários que predominantemente usam poucos aplicativos. Esses, em sua maioria, possuem uma frequência alta de uso de tais aplicativos, como por exemplo os usuários das aplicações da categoria Social. Assim, a identificação desses usuários pode ser utilizada pelas operadoras para identificar quais aplicações são mais populares e realizar aprimoramentos em seus serviços, como ações de marketing direcionadas, planos que melhor atendam a esses grupos, e melhorias quanto à infraestrutura de tráfego de rede.

5.1. Agrupamento

A primeira etapa do algoritmo visa separar os usuários do conjunto de dados estudado em grupos e entender esses grupos. Para realizar o agrupamento dos usuários, utilizamos técnicas de aprendizagem de máquina, como a análise de *clustering*. De acordo com [Jain et al. 1999], *clustering* é a classificação não supervisionada de padrões (observações, dados ou vetores de características) em grupos (*clusters*). Esta tarefa nos permite dividir os usuários em grupos que possuem características em comum. Assim, definimos os seguintes passos para o cálculo dos agrupamentos:

1. Para cada usuário u, obtemos o seu conjunto de transições T_u composto pelas transições $t_{i,j}$ entre aplicativos. Cada transição $t_{i,j}$ representa a probabilidade do usuário abrir o aplicativo j logo após ter usado o aplicativo i, em um intervalo de tempo menor que 5 horas. Desconsideramos as transições que ocorrem com grandes intervalos de tempo para evitar contabilizar casos em que o acesso ao *smartphone* foi interrompido por algum motivo, como durante o sono, por exemplo.
2. Assim, cada tupla T_u de transições entre n aplicativos $(t_{0,1}, t_{0,2}, ..., t_{n,n})$ representa um usuário e está associada a uma entropia e_{tu};
3. A entropia de Shannon [Shannon 2001] é calculada para cada tupla T_u com base nas probabilidades $t_{i,j}$. Ela representa a quantidade de informação produzida pelo usuário u, sendo que um usuário com valores pequenos de probabilidade, distribuídos entre vários aplicativos, terá um valor alto de entropia. Por outro lado, valores baixos de entropia indicam valores altos de probabilidades distribuídos entre poucos aplicativos;
4. Com o conjunto de transições obtido, calculamos os agrupamentos (*clusters*) dos usuários quanto à sua entropia. Usamos o algoritmo *X-Means* [Pelleg et al. 2000], que é uma extensão do *K-Means*. O *X-Means* realiza o agrupamento dos dados calculando automaticamente a quantidade ótima de *clusters* a gerar - neste caso, obtemos $X = 3$. Assim, os usuários foram divididos em 3 grupos C_t: aqueles em que as transições são predominantemente entre poucos aplicativos (1 a 2 aplicativos) ($C_{t,1}$), transições entre alguns aplicativos (3 a 5 aplicativos) ($C_{t,2}$), e transições entre muitos aplicativos (6 a 7 aplicativos) ($C_{t,3}$).
5. Além das transições, para cada usuário u, obtemos o seu número de acessos por aplicativo por dia, que compõem o conjunto de acessos A_u. Cada acesso a_i representa a probabilidade do usuário acessar o aplicativo i em algum dia arbitrário.

Figura 10. PDF dos agrupamentos



(a) Transições entre aplicativos

(b) Acessos por dia

6. Calculamos a entropia e_{au} para cada tupla A_u com base em suas probabilidades a_i;
7. Por fim, obtemos os agrupamentos C_a dos usuários quanto aos acessos por dia por aplicativo. Novamente, usamos o algoritmo *X-Means*, que desta vez nos forneceu $X = 2$ como quantidade ótima de *clusters*. Assim, dividimos os usuários em 2 grupos: aqueles com acessos a poucos aplicativos por dia (1 a 3) ($C_{a,1}$), e aqueles com acesso a muitos aplicativos por dia (mais que 3) ($C_{a,2}$).

