Análise de Perfis de Usuários de Música e Seus Impactos no Desempenho de Políticas de Substituição de *Cache*

Stéfani Pires[1,2], Francisco R. C. Araújo[1], Allan Freitas[2], Leobino Sampaio[1]

[1]Programa de Pós-Graduação em Ciência da Computação (PGCOMP)
Departamento de Ciência da Computação – Universidade Federal da Bahia (UFBA)
Salvador – BA – Brasil

[2]Instituto Federal da Bahia (IFBA) – Salvador – BA – (40.301-015)

{stefani.pires,franciscorca,leobino}@ufba.br, allanedgard@acm.org

Abstract. *The exploration of human behavior patterns is a central question for the development of new applications and technological solutions. However, few studies investigate how user habits can improve the performance of Information-Centric Network architectures. This work presents an analysis of the behavioral profiles of music users and how different profiles influence the performance of cache replacement policies. The results of an experimental study using ndnSIM with real traces from several users show that user habits are determining factors in choosing an optimized cache replacement policy. This research also reveals that the popularity distribution of the songs follows an approximation of Benford's Law, and it is possible to differentiate profiles for different users according to the behavior of Benford curve for these accessed songs.*

Resumo. *A exploração de padrões do comportamento humano é tema central e norteador no desenvolvimento de novas aplicações e soluções tecnológicas. No entanto, poucos trabalhos investigam como hábitos de usuários podem melhorar o desempenho de arquiteturas de Redes Centradas na Informação. Este trabalho apresenta uma análise de perfis comportamentais de usuários de música e como diferentes perfis influenciam o desempenho de políticas de substituição de cache. Os resultados de um estudo experimental utilizando o ndnSIM com traces reais de diversos usuários, mostram que os hábitos do usuário são fatores determinantes na escolha de uma política de substituição de cache otimizada. As investigações também revelam que a distribuição de popularidade das músicas segue uma aproximação da Lei de Benford, e é possível diferenciar o perfil dos usuários de acordo com o comportamento da curva de Benford das músicas acessadas.*

1. Introdução

Em anos recentes, o paradigma "ciente do humano" (do inglês, *Human-aware*) [Costa et al. 2018] tem se consolidado no campo das redes de computadores como forma de explorar dados comportamentais dos usuários para apresentar um melhor provimento de serviços de comunicação. O paradigma baseia-se na predição das necessidades e ações conscientes e inconscientes dos humanos para moldar dinamicamente requisitos de desempenho e viabilizar a implementação de serviços que consideram

as informações contextuais e comportamentais do usuário, tais como: mobilidade, personalidade, caráter, humor, interações sociais e rotinas diárias.

Em arquiteturas de Redes Centradas na Informação (do inglês, *Information-Centric Networking* – ICN) [Ioannou and Weber 2016], dados comportamentais dos usuários já tem sido explorados na elaboração de estratégias que visam melhorar o desempenho de tais redes, uma vez que as mesmas se baseiam num modelo de consumo/requisição centralizado no detentor do conteúdo. Na literatura recente, é possível encontrar trabalhos que propõem a identificação de padrões de comportamento de produtores, buscando minimizar efeitos do *handoff* [Lehmann et al. 2016, Araújo et al. 2018]; o uso de políticas de *cache* de acordo com contexto e perfil dos usuários [Ribeiro et al. 2018]; a escolha de localização de *cache* a partir das suas rotinas diárias [da Silva et al. 2016]; e o encaminhamento de interesses com base na previsão do deslocamento dos produtores [Araújo et al. 2018]. Apesar de tais iniciativas apresentarem contribuições em arquiteturas ICNs, as mesmas ainda exploram timidamente os perfis comportamentais dos usuários nas suas soluções.

Perfis comportamentais oferecem novos insumos no desenvolvimento de protocolos e estratégias em ICNs e contribuem para consolidação do conceito de redes centradas na informação cientes do humano. Para isso, é fundamental a identificação e entendimento de perfis dos usuários de conteúdos em tais redes de forma que se possa estabelecer uma associação com os serviços que as ICNs podem oferecer, sobretudo em cenários de mobilidade. Poucos trabalhos na literatura atual se propõem a fazer uma análise de perfis de usuários com tais objetivos. Em geral, são propostas que não discutem como tais perfis influenciam diferentes aspectos das redes de conteúdo, como por exemplo, as políticas de substituição de *cache*. Por tais motivos, este trabalho busca entender como perfis comportamentais de usuários podem influenciar as políticas de *cache* adotadas em arquiteturas ICNs.

Para alcançar os objetivos desta pesquisa, foram elencados dois padrões comportamentais de usuários de um serviço de música a partir de *datasets* reais. Os padrões comportamentais se baseiam em hábitos de ouvir música, estão expressados na forma de perfis e foram avaliados em relação às políticas de substituição de *cache* mais empregadas em ICNs. Em [Pires et al. 2018], iniciamos uma investigação preliminar sobre os impactos de tais perfis no desempenho das políticas de substituição de *cache*. No melhor do nosso conhecimento, foi o primeiro trabalho a caracterizar e avaliar como o comportamento de usuários pode influenciar o desempenho das políticas de substituição de *cache*. Os hábitos foram definidos através da observação empírica de dados reais de um aplicativo de *streaming online* de música, e especificam o nível de repetibilidade de músicas. No presente trabalho, complementamos a análise dos perfis de usuário através da correlação de dados de popularidade das músicas que o usuário costuma acessar, assim como, ampliamos o conjunto de políticas de substituição de *cache* analisadas. Por fim, propomos um modelo de identificação de tais perfis baseado na Lei de Benford [Benford 1938].

