**Universidade do Estado do Rio de Janeiro**
Centro de Tecnologia e Ciências
Escola Superior de Desenho Industrial

Augusto de Freitas Lohmann

# Sistemas de recomendação em sites de notícias: contextualização como proposta para o engajamento de usuários

Rio de Janeiro
2018

Augusto de Freitas Lohmann

# Sistemas de recomendação em sites de notícias: contextualização como proposta para o engajamento de usuários

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-graduação em Design, da Universidade do Estado do Rio de Janeiro. Área de concentração: Design e Tecnologia.

Orientador: Prof. Dr. André Soares Monat

Rio de Janeiro

2018

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/BIBLIOTECA CTC/G

L833s Lohmann, Augusto de Freitas.

Sistemas de recomendação em sites de noticias: contextualização como proposta para o engajamento de usuários / Augusto de Freitas Lohmann. - 2018.
147 f. : il.

Orientador: Prof. Dr. André Soares Monat.

Tese (Doutorado). Universidade do Estado do Rio de Janeiro, Escola Superior de Desenho Industrial.

1. Sistemas de recomendação - Teses. 2. Algoritmo - Teses. 3. Sites de notícias – Teses. 4. Jornalismo eletrônico - Teses. 5. Sites da web - Teses. I. Monat, André Soares. II. Universidade do Estado do Rio de Janeiro. Escola Superior de Desenho Industrial. III. Título.

CDU 7.048

Bibliotecária: Marianna Lopes Bezerra CRB7/6386

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta tese, desde que citada a fonte.

_____ _____

Assinatura Data

Augusto de Freitas Lohmann

**Sistemas de recomendação em sites de notícias: contextualização como proposta para engajamento de usuários**

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-graduação em Design, da Universidade do Estado do Rio de Janeiro. Área de concentração: Design e Tecnologia.

Aprovada em 29 de junho de 2018

Banca Examinadora:

_____

Prof. Dr. André Soares Monat (Orientador)

Escola Superior de Desenho Industrial - UERJ

_____

Prof. Dr. André Ribeiro

Escola Superior de Desenho Industrial – UERJ

_____

Prof. Dr. Edson Rufino

Universidade Federal do Espírito Santo – UFES

_____

Prof[a]. Dr[a]. Graciana Simoni Fischer

UniCarioca - Centro Universitário Carioca

_____

Prof. Dr. Sydney Freitas

Escola Superior de Desenho Industrial – UERJ

Rio de Janeiro

2018

## AGRADECIMENTOS

Aos meus pais. Obrigado pelo amor, carinho, educação, formação e incentivo.

Agradeço ao meu orientador, professor doutor André Monat, pelo estímulo, atenção, orientação e paciência ao longo de todo o processo.

Aos professores doutores pelo apoio, acompanhamento e críticas extremamente valiosas e orientações para o desenvolvimento do trabalho.

À ESDI na forma de seu corpo docente e funcionários, cuja atenção foi de grande valia.

Aos professores da ESDI pela atenção e incentivo durante todo o processo.

À minha família e amigos, pela torcida e por acompanhar de perto cada etapa do meu desenvolvimento.

Obrigado por tudo.

## RESUMO

LOHMANN, Augusto de Freitas. *Sistemas de recomendação em sites de notícias:* contextualização como proposta para o engajamento de usuários. 2018. 147f. Tese (Doutorado em Design) – Escola Superior de Desenho Industrial, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

Em um contexto de aumento exponencial na produção de conteúdo e de mudanças no modelo de consumo de notícias, os sistemas de recomendação de conteúdo apresentam-se como elementos que auxiliam na promoção do engajamento de usuários em sites de notícias. Entretanto, observa-se especificamente no âmbito dos sites de notícias o baixo desempenho dos sistemas de recomendação de conteúdo, em especial as reduzidas taxas de engajamento identificadas em acessos oriundos de fontes externas (redes sociais, buscadores e outros que já constituem a maior parte dos acessos).

Dessa forma, tem-se por objeto da pesquisa a recomendação de notícias nas páginas de conteúdo dos sites de notícias, com ênfase nos sites brasileiros. Nessa tese é apresentada a análise teórica e prática a respeito da utilização de sistemas de recomendação automatizados no contexto da notícia, de maneira a validar quais aspectos desses sistemas exercem influência sobre os índices de engajamento.

Para isso, são propostas três hipóteses a serem validadas a partir do resultado de pesquisas e experimentos práticos, a saber: 1) Mudanças no design de páginas de conteúdo (incluindo layout, usabilidade e/ou critérios dos elementos de recomendação de conteúdo) impactam nas métricas de engajamento do usuário com o site de notícias; 2) As variáveis “página de destino” e “origem do tráfego” influenciam de maneira relevante as métricas de engajamento dos sites de notícias; 3) O contexto interfere na probabilidade de um determinado usuário consumir uma notícia (Independente do nível de interesse por este conteúdo).

Além disso, são identificados problemas e fatores que trazem complexidade para a utilização dos sistemas de recomendação de conteúdo automatizados no contexto específico do jornalismo e das notícias. A partir de análises de métricas e pesquisas acerca do estado atual da recomendação nos sites de notícias brasileiros, é proposta a categorização dos critérios utilizados na recomendação de conteúdo destes sites, bem como cenários conceituais que apontem para soluções que promovam melhores resultados – em especial na questão do engajamento dos usuários - na utilização de sistemas de recomendação em sites de notícias.



Palavras-chave: Sistemas de Recomendação. Algoritmo. Engajamento. Contexto. Site de Notícias. Design. Conteúdo.

## ABSTRACT

LOHMANN, Augusto de Freitas. *Recommendation systems in news sites: contextualization as proposal for the engagement of users*. 2018. 147f. Tese (Doutorado em Design) – Escola Superior de Desenho Industrial, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

In a context of exponential increase in content production and changes in the news consumption model, content recommendation systems are seen as elements that help in promoting the engagement of users in news sites. However, the low performance of content recommendation systems, especially the low engagement rates identified in accesses from external sources (social networks, search engines and others that already make up most of the accesses).

Thus, the purpose of the research is the recommendation of news on the content pages of news sites, with emphasis on Brazilian sites. This thesis presents the theoretical and practical analysis regarding the use of automated recommendation systems in the news context, in order to validate which aspects of these systems influence the engagement indexes.

For this, three hypotheses are proposed to be validated from the results of research and practical experiments, namely: 1) Changes in the design of content pages (including layout, usability and / or criteria of content recommendation elements) impact on the metrics of user engagement with the news site; 2) The "landing page" and "traffic source" variables have a relevant influence on the engagement metrics of news sites; 3) The context interferes with the probability of a particular user consuming a news item (regardless of the level of interest in this content).

In addition, problems and factors that bring complexity to the use of automated content recommendation systems in the specific context of journalism and news are identified. Based on analyzes of metrics and research on the current scenario of the recommendation on Brazilian news sites, it is proposed to categorize the criteria used in recommending the content of these sites, as well as conceptual scenarios that point to solutions that promote better results - especially in the issue of user engagement - in the use of recommendation systems on news sites.



Keywords: Recommendation Systems. Algorithm. Engagement. Context. News Site. Design. Content.