Ao final destes passos, temos dois conjuntos de agrupamentos, C_t e C_a, que categorizam os usuários quanto às transições entre aplicativos e quanto aos acessos por aplicativo por dia. As Figuras 10(a) e 10(b) apresentam a função densidade de probabilidade (*PDF*) para as entropias de transição e de acesso, respectivamente. Também é possível ver a extensão de cada grupo. A Tabela 2 fornece mais detalhes, destacando os centróides e a quantidade de usuários por agrupamento. Pode-se perceber pelos gráficos que a distribuição das entropias de transição segue uma curva Normal ($\mu = 2.133, \sigma = 0.93$). Em relação à distribuição das entropias de acessos, percebe-se uma aproximação com a curva Normal ($\mu = 0.887, \sigma = 0.56$).

A seguir, utilizamos estes agrupamentos e suas características para identificar o perfil de uso de um usuário desconhecido.

5.2. Perfil de uso de um usuário desconhecido

O algoritmo proposto abaixo foi elaborado para analisar e categorizar o perfil de uso de um usuário desconhecido, utilizando para isso os agrupamentos definidos nesta seção. O objetivo é oferecer uma possibilidade para que o perfil de um usuário seja identificado.

O algoritmo recebe como entrada T_u (tupla de transição) e A_u (tupla de acesso) do usuário u e procede para os seguintes passos:

Tabela 2. Centróide e total de usuários de cada *Cluster*

Clusters		Centróide	Usuários
Transições	#1 (Poucos aplicativos)	0.805494	1252
	#2 (Alguns aplicativos)	2.094707	2287
	#3 (Muitos aplicativos)	3.096514	1821
Acessos	#1 (Poucos aplicativos)	0.623548	2018
	#2 (Muitos aplicativos)	1.469425	3342

1. Calcula-se a entropia de transições e_{tu} e de acessos e_{au} do usuário com base em T_u e A_u, respectivamente;
2. Para a distribuição quanto à entropia de transição (Figura 11(a)), calcula-se a distância d_{tu} de e_{tu} em termos do desvio padrão da distribuição das entropias obtida com base nos dados conhecidos, sendo:

$$d_{tu} = \frac{e_{tu} - \mu_t}{\sigma_t} \tag{1}$$

   onde μ_t é a média e σ_t o desvio padrão da entropia de transição da distribuição conhecida. Assim, temos quantos desvios e_{tu} está da média;
3. Se $d_{tu} < 0.5$, então e_{tu} se encontra mais à esquerda da distribuição (com entropia baixa e poucos aplicativos nas transições) e seguimos para o próximo passo. Caso contrário, o usuário é predominantemente de muitos aplicativos e não se encaixa no grupo escolhido. O ponto de corte de 0.5 foi calculado com base na distribuição da entropia de transições e nos agrupamentos obtidos pelos dados conhecidos (vistos na Figura 10(a)).
4. Mesmo que o cálculo anterior indique que o usuário pertença ao grupo que utiliza poucos aplicativos, ainda há chances de erros ocorrerem devido às margens de erro da distribuição. Para diminuir o risco de falsos-positivos (usuários de muitos aplicativos que podem ser indicados como usuários de poucos aplicativos), utilizamos também a entropia de acesso e_{au} para aumentar a precisão do algoritmo.
5. Para a distribuição quanto à entropia de acesso (Figura 11(b)), calcula-se a distância d_{au} de e_{au} em termos do desvio padrão da distribuição de entropias conhecidas com base nos dados existentes, assim como foi feito para d_{tu}:

$$d_{au} = \frac{e_{au} - \mu_a}{\sigma_a} \tag{2}$$

   onde μ_a é a média e σ_a o desvio padrão da entropia de acessos.
6. Se $d_{au} < 0.4$, então e_{au} se encontra mais à esquerda da distribuição e o usuário pode ser classificado como predominantemente de poucos aplicativos. O ponto de corte de 0.4 foi calculado com base na distribuição da entropia de acessos e nos agrupamentos obtidos pelos dados conhecidos (vistos na Figura 10(b)).
7. Se o usuário atendeu aos critérios dos passos 3 e 6, o algoritmo o classifica como predominante de poucos aplicativos. Os aplicativos que o usuário predominantemente utiliza são então escolhidos pela ordem de probabilidade de acesso; são escolhidos X aplicativos, com $1 \leq X \leq 3$, em que a soma das probabilidades de acesso por parte do usuário seja maior que 90%.