Resultados obtidos demonstram que a escolha da política de substituição de *cache* levando em consideração características dos usuários, pode aumentar a taxa de acerto das políticas em aproximadamente 30%. Os resultados também revelam a existência de uma correlação entre o comportamento de repetição de conteúdos, com o tipo do conteúdo acessado pelo indivíduo. Deste modo, este trabalho apresenta as seguintes

contribuições: (i) identificação de perfis de usuários e avaliação das relações com políticas de substituição de *cache* a partir de *datasets* reais; (ii) identificação de um modelo de distribuição de popularidade das músicas que segue uma derivação da Lei de Benford; (iii) modelo de identificação dinâmica de perfis de usuário através da correlação dos perfis com a Lei de Benford; (iv) discussões do uso do paradigma ciente do humano no desenvolvimento de arquiteturas de redes centradas na informação.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta uma visão geral de trabalhos relacionados ao tema de avaliação de políticas de *cache*, e também que incluam características dos usuários no decorrer do trabalho; a Seção 3 discorre sobre o método de análise proposto, com detalhamento dos processos de mineração utilizados para extração de amostras de dados, bem como do estudo experimental para avaliação das políticas, e também introduz a análise de correlação utilizando a Lei de Benford; a Seção 4 apresenta os resultados e discussões das contribuições do trabalho; e, por fim, a Seção 5 apresenta um sumário dos resultados alcançados e as considerações finais.

2. Trabalhos Correlatos

Muitos trabalhos apresentam avaliações de desempenho de políticas de substituição de *cache* em cenários distintos de rede. Em geral, os trabalhos exploram variações de parâmetros que possam influenciar diretamente o comportamento do *cache*, como, por exemplo, topologia da rede, popularidade do conteúdo, e tamanho médio dos arquivos transmitidos, contudo pouco exploram das características dos usuários, conforme discorremos a seguir.

O trabalho apresentado em [Neves et al. 2013] reproduz uma variedade de cenários com a intenção de encontrar a melhor política para *streaming* de mídia em cenários de Rede de Distribuição de Conteúdo (do inglês, *Content-Delivery Network* – CDN). Diferentes combinações de tamanhos de vídeo, modelos de popularidade, número de requisições, tamanho de *cache* e também duração da sessão dos usuários, são utilizadas. Embora o trabalho inclua um parâmetro relacionado ao comportamento dos usuários, por meio do tempo médio que um usuário assiste aos vídeos – classificado como variável discreta, i.e., sessão curta ou longa – a análise não explora essa perspectiva. Em resumo, os autores concluem não existir uma única estratégia de *cache* que atenda de forma ótima todas as combinações de fatores utilizados.

Uma outra aproximação pouco explorada de comportamento do usuário pode ser encontrada em [Rosensweig et al. 2013]. Os autores investigam a ergodicidade de redes de conteúdo e sua relação com as políticas de substituição de *cache*. Em um de seus exemplos de análise dos efeitos de diferentes estados iniciais, o trabalho sucintamente discorre sobre a influência dos padrões de requisição dos usuários, e conclui que pequenas alterações nos padrões de requisição podem gerar um impacto significativo no comportamento das políticas.

Em contrapartida, o trabalho de [Bernardini et al. 2014] envolve diretamente o usuário na definição de políticas de *cache*. Os autores propõem uma nova política de localização de *cache* que observa o número de conexões que um usuário possui em suas redes sociais. Os usuários com muitas conexões são considerados "influentes" e seus conteúdos recebem um tratamento diferenciado na rede, sendo replicados proativamente nos *caches* em direção às "conexões sociais" do usuário. Simulações com dados sintéticos

mostram um melhor desempenho desta nova política em relação à política padrão utilizada em redes ICN.

Em [Fricker et al. 2012] foi utilizado um cenário de topologia hierárquica de dois níveis para verificar o efeito de perfis de aplicação no desempenho das estratégias de *cache*. Os resultados apontam um melhor desempenho ao se armazenar conteúdos VoD (do inglês, *Video on Demand*) nos roteadores de borda e demais conteúdos em roteadores de núcleo. Dessa forma, além do perfil da aplicação, deve-se levar em consideração a localização do *cache* na topologia. Adicionalmente, comparando-se as políticas de substituição de *cache*, LFU obtém melhor desempenho que LRU para conteúdos homogêneos. Este conceito de perfis de aplicação também é observado em [Huo et al. 2016]. Os autores categorizam as aplicações em cinco classes de prioridade, variando de acesso *web* ordinário, até *streaming* multimídia. Os experimentos utilizam diferentes tamanhos de *cache* para cada classe e concluem que a alocação de recursos de *cache* baseada no tipo dos dados pode melhorar o desempenho da rede.

No trabalho de [Sun et al. 2014], os autores investigam o impacto do tráfego de vídeo no desempenho de políticas de localização e políticas de substituição de *cache*, com variações de tamanhos de *cache*. A análise foi realizada em um cenário de grande escala de uso de PPTV (sítio de *streaming* chinês), utilizando roteadores de 80K. O trabalho utilizou dados de 16 mil usuários e de 196 mil vídeos. Dentre os resultados, o trabalho conclui que a melhor combinação de políticas de inclusão e substituição de conteúdos depende do tamanho e da posição da *cache* na topologia.

Deste levantamento observa-se que a maioria dos trabalhos de avaliação de desempenho na literatura não investigam como diferentes hábitos de usuários podem influenciar seus resultados. O presente trabalho investiga além de fatores diretamente associados ao desempenho de políticas de *cache*, como topologia da rede ou tamanho dos arquivos, e inclui o estudo do comportamento dos usuários como um fator capaz de gerar um impacto relevante no desempenho das políticas de substituição de *cache*.