**SUMÁRIO**

**INTRODUÇÃO** ..... 10
1 **CONSUMO DE NOTÍCIAS NOS MEIOS DIGITAIS** ..... 16
1.1 **Contextualização e cenário atual do consumo de notícias** ..... 16
1.1.1 Analógico x Digital ..... 16
1.1.2 O declínio das homepages e o conteúdo como 'página de destino' ..... 19
1.1.3 Queda na audiência e no engajamento dos usuários ..... 23
1.2 **Sistemas de Recomendação** ..... 29
1.2.1 Sistemas de recomendação automatizados ..... 30
1.2.1.1 Desafios na utilização de sistemas de recomendação automatizados ..... 39
1.2.1.2 Recomendação de conteúdo em sites de notícias ..... 43
1.2.3 Sistemas de recomendação contextual ..... 50
2 **METODOLOGIA DE PESQUISA** ..... 54
2.1 **Levantamento de dados para categorização da recomendação de conteúdo nos sites de notícias brasileiros** ..... 54
2.2 **Análise de métricas de navegação: site O Globo** ..... 55
2.3 **Pesquisa com usuários: questionário** ..... 59
3 **LEVANTAMENTO DO ESTADO ATUAL DA RECOMENDAÇÃO DE CONTEÚDO EM SITES DE NOTÍCIAS BRASILEIROS E PROPOSTA DE CATEGORIZAÇÃO** .60
3.1 **Estado atual da recomendação de conteúdo em sites de notícias brasileiros.** ..... 62
3.1.1 G1 ..... 63
3.1.2 O Globo ..... 64
3.1.3 Terra ..... 66
3.1.4 UOL ..... 68
3.1.5 Folha de São Paulo ..... 70
3.1.6 R7 ..... 72
3.1.7 Estadão ..... 74
3.1.8 Yahoo Notícias ..... 75
3.2 **Proposta de categorização para a recomendação de conteúdo nos sites de notícias brasileiros** ..... 77
3.2.1 Editorial ..... 80

3.2.2 Patrocinado ..... 82
3.2.3 Assunto ..... 83
3.2.4 Autoria ..... 85
3.2.5 Cronologia ..... 86
3.2.6 Geolocalização ..... 88
3.2.7 Tipo de conteúdo ..... 90
3.2.8 Pessoal ..... 91
3.2.9 Relacionamento Social ..... 93
3.2.10 Emocional ..... 95
3.2.11 Popularidade ..... 96
3.2.12 Considerações ..... 98
4 **APLICAÇÃO DAS TÉCNICAS DE PESQUISA PARA VALIDAÇÃO DE HIPÓTESES** ..... 98
4.1 **Análise de Métricas do site do jornal O Globo** ..... 99
4.1.1 Comparativo entre páginas de matéria (antiga e nova) do site d'O GLOBO ..... 99
4.1.2 Comparativo entre as métricas de engajamento nos acessos iniciados pelas homepages *versus* os acessos iniciados pelas páginas de matéria ..... 106
4.1.3 Análise do impacto no engajamento de usuários em sites de notícias a partir das variáveis "Origem do Tráfego" e "Página de destino" ..... 109
4.2 **Pesquisa com Usuários** ..... 117
4.2.1 Questionário de avaliação da influência do contexto no consumo de notícias ..... 118
4.2.2 Resultados consolidados do questionário ..... 119
4.2.3 Análise individual das respostas ..... 127
5 **PROPOSIÇÃO TEÓRICA DE CENÁRIOS PARA A RECOMENDAÇÃO CONTEXTUAL DE CONTEÚDO NOTICIOSO** ..... 129
**CONSIDERAÇÕES FINAIS** ..... 137
**REFERÊNCIAS** ..... 142

## INTRODUÇÃO

Ao longo dos anos, o modelo de oferta de conteúdo baseado em uma curadoria editorial vigente nos modelos de jornalismo impresso e nos primórdios do jornalismo digital vêm perdendo espaço nos sites de notícias em decorrência do crescente aumento na produção de conteúdo noticioso nos meios digitais.

> "Publicadores sempre fizeram a curadoria de conteúdo para suas audiências, mas o volume e a variedade de conteúdo se expandiu, de tal forma que apresentar somente o que é relevante para seus leitores tem se tornado cada vez mais difícil". (HENDRIX, 2014)

Nesse processo, cabia ao corpo editorial atuar como "curadores" da notícia, ou seja, aqueles que eram responsáveis por selecionar o que deveria ou não receber destaque e ser apresentado para os usuários.

Em paralelo ao aumento na produção de conteúdo, o próprio modelo de consumo de notícias passou por profundas modificações em função das novas tecnologias, em especial da popularização do uso dos buscadores e das redes sociais como canais de acesso para a informação noticiosa. Nesse sentido, cada vez menos usuários buscam conteúdo ativamente acessando um site específico de notícias; na maior parte dos casos, a notícia é que "chega" até o usuário, a partir de resultados de buscas, ou de compartilhamentos e recomendações de amigos em suas redes sociais.

Estes acessos, agora iniciados por fontes externas, em geral se dão diretamente a partir de uma página de conteúdo específico do site de notícias, e não pelas "homepages", ou seja, a página inicial do site. Neste caso, a página de conteúdo assume a função de página inicial pela qual o usuário começa a sua navegação pelo site de notícias. Por definição, qualquer página por onde o usuário inicia a sua navegação em um site é chamada tecnicamente de "landing page" (em inglês) ou em sua versão em português "página de destino", termos que a partir de agora serão utilizados nesse contexto ao longo da tese.

Dessa forma, em um cenário onde o usuário não passa necessariamente pelas homepages em sua navegação, cabe às páginas de conteúdo a função de recomendar novas notícias que possam motivar os usuários a permanecerem naquele site.

Entretanto, como consequência direta da mudança no modelo de consumo de notícias em produtos digitais, os sites de notícias vêm apresentando quedas em suas receitas e também em suas métricas de acesso, em especial das chamadas “métricas de engajamento”, referentes ao tempo de navegação e a quantidade de interações de cada usuário de um mesmo site.

Nesse sentido, os elementos de recomendação de notícias das páginas de conteúdo adquiriram papel estratégico na função de manter os usuários por mais tempo ativos e engajados com sites de notícias. Em geral, é a partir das notícias sugeridas pelo site que o usuário decidirá por manter-se no site ou abandonar a navegação.

Em resumo, tem-se, portanto, um cenário onde cada vez mais os acessos a sites de notícias são originados por fontes externas, tendo como páginas de destino um conteúdo em específico. Tais acessos tem resultado em piora nas métricas de engajamento de sites de notícias em todo o mundo, o que vem gerando menor retorno financeiro para as empresas de jornalismo.

Nesse cenário, as mudanças no modelo de consumo de conteúdo aliada a queda no engajamento dos sites de notícias criaram um ambiente propício para a ampliação do investimento por parte das empresas de notícias em sistemas automatizados de recomendação de conteúdo.

Tais sistemas são baseados em lógicas algorítmicas que, a partir de um conjunto de dados e regras, tentam “antecipar” os interesses de cada usuário, recomendando conteúdos que sejam avaliados como de maior probabilidade de atrair o interesse e motivar a interação por parte do usuário.

Entretanto, apesar de diversas empresas dos mais variados domínios virem fazendo uso de sistemas de recomendação com relativo sucesso (HENDRIX 2014), no âmbito dos sites de notícias esses sistemas ainda não apresentam bons resultados. Nesse sentido, cada domínio ou área de conhecimento possui características e desafios específicos para a implementação de tais sistemas, e as soluções propostas para a recomendação de conteúdo devem, portanto, avançar no sentido de atender a essas especificidades.