O algoritmo entrega como resultado a indicação se o usuário predominantemente utiliza um número pequeno de aplicativos, e quais são esses aplicativos.

Para validar o algoritmo, foram geradas instâncias aleatórias de usuários com uma classe de uso conhecida. A Tabela 3 ilustra os resultados desta validação. Pode-se observar que o algoritmo obteve 97,6% de acurácia, portanto o mesmo pode ser considerado eficiente para a identificação de usuários com perfil de uso predominantemente de poucos aplicativos.

Tabela 3. Matriz de confusão para os testes realizados com o algoritmo proposto

		Classe apontada	
		Poucos aplicativos	Muitos aplicativos
Classe real	Poucos aplicativos	494	0
	Muitos aplicativos	24	482

6. Conclusão e Trabalhos Futuros

Neste artigo apresentamos a análise e a caracterização do comportamento de usuários de *smartphones* de algumas regiões do Brasil. Nós descrevemos os padrões de uso dos principais aplicativos móveis, com base em um conjunto de dados com milhões de registros de acessos por parte de milhares de usuários durante um ano.

Primeiro, definimos as métricas que foram usadas para avaliar os usuários quanto aos seus acessos. Foi possível observar que aplicações como redes sociais são responsáveis por grande parte da quantidade total de eventos gerados e do tempo total de uso. Observamos também que os usuários tendem a realizar acessos alternados entre suas aplicações prediletas e as redes sociais, reforçando a idéia de que esta categoria é a dominante na *Internet* móvel. Adicionalmente, vimos que apesar de ser responsável pela maior parte dos acessos, o *WhatsApp* representa somente uma pequena fatia do tráfego total de dados. Por fim, a análise temporal demonstrou que as aplicações seguem padrões bem definidos de hora e dia de uso, principalmente as pertencentes às categorias *Entretenimento* e *Escritório*. Esses resultados podem ser explorados pelos provedores de serviço em suas tomadas de decisão.

Com base nesta análise, propomos um algoritmo para identificar usuários que predominantemente usam poucos aplicativos. Esses, em sua maioria, possuem uma frequência alta de uso. Portanto, sua identificação pode ser utilizada pelas operadoras para realizar diversos aprimoramentos em seus serviços.

Trabalhos futuros incluem a análise do conjunto de dados quanto aos padrões de chamadas de voz realizadas, uma vez que este artigo focou no comportamento dos usuários quanto aos acessos a aplicativos. Com isso, será possível obter um resultado ainda mais completo quanto às características de usuários móveis.

Mais adiante, planejamos estudar o comportamento dos usuários durante a realização da Copa do Mundo 2014, que aconteceu no Brasil. Para isso, podemos considerar outros dados interessantes para a análise, como os padrões de mobilidade e a variação no sinal de rede. Assim, será possível avaliar também os impactos causados por um grande evento na infraestrutura de rede dos centros urbanos.

7. Agradecimento

Este trabalho contou com o apoio da Fapemig, CNPq e FUNARBE.

Referências

Benevenuto, F., Rodrigues, T., Cha, M., e Almeida, V. (2012). Characterizing user navigation and interactions in online social networks. *Information Sciences*, 195:1–24.

Chittaranjan, G., Blom, J., e Gatica-Perez, D. (2011). Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *2011 15th Annual International Symposium on Wearable Computers*, pages 29–36. IEEE.

Chittaranjan, G., Blom, J., e Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 17(3):433–450.

de Almeida Oliveira, R., Brandão, W. C., e Marques-Neto, H. T. (2015). Characterizing user behavior on a mobile sms-based chat service. In *Computer Networks and Distributed Systems (SBRC), 2015 XXXIII Brazilian Symposium on*, pages 130–139. IEEE.