3. Método Proposto de Análise de Perfis de Usuários

Para realizar a análise proposta de usuários e do comportamento de políticas de substituição de *cache*, utilizamos como estudo de caso o serviço de *streaming* de músicas *online* provido pelo `Last.FM`[1]. O processo proposto é apresentado na Figura 1, na qual:

1. Definimos classes de usuários baseadas nos hábitos de repetição de músicas observados de forma empírica em uma base de dados históricos do `Last.FM` (passo 1);
2. Utilizamos processos de mineração de dados para selecionar amostras de dados que representem requisições de músicas de cada classe (passo 2);
3. Com as amostras selecionadas, dividimos o processo em duas vertentes paralelas:
    - A primeira utiliza parte das amostras para a reprodução de um estudo experimental de avaliação de políticas de substituição de *cache* para diferentes tamanhos de *cache*;
    - A segunda vertente utiliza as amostras para uma análise de correlação de informações dos usuários com a popularidade das músicas acessadas.

As etapas do processo estão detalhadas nas subseções seguintes.

[1]Serviço utilizado por usuários de todo o mundo disponível em `http://www.last.fm`.

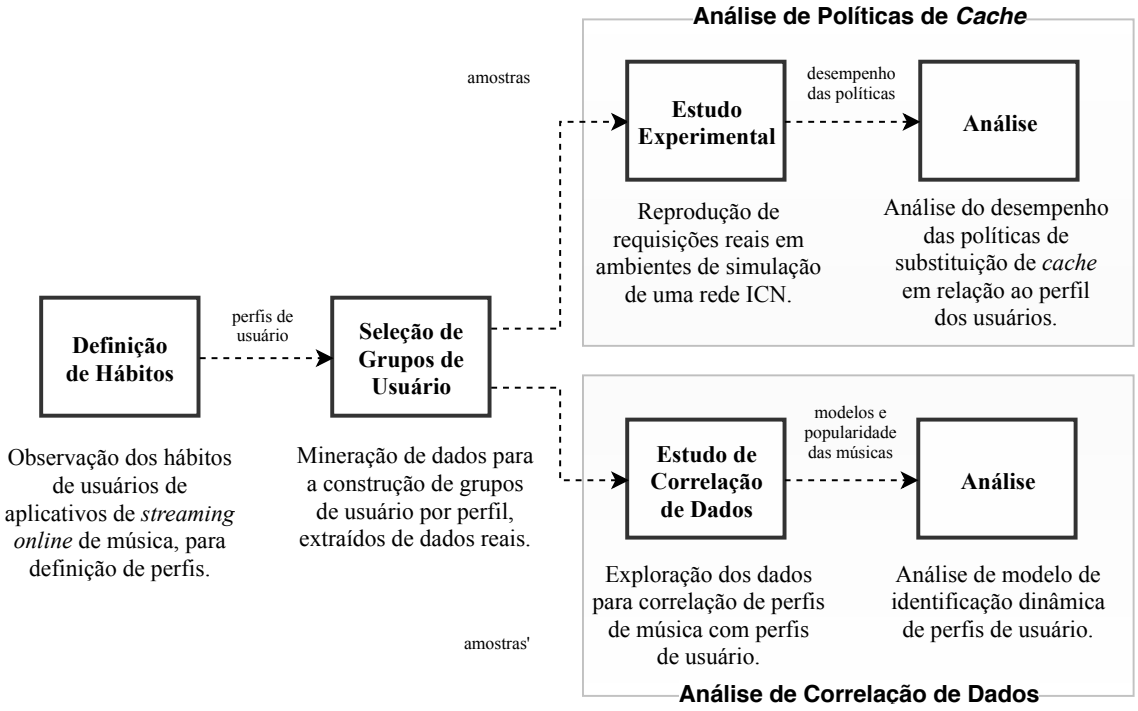


Figura 1. Sequência de atividades empregadas no processo de avaliação.

3.1. Definição de hábitos dos usuários

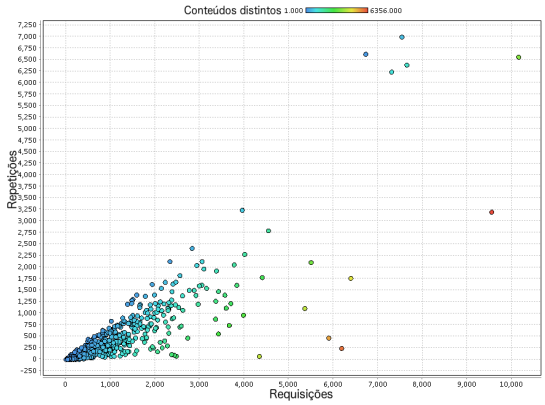
Para identificar diferentes características de usuários, o trabalho define dois perfis de usuários (P) baseados em seus hábitos de escutarem músicas, para construir *clusters* de usuários (C). Os hábitos foram mapeados de acordo com o nível de repetibilidade de músicas e definiu-se os seguintes perfis:

- P_1 – *usuários que frequentemente requisitam as mesmas músicas*: com este perfil inferimos o comportamento de pessoas metódicas e sistemáticas, que usualmente escutam *playlists* de músicas favoritas.
- P_2 – *usuários que geralmente não repetem as músicas*: deduzimos o comportamento de pessoas mais dinâmicas e impulsivas, que quase nunca repetem as músicas que são tocadas.

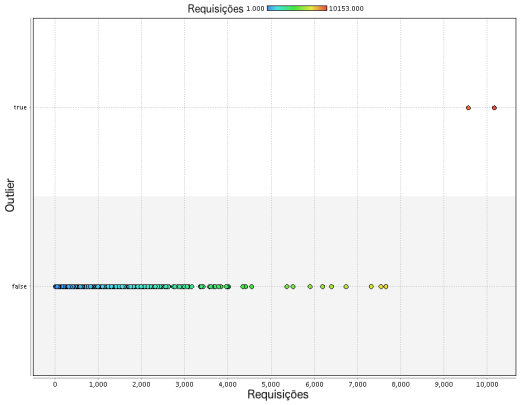
Dessa forma, o experimento se baseia em dois *clusters* de usuários (C_1 e C_2), construídos de acordo com os perfis citados acima (P_1 e P_2) respectivamente, em que $P \longrightarrow C$. Adicionalmente, foi criado um terceiro *cluster* (C_3) composto de usuários escolhidos randomicamente com comportamentos variados, para servir como base de comparação.