> Empresas como a Amazon, Spotify e Netflix usam ‘machine learning’ e outros algoritmos para recomendar seleções de conteúdo que atingem com grande precisão as preferências de cada usuário. No Netflix, por exemplo, 75% da atividade de seus usuários se dá a partir de recomendações. O equivalente a isso no âmbito das notícias ainda não acontece por 2 razões: Em primeiro lugar, um sistema de

> recomendação demanda uma grande quantidade de dados sobre as preferências do usuário. Ao contrário de livros, música e filmes, leitores consomem notícias de múltiplas fontes, então os dados relativos ao interesse de uma pessoa perpassam por diferentes sites. Segundo, a construção de sistemas de recomendação exige um esforço massivo, e requer cientistas de dados altamente qualificados. Poucas organizações de notícias fizeram esse investimento até o momento. (HENDRIX, 2014)

Dessa forma, para o avanço e consequente melhoria nos resultados da utilização de sistemas de recomendação no âmbito das notícias, é fundamental se aprofundar no que Ricci (2014) chama de “desafios específicos de domínio”. Para o autor, em diferentes domínios, o contexto em que cada recomendação acontece é fator chave para o desenvolvimento de sistemas automatizados mais assertivos, ou como ele denomina, “sistemas de recomendação contextual”. Entretanto, a inclusão de aspectos contextuais nos sitemas de recomendação ainda não é um padrão, em especial nos sites de notícias brasileiros onde muito pouco do contexto de cada usuário é levado em conta no momento da recomendação de conteúdo.

Dessa maneira, entende-se que por ser uma área de estudos bastante nova, pesquisas a respeito dos sistemas de recomendação no âmbito da notícia se justificam por ser este um campo de interesse elevado e aberto a constantes contribuições. Ao mesmo tempo que resolvem uma série de questões e ajudam os usuários a lidar com a sobrecarga de informações ao fornecer recomendações, conteúdo e serviços em níveis cada vez mais personalizados, os sistemas de recomendação automatizados ainda não foram capazes de promover nos acessos iniciados por origens externas o mesmo nível de engajamento dos acessos iniciados diretamente pelas homepages dos sites de noticias. Tal fenômeno pode ser identificado em sites de notícias de todo o mundo, e nos sites brasileiros isso não é diferente. É necessário, portanto, trabalhar em caminhos para a melhoria desses sistemas de recomendação no domínio em questão.

Apresenta-se, portanto, o **problema de pesquisa:** como os sistemas de recomendação de conteúdo podem reverter as baixas taxas de engajamento em sites de notícias, em especial em acessos oriundos de fontes externas (que atualmente representam a maioria). Para esta tese, o escopo dos sites avaliados será limitado ao universo dos sites de notícias brasileiros, com especial atenção aos sites mais representativos e de maior acesso. Apesar de entender que este problema se faz presente nos sites e produtos digitais de notícias em âmbito

mundial, a quantidade e a variedade de produtos existentes torna necessária a escolha de um recorte específico para análise mais aprofundada da questão.

O **objeto da pesquisa** é, portanto, a recomendação de notícias nas páginas de conteúdo em sites brasileiros. Dentre possíveis aspectos do Design acerca da recomendação de notícias (critérios para a recomendação, usabilidade ou mesmo aspectos visuais da interface), o enfoque desta abordagem se dará essencialmente sobre os critérios adotados pelos sistemas automatizados de recomendação de conteúdo. Será estudado o design da recomendação de conteúdo pela ótica dos critérios que definem quais conteúdos adicionais são recomendados para cada usuário no momento do acesso à uma página de notícia. Os demais aspectos do Design citados irão permear este trabalho, mas não serão objetos de estudo.

Ainda em relação aos critérios para a recomendação automatizada de conteúdo de notícias, será estudado como o Design pode fornecer ferramentas e abordagens para auxiliar na solução de problemas típicos dos sistemas de recomendação no âmbito específico das notícias, visando a melhoria nos resultados alcançados por estes sistemas.

Como **metodologia de pesquisa**, conforme Lakatos (2003), após levantamento bibliográfico e revisão sistemática acerca do tema, foram formuladas as seguintes **hipóteses** a respeito do problema de pesquisa identificado:

- **Hipótese 1**: Mudanças no design de páginas de conteúdo (incluindo layout, usabilidade e/ou critérios dos elementos de recomendação de conteúdo) impactam nas métricas de engajamento do usuário com o site de notícias;
- **Hipótese 2**: As variáveis “página de destino” (termo utilizado para a página de entrada do usuário em um site) e “origem do tráfego” (de onde o usuário veio antes do acesso ao site) influenciam de maneira relevante as métricas de engajamento de sites de notícias. Acessos iniciados diretamente pelo conteúdo como “página de destino” tendem a ter métricas de engajamento piores em comparação a acessos iniciados por homepages como “página de destino”. Da mesma maneira, diferentes “origens de tráfego” impactariam no engajamento dos usuários de sites de notícias;
- **Hipótese 3**: O contexto (pessoal, local ou mundial) interfere na probabilidade de um determinado usuário consumir uma notícia (Independente do nível de interesse por este mesmo conteúdo).

Para a **validação das hipóteses**, como procedimentos metodológicos foram analisadas métricas de utilização de um site de notícias brasileiro específico (O

Globo). Além disso, será aplicada uma pesquisa estruturada em formato de questionário com usuários de sites de notícias.

Como **objetivo geral** da tese está a proposição teórica de cenários para a recomendação de conteúdo em sites de notícias brasileiros a partir da formulação de critérios que incluam questões do contexto de cada usuário no momento do consumo da notícia.

Como **objetivos específicos**, pretende-se identificar e compreender fatores responsáveis pelos resultados insatisfatórios apresentados na utilização de sistemas de recomendação de conteúdo automatizados nos sites de notícias do Brasil. Além disso, objetiva-se fazer um levantamento acerca do estado atual dos sistemas de recomendação dos sites de notícias brasileiros. O terceiro objetivo específico é a compreensão acerca do estado atual da recomendação, apresentando uma proposta de categorização para os critérios da recomendação de conteúdo a partir da identificação dos padrões utilizados nos sites de notícias do Brasil.

Dessa maneira, após este capítulo introdutório, o capítulo 1 traz a fundamentação teórica acerca das temáticas do consumo de notícias nos meios digitais e dos sistemas de recomendação (em geral e especificamente no âmbito do jornalismo).

No capítulo 2, são apresentados e detalhados os procedimentos metodológicos, a saber: levantamento de dados e categorização da recomendação de conteúdo; análise de métricas de navegação e pesquisa com usuários.

O capítulo 3 traz o resultado do levantamento acerca do estado atual dos sites de notícias brasileiros no que diz respeito a utilização de sistemas de recomendação nas páginas de conteúdo. Nesse levantamento, foram identificados os elementos de recomendação de conteúdo de alguns dos principais sites de notícias do Brasil (em audiência e representatividade), bem como os critérios que definem seu funcionamento, no intuito de reconhecer padrões a fim de propor uma categorização da recomendação de conteúdo que possa ser aplicada especificamente aos sites de notícias.