Do, T. M. T., Blom, J., e Gatica-Perez, D. (2011). Smartphone usage in the wild: a large-scale analysis of applications and context. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 353–360. ACM.

Farseev, A., Nie, L., Akbari, M., e Chua, T.-S. (2015). Harvesting multiple sources for user profile learning: a big data study. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 235–242. ACM.

Ferdous, R., Osmani, V., e Mayora, O. (2015). Smartphone app usage as a predictor of perceived stress levels at workplace. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*, pages 225–228. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Fernandez, M., Scharl, A., Bontcheva, K., e Alani, H. (2014). User profile modelling in online communities. In *Proceedings of the Third International Conference on Semantic Web Collaborative Spaces-Volume 1275*, pages 1–15. CEUR-WS. org.

Fiadino, P., Casas, P., Schiavone, M., e D'Alconzo, A. (2015). Online social networks anatomy: On the analysis of facebook and whatsapp in cellular networks. In *IFIP Networking Conference (IFIP Networking), 2015*, pages 1–9. IEEE.

Gonçalves, G. D., Vieira, A. B., da Silva, A. P. C., e Almeida, J. M. (2016). Trabalho colaborativo em serviços de armazenamento na nuvem: Uma análise do dropbox.

Iwata, M., Miyamoto, H., Hara, T., Komaki, D., Shimatani, K., Mashita, T., Kiyokawa, K., Uemukai, T., Hattori, G., Nishio, S., et al. (2013). A content search system considering the activity and context of a mobile user. *Personal and ubiquitous computing*, 17(5):1035–1050.

Jain, A. K., Murty, M. N., e Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jin, L., Chen, Y., Wang, T., Hui, P., e Vasilakos, A. V. (2013). Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, 51(9):144–150.

Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.-M.-T., Dousse, O., Eberle, J., Miettinen, M., et al. (2012). The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, number EPFL-CONF-192489.

LeRouge, C. e Ma, J. (2010). User profiles and personas in consumer health technologies. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE.

Li, H., Lu, X., Liu, X., Xie, T., Bian, K., Lin, F. X., Mei, Q., e Feng, F. (2015). Characterizing smartphone usage patterns from millions of android users. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 459–472. ACM.

Malmi, E. e Weber, I. (2016). You are what apps you use: Demographic prediction based on user's apps. *arXiv preprint arXiv:1603.00059*.

Naboulsi, D., Fiore, M., Ribot, S., e Stanica, R. (2015). Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161.

Pavan, M., Mizzaro, S., e Scagnetto, I. (2015). Mining movement data to extract personal points of interest: A feature based approach.

Pelleg, D., Moore, A. W., et al. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, volume 1.

Seneviratne, S., Seneviratne, A., Mohapatra, P., e Mahanti, A. (2014). Predicting user traits from a snapshot of apps installed on a smartphone. *ACM SIGMOBILE Mobile Computing and Communications Review*, 18(2):1–8.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

TELECO (2016). Estatísticas de celulares no brasil. `http://www.teleco.com.br/ncel.asp`. Acessado em 21/11/2016.

Wang, H., Xu, F., Li, Y., Zhang, P., e Jin, D. (2015). Understanding mobile traffic patterns of large scale cellular towers in urban environment. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 225–238. ACM.

Xavier, F. H. Z., Silveira, L. M., Almeida, J. M. d., Ziviani, A., Malab, C. H. S., e Marques-Neto, H. T. (2012). Analyzing the workload dynamics of a mobile phone network in large scale events. In *Proceedings of the first workshop on Urban networking*, pages 37–42. ACM.

Xu, R., Frey, R. M., Vuckovac, D., e Ilic, A. (2015). Towards understanding the impact of personality traits on mobile app adoption-a scalable approach. In *23rd European Conference on Information Systems*.

Yang, J., Qiao, Y., Zhang, X., He, H., Liu, F., e Cheng, G. (2015). Characterizing user behavior in mobile internet. *IEEE Transactions on Emerging Topics in Computing*, 3(1):95–106.