3.2. Seleção de grupos de usuário

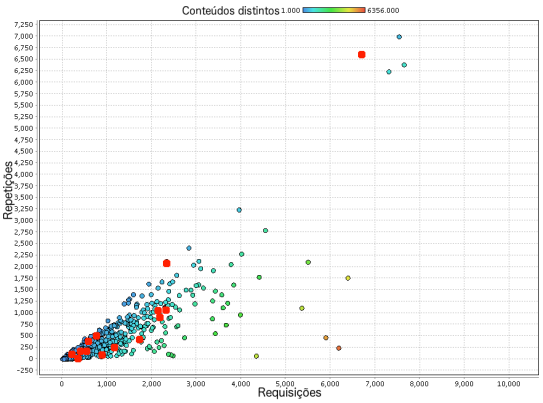
Conduzimos a seleção de *clusters* utilizando processos de mineração de dados para identificar grupos de usuários similares na base do `Last.FM`, de acordo com cada perfil. A base de dados contém todos os registros de requisições de músicas em um período de aproximadamente quatro anos, com mais de 19 milhões de registros. Cada registro descreve a música, o dia e hora da requisição, o artista correspondente, e o usuário que requisitou a música. A preparação dos dados objetivou a geração de três *clusters* que serviram como dados de entrada para as simulações. Os conjuntos (C_1 e C_2) são formados



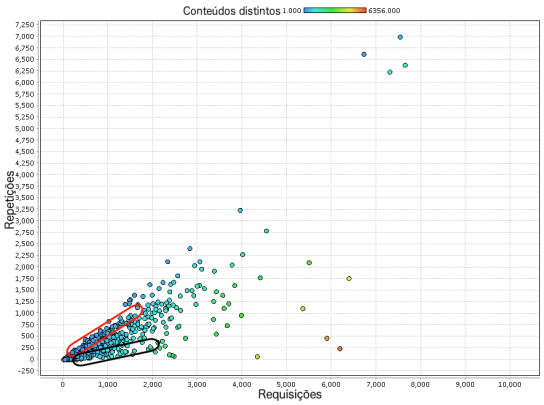
(a) Exemplo de base de dados gerada a partir de uma amostra de período de 1 mês.



(b) Exemplo de identificação de *outliers* baseado no número total de requisições.



(c) Exemplo de seleção randômica de 20 usuários sem distinção de hábitos.



(d) Exemplo de duas áreas para seleção de *clusters* de diferentes perfis de usuários.

Figura 2. Construindo *clusters* de usuários: exemplo de seleção de conjunto de dados.

por usuários que apresentam os comportamentos definidos em P_1 e P_2, respectivamente. Todo o processo se baseou em técnicas de pré-processamento de mineração de dados e foi conduzido da seguinte maneira:

- *Seleção de amostras por tempo*: filtragem de 10 amostras aleatórias de intervalos de tempo em períodos de um mês, totalizando o período de 10 meses. Cada mês possui uma média de 594,483 requisições;
- *Extração de hábitos de usuários*: Para cada amostra de tempo, montamos uma base de dados que mapeia os hábitos dos usuários da amostra de acordo com o total de repetições feitas por cada usuário. A Figura 2(a) ilustra um exemplo de um período de um mês. Cada ponto no gráfico representa um usuário distinto, o eixo *X* representa o total de músicas requisitadas (por usuário), e o eixo *Y* retrata o total de repetições de músicas feitas pelo usuário no mesmo período, conforme definição a seguir: seja M o conjunto de n músicas distintas acessadas por um usuário, e $m_i \in M$, com $0 < i < n + 1$. Seja $Q_i(m_i)$ a quantidade de acessos à música m_i. O total de repetições de um usuário é definido como o somatório de Q_i, para todo $Q_i > 1$;
- *Limpeza dos dados – Identificação e remoção de outliers:* Com o objetivo de

mitigar um possível viés pela aleatoriedade na seleção de usuários, analisamos todos os usuários em relação ao total de requisições, para excluir os potenciais *outliers* pertencentes às amostras. Aplicamos um algoritmo de pré-processamento de dados [Ramaswamy et al. 2000] que detecta *outliers* baseando-se na distância Euclidiana entre cada registro e os seus k vizinhos mais próximos. Uma vez que é difícil definir um limiar que qualifique um usuário como *outlier* devido à subjetividade do contexto, e considerando que idealmente deveríamos realizar interferências mínimas nos dados, nós configuramos os parâmetros da distância Euclidiana com valores mínimos. A Figura 2(b) mostra o resultado de uma amostra de dados classificada com o número de dois *outliers*;

- *Seleção de grupos de dados*: Após a filtragem de *outliers*, realizamos a seleção randômica de usuários. Para cada amostra, três conjuntos de 20 usuários foram selecionados. O primeiro conjunto foi construído sem distinção de hábitos de repetição de músicas, para ser utilizado como base na avaliação de desempenho. Dessa forma, os usuários foram selecionados randomicamente utilizando toda a amostra. A Figura 2(c) ilustra o exemplo de uma seleção desse primeiro conjunto. Os outros dois conjuntos são destinados a refletir hábitos distintos, relacionados aos perfis de usuários mapeados: um conjunto para representar o perfil P_1 (extraídos da área superior em destaque na Figura 2(d)), e outro conjunto para representar o perfil P_2 (extraídos da área inferior em destaque na Figura 2(d)).