No capítulo 4 são detalhados e apresentados os procedimentos metodológicos propostos para a validação das hipóteses, a saber: análise das métricas de navegação e pesquisa com usuários. Para isso, são apresentados os resultados obtidos a partir de análises de métricas do site de notícias O Globo. Apresenta-se também o resultado da pesquisa estruturada aplicada a usuários

acerca da influência das questões contextuais em seu comportamento de consumo de notícias.

No capítulo 5, a partir da validação das hipóteses e dos resultados obtidos nas etapas de pesquisa, é feita a proposição teórica de cenários onde a recomendação de conteúdo leve em conta critérios contextuais de cada usuário, aumentando o grau de assertividade da recomendação e consequentemente o engajamento em sites de notícias.

O capítulo 6 apresenta as conclusões obtidas a partir desta pesquisa e as indicações para futuros estudos acerca da temática da recomendação de conteúdo. Ao final são listadas as referências utilizadas para o embasamento teórico da abordagem e dos procedimentos metodológicos utilizados ao longo desta pesquisa.

# 1 **CONSUMO DE NOTÍCIAS NOS MEIOS DIGITAIS**

## 1.1 **Contextualização e cenário atual do consumo de notícias**

### 1.1.1 Analógico x Digital

Aproximadamente três décadas após o início da transposição do conteúdo dos jornais impressos para o formato digital, os sites e aplicativos de notícias oferecem uma maior variedade de fluxos de navegação e narrativas, se comparado ao modelo de consumo de conteúdo convencional do jornalismo impresso. Ao invés de um único fluxo de consumo da informação (a leitura sequencial do conteúdo, baseada em um critério editorialmente definido), nas versões digitais o usuário passa a ter uma maior oferta de possibilidades de fluxos a partir de qualquer ponto do consumo do conteúdo informacional.

Tal modelo de estruturação da informação e de navegação descentralizada é característico da internet como um todo (e dos sites de notícias em especial). Segundo Gonçalves (2009), a internet funciona como uma rede em que cada conteúdo pode ser acessado a partir de qualquer ponto daquela estrutura, e dali pode-se atingir qualquer outro ponto sem que necessariamente seja seguido um determinado caminho pré-estabelecido.

> “A rede pode, assim, ser vista como um rizoma, constantemente estratificada, à medida que a ela se acrescentam conteúdos de toda sorte, fazendo que não possa ser classificada dentro de conceituações dicotômicas, dualistas, ou mesmo moralistas, levando-se em consideração a multiplicidade dos conteúdos que a compõem. À cada tentativa de corte ou ruptura, por mais que se queira atingir uma conexão essencial, a rede se recria, se reconfigura e permanece inalterada”. (GONÇALVES, 2009)

Da mesma forma, Leão (1999) associa essa multiplicidade de fluxos e caminhos possíveis na navegação na internet com a metáfora de um “labirinto”, dado os aspectos de complexidade envolvidos na experiência hipertextual:

> "O processo de desenvolvimento de um sistema hipermidiático envolve uma série de questões que se avolumam e fazem emergir uma complexidade. ... apesar de o termo labirinto estar, de um modo geral, associado à ideia de confusão, de estar

> perdido e de erro, quando é visto como metáfora da complexidade, seu sentido se expande." (LEÃO, 1999)

Nesse sentido, constata-se o surgimento de um novo modelo de consumo de conteúdo no ambiente digital. Tal modelo se faz presente igualmente no âmbito das notícias, provocando assim uma necessidade de ruptura no modelo vigente de organização e oferta de conteúdo. No jornalismo impresso, o conteúdo é organizado de uma única maneira (definida pela equipe editorial do jornal), limitando a participação do leitor no processo de construção do fluxo narrativo e oferecendo uma estrutura estática e sequencial de conteúdos a serem consumidos. Já no meio digital as possibilidades trazidas pela lógica das estruturas em rede são múltiplas e dinâmicas, permitindo a cada usuário estabelecer e exercer seu próprio modelo de fluxo de consumo do conteúdo noticioso.

Entretanto, ao mesmo tempo em que encontra mais liberdade, o usuário vê diminuída sua percepção de finitude sobre uma hipotética totalidade do conteúdo e sobre as diferentes formas de acessá-lo e consumi-lo. Enquanto no meio físico do jornal impresso o modelo de edições periódicas traz a noção de um conteúdo finito, da completude da absorção das informações selecionadas pelo corpo editorial do jornal, no ambiente virtual/digital ocorre a ruptura desse modelo de edições fechadas e limitadas. O aumento na produção e na oferta de conteúdo de maneira praticamente ilimitada, em paralelo aos inúmeros fluxos de consumo possíveis, diminui nos usuários a percepção da finitude do conteúdo.

Este excesso de oferta de conteúdo e de caminhos possíveis, se tratados de maneira desordenada, tende a diminuir o interesse do usuário no produto noticioso, conforme resultado apresentado em pesquisa da Associated Press (2008):

> “Os participantes desta pesquisa mostraram sinais de fadiga de notícias; Ou seja, eles pareciam debilitados pela sobrecarga de informações e experiências de consumo de notícias insatisfatórias. Muitos estavam tão sobrecarregados e inundados pelas notícias que eles simplesmente não sabiam o que fazer. Os participantes com fadiga de notícias até tentaram verificar todas as notícias, mas regularmente e repetidamente ficavam insatisfeitos. E quanto mais eles se sentiam sobrecarregados ou insatisfeitos, menos esforço eles estavam dispostos a colocar no consumo de novos conteúdos”. (Associated Press, 2008)

Lafrance (2012) corrobora a ideia de uma sobrecarga de informação que incomoda e provoca cansaço nos usuários:

> “Os jornais impressos tinham a ideia certa: dar aos seus leitores uma sensação de conclusão quando eles terminavam de paginar uma edição. As pessoas têm essa sobrecarga de informações; elas estão verificando muitas fontes. Estão um pouco cansadas de ter que passar por fluxos infinitos de informações, nunca tendo uma sensação de conclusão da tarefa."(LAFRANCE, 2012)

Tal cenário remete a um caráter de aleatoriedade e de infinitude de possibilidades, enquadrando de maneira eficiente o funcionamento da internet como rede sob um olhar técnico e sistêmico da organização do conteúdo, mas pouco elucidativo quando se pensa pelo olhar da oferta desse mesmo conteúdo ao usuário final. Para este, não é interessante a multiplicidade de possibilidades tendendo ao infinito sem qualquer direcionamento ou indicativos de caminhos a seguir. Conforme afirma Ingram (2011) a curadoria e a “filtragem” de conteúdo são cada vez mais necessários:

> “O aumento de fontes de informação em tempo real produziu uma onda de conteúdo tão imparável que precisamos de curadoria e filtragem mais do que nunca antes. (INGRAM, 2011)

Nesse contexto, os sites e produtos de notícias são resultados de projetos onde suas interfaces e fluxos de navegação são soluções pensadas de maneira a oferecer a melhor experiência possível para os usuários. Parte dessa tarefa passa por deixar clara a existência dos caminhos e fluxos existentes para que o usuário dê sequência a sua navegação.

Entretanto, no caso específico dos produtos de notícias, muitos ainda são projetados com base em uma lógica de usuários iniciando suas experiências a partir das homepages, ou páginas iniciais dos sites (cujo paralelo pode ser traçado com as capas dos jornais de papel). Nesta lógica, segundo Ingram (2008), o equívoco está na visão de que todo usuário navegaria a partir de um mesmo ponto de partida (a homepage) e a partir daí seguiria por uma ordem linear de consumo similar a dos jornais impressos:

> “Grande parte dos sites continua sendo projetada como se as pessoas lessem notícias digitais da mesma forma que liam nos jornais impressos — começando pelo “início” e acessando págna por página até o “final” (obviamente, muitas pessoas sequer lêem os jornais impressos dessa forma, mas isso é outra questão). Na realidade, as pessoas vêm de diferentes lugares, caem em uma notícia e depois desaparecem, se localizando a partir de links e conteúdos relacionados”. (INGRAM, 2008)

Stray (2010) corrobora com esta visão e aponta para o mesmo equívoco de projeto nos produtos de notícias, por vezes ainda favorecendo o modelo de consumo linear e estático:

> “Um problema profundo dos produtos de jornalismo existentes é que eles não são originalmente projetados para serem pessoais. Talvez isso seja decorrente da mentalidade clássica "vamos dizer-lhes o que eles precisam saber" que existe entre os editores. Isso faz sentido para um produto de papel, onde todos lêem exatamente a mesma cópia, e você pode realmente "terminar" de consumir o conteúdo quando você ficar sem páginas para ler. Mas não faz sentido quando as páginas são efetivamente infinitas, e não há nenhuma razão para supor que a minha leitura vá ser semelhante à sua”. (STRAY, 2010)

Como decorrência disso, em muitos casos as homepages são editadas e projetadas de acordo com uma percepção do que é ou não relevante para o usuário do ponto de vista do corpo editorial da empresa de notícias, e como se a navegação de cada usuário fosse necessariamente ter início a partir desta mesma página. Entretanto, os números mostram que a lógica de uma navegação inteiramente baseada em homepages e uma curadoria editorial vem perdendo força no modelo de consumo dos usuários de sites e demais produtos de notícias.

### 1.1.2 O declínio das homepages e o conteúdo como ‘*página de destino’*

A figura 1 apresenta os resultados de pesquisa realizada pelo “The Pew Research Center's Project for Excellence in Journalism”, apresentada por Mitchell (2012). Nestes resultados, observa-se que apesar das redes sociais ainda não serem a principal origem para o consumo de notícias, apenas 36% dos usuários norte-americanos acessava diretamente um site de notícias (ou seja, via tráfego direto) em 2012. Na prática, os 64% restantes acessavam tendo como origem redes sociais, buscadores ou outras fontes externas.

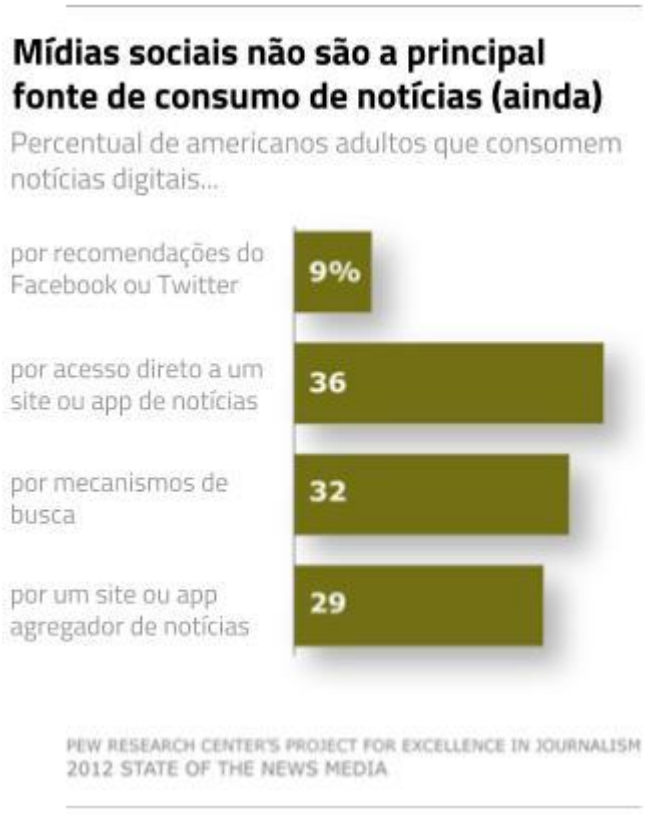


Figura 1 – Percentual de adultos americanos que acessa notícias em dispositivos digitais por diferentes origens.
Fonte: adaptado de "MITCHELL, Amy; ROSENTIEL, Tom. 2013".

No Brasil, estudos mais recentes mostram que esse padrão, além de crescente, é ainda mais acentuado. Conforme levantamento do Instituto Reuters de Estudos sobre o Jornalismo (Newman, 2016), 72% da população brasileira utiliza as redes sociais como fontes para consumo de notícias:

> “De acordo com um levantamento anual do Instituto Reuters de Estudos sobre Jornalismo, entre 26 países pesquisados, o Brasil é o que mais cresce em uso das redes sociais como principal fonte de informação.
> No país, 72% dos entrevistados afirmaram utilizar as redes sociais como fontes de notícias, percentual que só fica atrás da Grécia (74%) e da Turquia (73%). A variação entre os países analisados é grande: esse índice é de apenas 31% na Alemanha, e 35% no Reino Unido. Nos Estados Unidos e na Europa, ele é de 46%, com média geral sendo 51%.”
> http://www.digitalnewsreport.org/survey/2016/overview-key-findings-2016/#sharetab-0 (NEWMAN, 2016)

Ainda de acordo com o mesmo estudo, o Brasil é líder na tendência de mudança na origem do tráfego de notícias:

> “Informações divulgadas pelo site Convergência Digital apontam que a maior mudança na mídia digital foi o crescimento de notícias acessadas por redes sociais

> como Facebook, Twitter, Instagram e Snapchat. Além de estar no grupo dos que já têm alta adesão a essa prática, o Brasil lidera no crescimento dessa tendência, tendo praticamente dobrado em um ano o percentual daqueles que dizem ter nas redes sociais a principal fonte de informação – 10% em 2015 para 18% em 2016." http://www.digitalnewsreport.org/survey/2016/overview-key-findings-2016/#sharetab-0 (NEWMAN, 2016)

É justamente o maior acesso a partir de origens externas um dos principais fatores que promovem o declínio da homepage em sua função original de porta de entrada dos sites de notícias. Tal fato se deve a própria natureza do comportamento dos usuários: em um buscador, por exemplo, é comum a maior parte dos usuários buscar por um assunto específico, e dessa maneira os resultados apresentados tendem a ser páginas de conteúdos específicos, e não homepages que agregam diversos conteúdos, como acontece nos sites de notícias.

Da mesma forma, em uma rede social, a maioria dos usuários compartilham, consomem e comentam os conteúdos de seu interesse, em detrimento do que é ou não oferecido pelas homepages dos sites. Dessa maneira, a pessoa que faz uso de redes sociais como fonte para o consumo de notícias tem no compartilhamento de outros usuários o fator que estimula e direciona seu acesso a informação. Também neste caso, o usuário tende a acessar diretamente um conteúdo específico, sem passar pela homepage do site de notícias.