3.3. Análise de políticas de *cache*

Após montagem dos conjuntos de usuários, reproduzimos todas as requisições de músicas pertencentes a cada conjunto em ambiente de simulação conduzidos durante um estudo experimental. As amostras foram utilizadas para avaliação das políticas de substituição de *cache* em uma rede NDN (do inglês, *Named Data Networking*). A NDN é uma das propostas mais populares de arquitetura ICN. Dessa forma, esta seção apresenta os detalhes do estudo experimental realizado em ambiente de simulação com dados reais, incluindo o ambiente de simulação com sua devida parametrização e as métricas avaliadas.

3.3.1. Ambiente de simulação

O estudo experimental foi realizado em um ambiente simulado utilizando o ndnSIM. O ndnSIM é um simulador específico para NDN, o qual é desenvolvido com base no NS-3[2], e atualmente é um dos principais simuladores para experimentações em NDN [Mastorakis et al. 2017]. O código do ndnSIM foi alterado para interpretar os dados reais, no qual cada requisição de música corresponde a um pacote de Interesse enviado para a rede. O simulador reproduz a mesma sequência e tempo das requisições de música, exatamente como armazenadas nas amostras.

O cenário avaliado consiste de um agrupamento de usuários (os consumidores dos conteúdos) conectado a um roteador NDN com capacidade de *cache* (configurado com uma política de substituição de *cache*). O roteador, por sua vez, é conectado a um servidor de músicas (o produtor do conteúdo), e intermedeia as solicitações entre os consumidores e o produtor.

[2] `https://www.nsnam.org/`

Cada execução de simulação reproduz as requisições de um dos *clusters* C_1, C_2 ou C_3, separadamente. O processo de comunicação segue o padrão da arquitetura NDN: para cada requisição de um usuário, a música é solicitada ao roteador. Se estiver presente no *cache* do roteador, a taxa de acerto do *cache* é incrementada, e a música é imediatamente encaminhada ao usuário. Caso a música não esteja presente, a requisição é encaminhada ao produtor. A Tabela 1 sumariza os parâmetros utilizados nos experimentos.

Tabela 1. Parâmetros da Simulação.

Parâmetro	**Valor**
Quantidade de usuários (por *cluster*)	20
Quantidade de roteadores	1
Média de conteúdos distintos (*cluster* 1)	3221
Média de conteúdos distintos (*cluster* 2)	10125
Média de conteúdos distintos (*cluster* 3)	7059
Política de *cache*	FIFO, LRU, LFU e LFU-DA[3]
Capacidade de *cache*	5%, 15% e 30% de conteúdos distintos
Taxa de dados	1Mbps
Atraso	10ms
Tempo de simulação	varia de acordo com o *trace*

3.3.2. Métricas de desempenho

A métrica utilizada para avaliar as políticas foi a taxa de acerto de *cache*, pois mede a capacidade do *cache* em resolver as requisições localmente, ao invés de requisitar conteúdos do servidor. Quanto maior a taxa de acerto, mais eficiente é a técnica, uma vez que possibilita economia de uso de largura de banda da rede. A Equação 1 especifica a taxa de acerto, onde *acertoTotal* é o número de requisições satisfeitas pelo roteador, e *totalRequisições* representa o número de requisições recebidas pelo roteador.

A taxa de acerto foi medida para cada amostra de 1 mês separadamente (totalizando 10 meses), e os resultados correspondem a uma média das medições de cada mês.

$$TaxaAcerto = \frac{acertoTotal}{totalRequisições} \times 100 \quad (1)$$

As medições da taxa de acerto obtidas com as combinações dos fatores dos experimentos estão detalhadas na seção de resultados (Seção 4.1), bem como a análise correspondente à influência do perfil dos usuários.

3.4. Análise de correlação de dados

Os hábitos de repetição de música mapeados no trabalho remetem intuitivamente à popularidade do conteúdo. Dessa forma, investigamos também qual a relação dos hábitos dos usuários com a popularidade dos conteúdos que ele acessa. Nessa segunda parte

[3]LFU *with Dynamic Aging*, proposto em [Arlitt et al. 2000].

do estudo, realizamos um novo processamento nas amostras de um mês extraídas anteriormente. Para cada música presente na amostra, calculamos a sua popularidade contabilizando todas as requisições de todos os usuários presentes. Após análise de combinação dos dados, observamos que a distribuição de popularidade das músicas segue uma aproximação da Lei de Benford.

A Lei de Benford é uma distribuição de probabilidade P observada empiricamente em dados numéricos de diversos processos naturais. Segundo Benford, dado um conjunto numérico em base decimal, a proporção do primeiro dígito d de um número qualquer é aproximadamente igual à função de probabilidade:

$$\begin{aligned} P(d) &= \log_{10}(d+1) - \log_{10}(d) \\ &= \log_{10}\left(\frac{d+1}{d}\right) \\ &= \log_{10}\left(1+\frac{1}{d}\right) \end{aligned} \tag{2}$$

onde $\forall d \in \{1, 2, 3, ..., 9\}$. Dessa forma, calculamos a proporção de dígitos d em uma base contendo o valor numérico da popularidade de todas as músicas das amostras. A Equação 2 pode ser facilmente compreendida observando a Figura 3 que ilustra que a quantidade de $P(d)$ é proporcional ao espaço entre d e $d+1$ na escala logarítmica.