O gráfico a seguir (figura 2) mostra um exemplo prático e real do declínio da audiência nas homepages do site do NY Times entre 2011 e 2013.

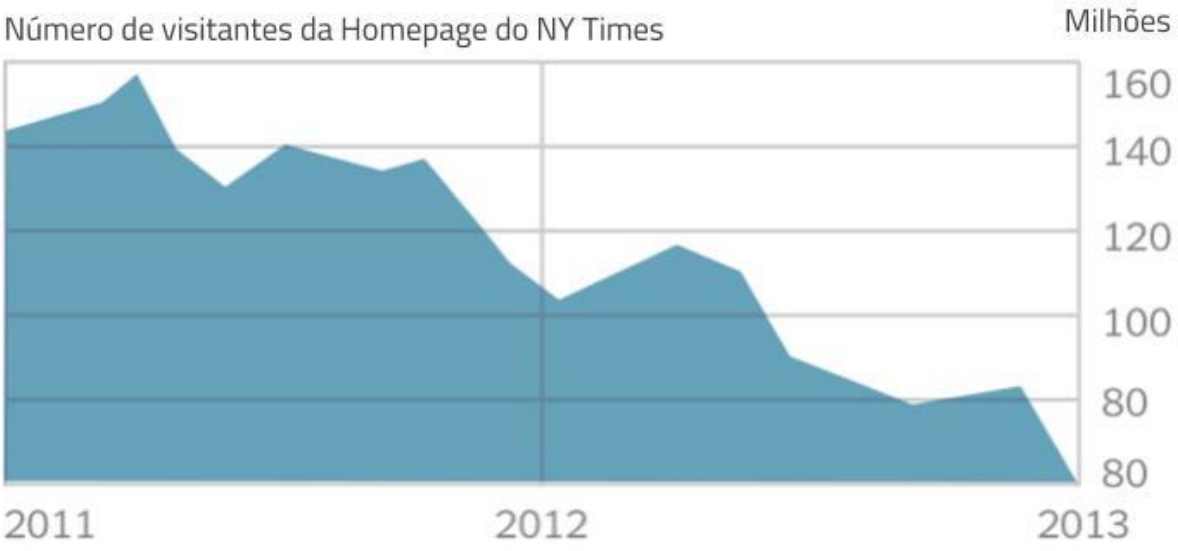


Figura 2: Tráfego da homepage do NY Times caiu pela metade em 2 anos (SEWARD, 2014)

O exemplo pontual do NY Times somente exemplifica e deixa mais clara uma realidade que é comum na grande maioria dos produtos de notícia. A queda na audiência via homepages de fato é constante e afeta também os dispositivos móveis, mostrando ser um novo padrão de comportamento que independe do dispositivo de acesso, conforme analisa Seward (2014):

> "O tráfego da homepage vem caindo, mês a mês, há anos. Mesmo nos nossos aplicativos para dispositivos móveis, que são basicamente réplicas das nossas homepages, o tráfego vem caindo". (SEWARD, 2014)

Em artigo publicado no site "Nielsen's Lab", Adrianne Lafrance (2012) confirma a mudança de papel das homepages, que perdem cada vez mais sua relevância como direcionadores de tráfego:

> "Quanto mais as pessoas entram nos sites por fontes externas — via mecanismos de busca, links recebidos por email, Facebook ou Twitter — as empresas de notícias estão vendo que suas homepages não são mais as portas de entrada de antigamente" (LAFRANCE, 2012).

No artigo de Lafrance (2012), o site americano de notícias "The Atlantic" é citado como exemplo onde as estatísticas confirmam a transição do papel de porta de entrada das homepages para as matérias:

> "A página de matéria é hoje a principal porta de entrada do 'The Atlantic'. O velho mantra de que cada página precisa ser uma homepage nunca foi tão verdadeira. As pessoas vêm em função de um conteúdo, e o objetivo é dar a elas uma experiência de leitura limpa e interessante — elegante, não muito poluída, com um pouco de arte — e além disso ter a certeza de que estamos oferecendo ao usuário os demais conteúdos do nosso site". (LAFRANCE, 2012).

Portanto, se inicialmente a experiência de consumo de notícias envolvia a repetição contínua do fluxo "homepage > matéria > homepage", essa mudança de tendência tem feito com que cada vez mais usuários acessem aos sites vindo de fontes externas diretamente pela área de um conteúdo específico – normalmente uma página de matéria, conforme representado na figura 3.

Figura 3 – Comparativo entre os acessos iniciados via homepages e os acessos iniciados via fontes externas.

Ao iniciar a navegação diretamente pelo conteúdo, este – e não mais a homepage – passa a exercer verdadeiramente a função de ‘página de destino’, que conforme definição apresentada é a página de entrada de um usuário em um site.

Stray (2010) afirma que cada vez mais as páginas de conteúdo devem ser projetadas como uma possível página de entrada do usuário, de maneira a estimular a sequência da sua navegação e a interação com a informação de seu interesse a partir de uma experiência de consumo mais pessoal.

> “As páginas de destino dos sites de notícias devem conter mais hiperlinks do que qualquer outra coisa (e não somente links internos para SEO, mas links que sejam úteis para o usuário). A experiência de consumo de notícias precisa ser cada vez mais pessoal. Tem que ser fácil para os usários encontrarem conteúdos do seu interesse, não importa de que maneira”. (STRAY, 2010)

Entretanto, como será apresentado na sequência, as páginas de conteúdo dos sites de notícias parecem não exercer de maneira satisfatória seu papel de “página de destino”, tendo em vista as métricas bem inferiores em relação aos acessos iniciados pelas homepages destes mesmos sites.

### 1.1.3 Queda na audiência e no engajamento dos usuários

A avaliação de métricas é fator constante no modelo de negócios de sites, aplicativos e produtos digitais de maneira geral, em prática conhecida como ‘web

analytics'. Carneiro (2008) define a prática do Web Analytics como o "monitoramento dos websites para que as empresas possam ter um melhor entendimento das complexas interações entre as ações dos visitantes e as ofertas que o site possui, assim como prover informações para aumentar a lealdade dos usuários" (CARNEIRO, 2008).

Turchi (2009) corrobora a importância das métricas para entender não apenas o próprio produto digital, mas também o comportamento de seus usuários, possibilitando o projeto de soluções que atender às suas espectativas:

> "Métricas são análises fundamentais para melhor se conhecer, controlar e aperfeiçoar as ações on-line. Podem ser mensurados o retorno, as conversões, as visitas (pageviews), a interação, as pessoas impactadas e as impressões". (TURCHI, 2009).

A partir do conhecimento detalhado do comportamento dos usuários de um site torna-se possível a prática de traçar estratégias segmentadas para cada tipo de público, segundo Okada (2011):

> O desenvolvimento de métricas para múltiplas mídias elevam a quantidade de informações úteis sobre o perfil de consumo possibilitando a otimização das estratégias digitais a públicos segmentados (OKADA, 2011).