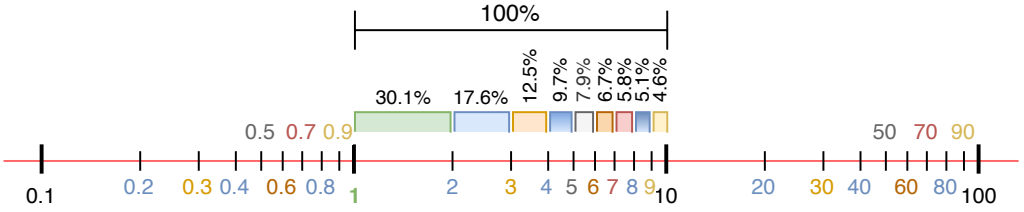


Figura 3. Representação da proporção da ocorrência de d na escala logarítmica.

Também conhecida como Lei do primeiro dígito, possui aplicação prática em diversas áreas do conhecimento e é utilizada para diferentes finalidades. Por exemplo, na computação especificamente, pode ser empregada na detecção de anomalias na rede [Arshadi and Jahangir 2014] ou identificação de mensagens escondidas em imagens [Pérez-González et al. 2007].

A investigação conduziu à evidências de uma correlação entre o perfil dos usuários e a popularidade das músicas, e a um modelo de popularidade capaz de identificar o perfil dos usuários. Os resultados da análise de correlação estão detalhados na Seção 4.2.

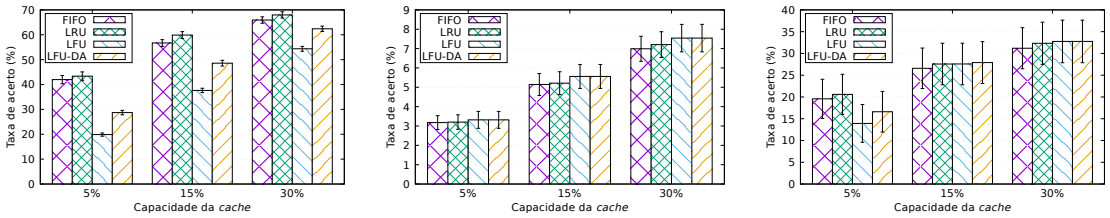
4. Resultados e Discussão

Esta seção apresenta os resultados e a análise dos experimentos realizados. Inicialmente, discute-se como as políticas de *cache* são influenciadas pelos perfis identificados. Em seguida, apresenta-se uma avaliação da distribuição de popularidade das músicas acessadas em relação à lei de Benford. Por fim, a seção é finalizada apresentando um resumo de tais análises.

4.1. Resultados da análise de políticas de *cache*

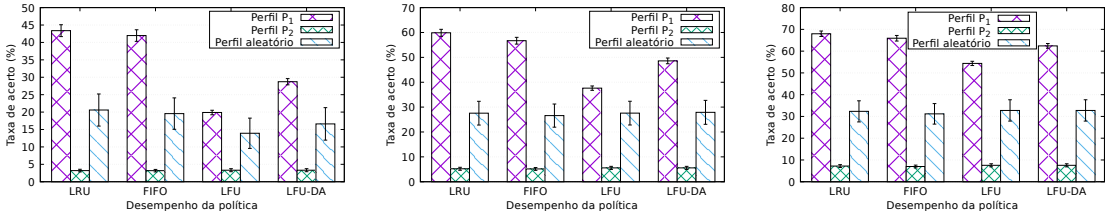
As Figuras 4 e 5 ilustram os resultados obtidos no experimento (média e intervalo de confiança de 95%). A Figura 4 agrupa os resultados por perfil de hábitos dos usuários, e é possível observar as diferentes medições da taxa de acerto de *cache* para cada política avaliada. A Figura 4(a) descreve a taxa de acerto obtida quando o agrupamento é formado por usuários do perfil P_1 (usuários que frequentemente repetem as mesmas músicas). Para esse perfil, as políticas que levam em consideração o tempo, como LRU e FIFO, possuem desempenho consideravelmente maior que as demais, com LRU apresentando o melhor desempenho. O mesmo não acontece para o agrupamento composto de usuários do perfil P_2, como pode ser observado na Figura 4(b). Embora as técnicas LFU e LFU-DA apresentem taxas de acerto ligeiramente maiores, o desempenho de todas as técnicas é considerado estatisticamente equivalente, devido ao intervalo de confiança.

Os resultados revelam que todo o conjunto de políticas sofreu alterações no desempenho apenas variando o perfil do usuário, isso indica que o perfil do usuário pode influenciar no desempenho da política de substituição de *cache*. Aliado à influência dos hábitos dos usuários, o agrupamento permite a obtenção de resultados ótimos de acordo com cada perfil, quando comparado com o grupo de usuários escolhidos aleatoriamente (com hábitos distintos) (Figura 4(c)).



(a) P_1: hábito de muitas repetições. (b) P_2: hábito de poucas repetições. (c) Base comparativa: hábitos aleatórios.

Figura 4. Taxa de acerto de *cache* de acordo com os hábitos dos usuários.



(a) 5% de capacidade de *cache*. (b) 15% de capacidade de *cache*. (c) 30% de capacidade de *cache*.

Figura 5. Taxa de acerto de acordo com a capacidade de armazenamento da *cache*.

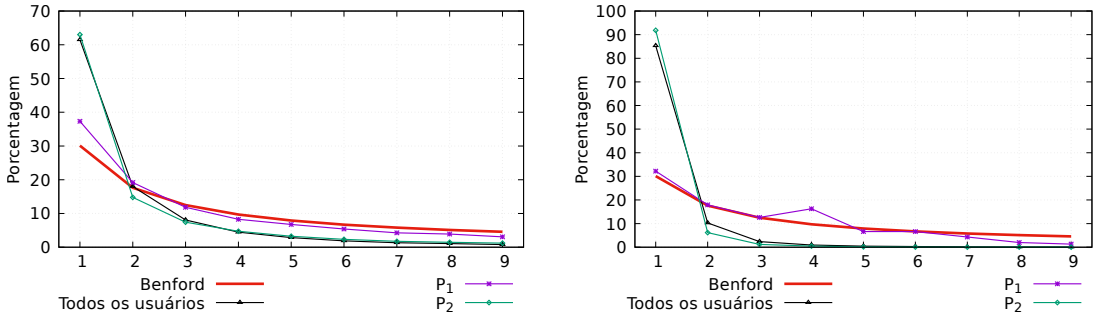
A Figura 5 apresenta os mesmos resultados de taxa de acerto de *cache*, mas sob a perspectiva da capacidade do *cache*. Independente da capacidade do *cache*, as vantagens em agrupar usuários que frequentemente repetem as músicas é sempre maior quando comparada ao grupo base, de hábitos distintos. Essa perspectiva ratifica a ideia de que agrupar usuários com comportamentos similares é a melhor opção para obter altas taxas de acerto