No contexto dos produtos de notícias, Okada (2011) ressalta ainda a importância dada pelas organizações para a mensuração das ações na tentativa de criar maior valor para seus usuários:

> "Vivemos a era da recomendação com a disseminação exponencial de conteúdo relevante, ascensão do prosumerismo e crowdsourcing. A compreensão deste cenário complexo e o conhecimento das tecnologias disponíveis constituem um diferencial competitivo para as organizações. Ressalta-se, portanto a importância da mensuração de ações de marketing para criar valor para os consumidores, visto que as métricas relacionadas à satisfação dos consumidores e a fidelidade tem ganhado maior atenção por parte das organizações" (OKADA, 2011)

Através da mensuração do comportamento e das ações dos usuários, observa-se que não obstante a queda na quantidade de acessos via homepages dos sites de notícias (apresentada no subcapítulo anterior), o novo cenário apresenta ainda queda consistente na qualidade desses acessos. Tal queda pode ser verificada em uma análise mais aprofundada das métricas de engajamento.

Para estabelecer o conceito de “engajamento” no âmbito dos sites e produtos digitais, é importante ressaltar que o termo tem sido utilizado em diversas disciplinas acadêmicas, conforme cita Marra (2013).

> “Na última década, o termo engajamento tem sido tratado em algumas disciplinas acadêmicas: Sociologia, Ciência Política, Psicologia e Comportamento Organizacional (BRODIE et al., 2011). Na Sociologia, o foco está no engajamento cívico; na Ciência Política, no engajamento do Estado e na abrangência desse engajamento. Na Psicologia, o interesse está no engajamento social, de tarefa e ocupacional; a Psicologia Educacional centra-se no engajamento do estudante. Já na disciplina Comportamento Organizacional o objeto de estudo é o engajamento do funcionário (BRODIE et al., 2011; HOLLEBEEK, 2011b). Por fim, no Marketing, o interesse reside no relacionamento dos consumidores com as empresas.” (MARRA, 2013)

Em abordagem mais subjetiva, Higgins e Scholer (2009) definem o engajamento como o grau de envolvimento ou absorção em relação a algo:

> “Engajamento é um estado de estar envolvido, ocupado, totalmente absorvido ou absorto em algo (ou seja, mantendo atenção), que gera as consequências de uma atração especial ou força de repulsão. Os indivíduos mais engajados vão se aproximar ou repelir um alvo quando mais valor for adicionado ou subtraído dele.” (HIGGINS E SCHOLER, 2009)

Ao contextualizar o uso do termo aos meios digitais, Rez (2017) apresenta uma definição de “engajamento” que aponta para a mensuração de métricas comuns, como “número de visitas”, “tempo gasto por página” ou “número de páginas vistas” como fatores que definiriam um menor ou maior grau de engajamento:

> “Engajamento é envolvimento, interação, relacionamento com a marca, que vai além do número de seguidores em uma rede social ou likes em uma postagem. O aspecto do envolvimento pode ser mensurado através das métricas comuns, como número de visitas ao site, tempo gasto por página, número de páginas vistas”.
> (REZ, 2017)

Okada (2011) faz referência ao termo “métricas de engajamento” como o resultado analítico prático do envolvimento dos usuários, citando sua importância cada vez maior para sites e produtos digitais:

> “As métricas de engajamento devem ganhar cada vez mais espaço nas análises, [...] devido a uma demanda crescente pelo entendimento dos resultados que estão além da conversão principal, em um cenário de ações cada vez mais interativas e multifacetadas” (OKADA, 2011)

É justamente nessas métricas, as chamadas “métricas de engajamento” citadas anteriormente, em que é possível observar uma queda consistente à medida que os acessos via homepages perdem espaço.

O resultado de um estudo acerca da origem dos acessos aos sites de notícias norte americanos feito pela “The Pew Research Center's Project for Excellence in Journalism” apresentado por Mitchel (2014) deixa claro que o tempo gasto por página (uma das métricas que caracterizam o engajamento do usuário) vêm apresentado quedas consistentes em acessos que tem por origem as redes sociais ou os buscadores:

> “Neste estudo sobre o tráfego de internet dos EUA para 26 dos sites de notícias mais populares, os visitantes diretos - aqueles que digitam o endereço específico (URL) ou têm o endereço marcado - passam muito mais tempo no site de notícias, visualizam muito mais páginas de conteúdo e voltam com muito mais freqüência do que os visitantes que chegam de um mecanismo de pesquisa ou uma referência no Facebook.” (MITCHEL, 2014)

Na figura 4, observa-se que todas as métricas de engajamento de quem acessa diretamente um site de notícias via tráfego direto são melhores em relação aos usuários cuja origem do tráfego se dá a partir do Facebook ou de mecanismos de busca.

Figura 4 - Pesquisa do Pew Research Center comparando métricas de usuários que acessam por tráfego direto ou vindo por facebook e buscadores.

Na imagem observa-se que a duração da visita de quem acessa sites de notícias por tráfego direto é cerca de 275% maior do que os acessos iniciados via facebook ou buscadores. Da mesma forma, o número de páginas por visita é quase 590% maior entre quem acessa via tráfego direto.

Os dados mostram ainda que, apesar das diferenças entre os sites de notícias analisados, o mesmo padrão de queda de engajamento pode ser identificado em todos eles.

> “Esse nível mais alto de engajamento dos visitantes diretos é verdadeiro em todo o mix de sites estudados, daqueles que estão entre os mais compartilhados no Facebook, como o breitbart.com, para aqueles cujo tráfego é fortemente impulsionado pelo tráfego dos mecanismos de pesquisa, como abcnews.go.com (o site da ABC News); e daqueles com um pequeno público total (mrconservative.com) para agregadores (yahoonews.com). Até mesmo sites como o nativo digital buzzfeed.com e o npr.org, da National Public Radio, que têm um alto nível de tráfego no Facebook, tiveram um envolvimento muito maior daqueles que entraram diretamente.” (MITCHEL,2014)

Apesar do padrão consistente de queda de engajamento, foram encontradas grandes diferenças nas métricas de engajamento gerais entre os sites analisados, trazendo indícios de que a experiência oferecida pelo site é capaz de impactar no engajamento de seus usuários.

> Dentro dessas descobertas gerais, alguns sites de notícias exibem níveis mais altos de engajamento geral de visitantes do que outros. O visitante médio do Foxnews.com (um site operado pelo Fox News Channel) passou quase oito minutos por visita. Em comparação, o visitante médio da CNN gastou 1 minuto e 30 segundos no site, enquanto os visitantes do nbcnews.com ficaram cerca de quatro minutos em média. (MITCHEL,2014)

Do ponto de vista mercadológico, quanto maior o engajamento dos usuários em um site, maior seu potencial de monetização, ou seja, de transformar os acessos desses usuários em lucro. Dessa forma, a maior parte dos sites de notícias do Brasil e do Mundo vêm buscando formas de ampliar o engajamento e promover melhorias em suas métricas de utilização.

Nesse sentido, o artigo "What you think you know about the web is wrong" (Haile, 2015) reforça o tempo de permanência em um site como uma estatística de fundamental importância, interferindo de maneira decisiva na possibilidade de retorno e fidelização do usuário ao portal noticioso:

> “Pesquisas realizadas pelo site Chartbeat indicaram que usuários que permaneceram por pelo menos 3 minutos navegando em um site tem uma tendência a retornar ao mesmo site 2 vezes maior em relação a usuários que ficam por menos de 3 minutos.” (Haile, 2015)

É necessário, portanto, buscar pela ampliação do engajamento dos usuários cuja página de destino mais comum passa a ser o próprio conteúdo, e cuja navegação acontece cada vez menos orientada à organização de leitura sugerida pelas homepages. Torna-se, portanto, fundamental para o produto noticioso o projeto de páginas de conteúdo capazes de exercer essa “nova função” que lhe foi atribuída – a de recomendação de conteúdos relevantes de maneira cada vez mais pessoal e assertiva, mantendo o interesse do usuário e estimulando a continuidade da navegação.