de *cache*, em particular para grupos de usuários que requisitam as mesmas músicas repetidamente. Além disso, os resultados revelam que a escolha da política de substituição de *cache* independe da capacidade do *cache*, mas está diretamente relacionada ao perfil do usuário. Ou seja, a melhor política pode ser diferente para cada perfil, mas para a maioria dos casos, permanece a mesma para tamanhos diferentes de *cache*.

4.2. Resultados de correlação de dados e identificação dinâmica de perfil

A Figura 6 retrata a distribuição dos dígitos conforme a Lei de Benford, e também a distribuição de popularidade de todas as músicas acessadas por todos os usuários presentes em amostras de períodos de um mês (Figura 6(a)) e um dia (Figura 6(b)). Os valores do eixo *X* remetem ao primeiro dígito do valor total de requisições de uma música por todos os usuários (popularidade global da música). É possível observar que a proporção de popularidade das músicas segue uma derivação da Lei de Benford, com alteração significativa para as músicas com popularidade cujo primeiro dígito é 1 (em sua maioria, são as músicas de baixa popularidade).

Para investigar a existência de relação entre o hábito do usuário e a popularidade da música que ele acessa, acrescentamos à análise as músicas acessadas por perfil de usuário (P_1 e P_2). Conforme a Figura 6, os perfis são similares para os conteúdos com "alta popularidade", mas diferem na proporção de acesso para os conteúdos de "baixa popularidade".



(a) Amostra de um mês. (b) Amostra de um dia.

Figura 6. Distribuição dos dígitos numéricos em relação à popularidade dos conteúdos.

Avaliando o comportamento das curvas, é possível diferenciar o hábito do usuário a partir de sua proporção de acesso de conteúdos com "baixa popularidade". Dessa forma, é possível inferir o perfil do usuário observando apenas a popularidade global do conteúdo que ele acessa, sem a necessidade de contabilizar o nível de repetição de cada indivíduo (viabilizando a identificação dinâmica de perfis de usuário sem maiores custos computacionais). Por exemplo, derivando as seguintes regras:

- Usuários com perfil P_1 acessam menos de 40% de conteúdos com popularidades cujo primeiro dígito significativo é 1.
- Usuários com perfil P_2 acessam mais de 60% de conteúdos com popularidades cujo primeiro dígito significativo é 1.

Seguindo essa correlação, a soma dos quadrados residuais (SQR) foi utilizada como métrica para a distinção de perfis de usuário. SQR, também conhecida como soma

dos quadrados dos erros (SQE), é uma medida de variação entre os valores de dados observados e valores de um modelo estimado. Pode ser comparada à distância Euclidiana entre os valores reais e estimados, e reflete a margem de erro de um modelo. Nesta análise, utilizamos a métrica SQR para avaliar a distância de cada perfil de usuário em relação ao modelo de Benford, conforme a equação seguinte:

$$SQR = \sum_{d=1}^{9} (y_d - \hat{y}_d)^2 \tag{3}$$

onde $\hat{y}_d$ é a proporção esperada de frequência das músicas com popularidade com primeiro dígito d, conforme Benford, e y_d é o valor obtido nas amostras de dados reais. De acordo com a Equação 3 e com as medições em intervalos de um dia, tem-se a Figura 7.

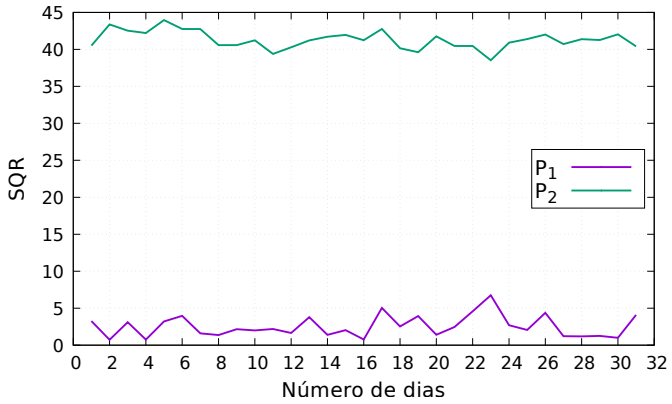


Figura 7. Soma dos quadrados residuais por perfil de usuário.

A Figura 7 contém os valores de SQR medidos por perfil de usuário. Os usuários do perfil P_1 (muitas repetições) possuem medições de SQR notadamente distintas e distantes das medições de SQR dos usuários do perfil P_2 (poucas repetições). A diferença de comportamento entre os perfis é refletida na diferença de distribuição de popularidade das músicas que cada perfil costuma acessar, e também pode ser calculada através do método SQR por um sistema dinâmico de classificação de usuários na rede, que permitiria determinar o perfil do usuário observando apenas a proporção da popularidade do conteúdo que o usuário requisita.

4.3. Síntese dos resultados

Em resumo, os resultados comprovam que os hábitos dos usuários exercem influência no desempenho das políticas de substituição de *cache*, e dessa forma, características comportamentais do usuário devem ser inseridas no processo de decisão de qual política utilizar na rede. O resultado dos experimentos mostra um aumento de aproximadamente 30% na taxa de acerto de *cache* das políticas, quando levado em consideração o hábito de usuário predominante na rede. Comprovados os benefícios em se observar comportamentos do usuário, o estudo vai além e propõem um modelo de identificação dos perfis de usuário como forma de viabilizar a implantação da solução em ambiente operacional. O modelo utiliza uma correlação dos hábitos mapeados, com o padrão de distribuição de popularidade dos conteúdos.

5. Conclusão

O paradigma *ciente do humano* é um tema emergente no desenvolvimento de soluções de redes de computadores. Nesse paradigma, as características do comportamento humano podem ser incorporadas em processos e aplicações, tornando a rede eficientemente adaptada às especificidades dos usuários. Sendo assim, o usuário deixa de ser visto como um elemento genérico da rede, e é introduzido como um novo aspecto de contexto capaz de influenciar o desempenho da rede e as decisões de protocolos e soluções utilizadas. No entanto, existe uma lacuna na literatura sobre a investigação de métodos de identificação dos hábitos dos usuários, assim como avaliações de como os diferentes hábitos podem influenciar o desempenho das redes.

O presente trabalho contribui com uma investigação do conceito de redes centradas na informação cientes do humano, e apresenta uma análise de hábitos de usuários de aplicativo de *streaming* de música *online*. Os resultados mostram que o desempenho das políticas de substituição de *cache* pode ser otimizado quando a política é escolhida de acordo com o hábito do usuário. A análise dos dados também revela que a distribuição de popularidade do conteúdo de músicas segue uma aproximação da Lei de Benford. Essa acordância permite uma geração de tráfego sintético para testes futuros, que possa refletir fielmente o padrão de popularidade do tráfego de conteúdos utilizado. Adicionalmente, a análise conclui que hábitos distintos possuem padrões característicos de acordância com a Lei de Benford e, dessa forma, é possível inferir o hábito de um usuário a partir da observação da popularidade das músicas acessadas.

Agradecimentos

Os autores agradecem o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Referências

Araújo, F. R. C., de Sousa, A. M., and Sampaio, L. N. (2018). Armazenamento oportunista em redes de dados nomeados sem fio como suporte à mobilidade de produtores. In *XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) 2018*, Campos do Jordão, SP.

Arlitt, M., Cherkasova, L., Dilley, J., Friedrich, R., and Jin, T. (2000). Evaluating content management techniques for web proxy caches. *ACM SIGMETRICS Performance Evaluation Review*, 27(4):3–11.

Arshadi, L. and Jahangir, A. H. (2014). Benford's law behavior of internet traffic. *Journal of Network and Computer Applications*, 40:194–205.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American philosophical society*, pages 551–572.

Bernardini, C., Silverston, T., and Festor, O. (2014). Socially-aware caching strategy for content centric networking. In *Networking Conference, 2014 IFIP*, pages 1–9. IEEE.

Costa, R. L., Sampaio, L. N., Ziviani, A., and Viana, A. (2018). Humanos no ciclo de comunicação: facilitadores das redes de próxima geração. In *Livro de Minicursos do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) 2018*, Campos do Jordão, SP.

da Silva, V. B. C., Campista, M. E. M., and Costa, L. H. M. K. (2016). Trac: A trajectory-aware content distribution strategy for vehicular networks. *Vehicular Communications*, 5:18 – 34.

Fricker, C., Robert, P., Roberts, J., and Sbihi, N. (2012). Impact of traffic mix on caching performance in a content-centric network. In *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*, pages 310–315. IEEE.

Huo, R., Xie, R., Zhang, H., Huang, T., and Liu, Y. (2016). What to cache: differentiated caching resource allocation and management in information-centric networking. *China Communications*, 13(12):261–276.

Ioannou, A. and Weber, S. (2016). A survey of caching policies and forwarding mechanisms in information-centric networking. *IEEE Communications Surveys & Tutorials*, 18(4):2847–2886.

Lehmann, M. B., Barcellos, M. P., and Mauthe, A. (2016). Providing producer mobility support in NDN through proactive data replication. In *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, pages 383–391. IEEE.

Mastorakis, S., Afanasyev, A., and Zhang, L. (2017). On the evolution of ndnsim: An open-source simulator for ndn experimentation. *SIGCOMM Comput. Commun. Rev.*, 47(3):19–33.

Neves, M., Rodrigues, M., Azevêdo, E., Sadok, D., Callado, A., Moreira, J., and Souza, V. (2013). Selecting the most suited cache strategy for specific streaming media workloads. In *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on*, pages 792–795. IEEE.

Pérez-González, F., Heileman, G. L., and Abdallah, C. T. (2007). A generalization of benford's law and its application to images. In *Control Conference (ECC), 2007 European*, pages 3613–3619. IEEE.

Pires, S. S., Ribeiro, A. V., de Sousa, A. M., Freitas, A. E. S., and Sampaio, L. N. (2018). On evaluating the influence of user's music listening habits on cache replacement policies. In *IEEE Symposium on Computers and Communications (ISCC)*, pages 930–933. IEEE.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In *ACM Sigmod Record*, volume 29, pages 427–438. ACM.

Ribeiro, A. V., Sampaio, L. N., and Ziviani, A. (2018). Affinity-based user clustering for efficient edge caching in content-centric cellular networks. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00474–00479.

Rosensweig, E. J., Menasché, D. S., and Kurose, J. (2013). On the steady-state of cache networks. In *INFOCOM*, pages 863–871.

Sun, Y., Fayaz, S. K., Guo, Y., Sekar, V., Jin, Y., Kaafar, M. A., and Uhlig, S. (2014). Trace-driven analysis of icn caching algorithms on video-on-demand workloads. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pages 363–376. ACM.