Seward (2014) destaca a mudança do modelo de consumo baseado na “Pull Media” (a mídia que o usuário busca ativamente) para um consumo que se dá pela “Push Media” (a mídia que chega até o usuário, que “encontra” o usuário):

> “‘O modelo de Pull media’ foi rapidamente sendo substituído pelo modelo de ‘push media’, conforme deixa claro o relatório sobre o NY Times. A informação — atualizações de status, fotos dos seus amigos, vídeos e até mesmo as notícias — vem até você; elas te encontram. (SEWARD, 2014)

Dessa maneira, não é coerente nos dias de hoje apostar em um modelo baseado no controle editorial da navegação e do fluxo de consumo de conteúdo de cada usuário. O relatório da pesquisa “A New Model for News: Studying the Deep Structure of Young-Adult News Consumption” apresentada em 2008 pela Associated Press (associação norte-americana de imprensa) aponta que cada conteúdo em um site de notícias possui diferentes portas de entrada, e que cada usuário chegará com variadas perspectivas e vindo de diferentes plataformas de origem.

Caberia, portanto, aos editores criarem não só bons conteúdos, mas principalmente encontrar as melhores formas de conectar esses pontos de entrada, oferecendo ao usuário uma experiência que faça com que ele tenha vontade de continuar consumindo as demais notícias.

> “Uma questão-chave, demonstrada neste estudo, é que os consumidores estão acessando os pontos de entrada às notícias tanto por acaso como por design. Os editores humanos em um site específico podem definir caminhos de uma notícia para outra e adicionar links para outra cobertura de mídia, botões para compartilhar com um amigo, etc. Mas essa atividade cobre apenas uma pequena fração das muitas rotas tortuosas que os consumidores agora estão habilitados a tomar através das notícias. Hipoteticamente, um novo valor pode ser criado se os produtores e

> distribuidores de notícias puderem descobrir formas de ajudar os consumidores a conectar os pontos de forma mais coerente. Mas onde você começa? Este estudo confirmou a necessidade de manter duas faixas de trabalho paralelas. Primeiro, crie um conteúdo mais atraente para os pontos de entrada chave. Em segundo lugar, e mais difícil, crie as conexões que irão transportar os consumidores para esse conteúdo tanto na plataforma de mídia como na marca. Como a AP descobriu, é um trabalho para pessoas e máquinas". (Associated Press, 2008)

É justamente na "criação das conexões que irão transportar os consumidores para este conteúdo" que está o maior desafio das pessoas que atuam como profissionais da área (designers, jornalistas, editores e programadores) e das "máquinas": recomendar conteúdos de maneira assertiva e no momento certo para atrair novos usuários ou mantê-los por mais tempo consumindo conteúdo relevante.

É nesse sentido que Jolly (2014) também reforça que os acessos oriundos de fontes externas e a consequente queda no engajamento dos sites de notícias proporcionaram o momento propício para o investimento nos chamados "sistemas de recomendação de conteúdo".

> "O tráfego via homepages continua a cair nos sites de notícias. Isso é resultado da mentalidade oriunda do crescimento das mídias sociais, quando as pessoas começaram a acessar conteúdos recomendados por seus amigos: "se a notícia é importante, ela irá me encontrar". Dessa forma, cada vez mais os leitores estão descobrindo as notícias via mídias sociais, email ou aplicativos de leitura.
> As empresas de notícias estão atentas a isso, investindo na implementação de sistemas baseados em algoritmos que podem modificar a experiência do site em funçao da origem do tráfego de cada usuário, por exemplo". (JOLLY, 2014)

A seguir, serão discutidos conceitos referentes a lógica de funcionamento dos sistemas de recomendação de conteúdo, bem como os desafios de sua implementação num âmbito geral e mais especificamente no âmbito do conteúdo noticioso.

## 1.2 **Sistemas de Recomendação**

Para Resnick (1997), sistemas de recomendação são "sistemas que auxiliam no aumento da capacidade e eficácia do processo de indicação já bastante conhecido na relação social entre seres humanos".

Já Ricci (2011) define os sistemas de recomendação como "ferramentas e técnicas de software que fornecem sugestões para que os itens sejam úteis para um

usuário". Tais "sugestões" fazem referência a qualquer processo de tomada de decisão por parte do usuário: definir um item para compra, uma música para ouvir, ou uma notícia para ler.

Em seu conceito mais simples, sugerir ou recomendar diz respeito a oferecer uma lista classificada de itens para um determindado usuário. Para identificar e recomendar essa "lista de itens", os sistemas devem prever quais deles são mais adequados para cada usuário. Para isso, diferentes recursos podem ser utilizados com o objetivo de coletar dados dos usuários que ajudem a indicar suas preferências e, dessa maneira, atingir um maior grau de assertividade na sua função de recomendação de conteúdo.

Shardanand (1995) afirma que o mecanismo de recomendações para a tomada de decisão é uma prática comum entre os seres humanos, mesmo muito antes dos computadores serem inventados. Entretanto, a história dos sistemas de recomendação na computação começou já na segunda metade do século XX.

Um dos primeiros sistemas de recomendação que se tem conhecimento denominou-se 'Tapestry', desenvolvido no início dos anos 90 (Goldberg et all, 1992). Seus responsáveis cunharam o termo "Filtragem Colaborativa" para definir um tipo de sistema específico no qual a filtragem da informação era realizada com o auxílio humano, ou seja, através da colaboração entre grupos de pessoas. Apesar de vários pesquisadores rapidamente adotarem o termo, Resnick (1997) defendeu em seu artigo intitulado "Recommender Systems, Communications of the ACM 40" o termo "sistemas de recomendação" como uma terminologia mais genérica e abrangente - já que podem existir sistemas de recomendação sem necessariamente colaboração entre pessoas.

### 1.2.1 Sistemas de recomendação automatizados

Os primeiros sistemas de recomendação automatizados na internet surgiram em meados dos anos 2000, a partir da sistematização de uma lógica comportamental: pessoas geralmente dependem de recomendações fornecidas por outros na tomada de suas decisões diárias (RICCI et all, 2011).

Jones (2013) ressalta a importância dos sistemas de recomendação automatizados na forma como se realiza a interação e a navegação dos usuários com os sites e produtos digitais nos dias atuais:

> “Os sistemas de recomendação mudaram a forma como os sites inanimados se comunicam com seus usuários. Em vez de fornecer uma experiência estática em que os usuários pesquisam e potencialmente compram produtos, os sistemas recomendadores aumentam a interação para fornecer uma experiência mais rica. Os sistemas de recomendação identificam recomendações de forma autônoma para usuários individuais com base em compras e buscas passadas e no comportamento de outros usuários”. (JONES, 2013)

A automatização da recomendação de conteúdo se dá a partir de sistemas baseados em algoritmos. Diakopoulos (2013) define um algoritmo como “uma série de etapas empreendidas para resolver um problema específico ou realizar um resultado definido”. (DIAKOPOULOS, 2013)

Do ponto de vista conceitual, em sistemas de recomendação de conteúdo automatizados baseados em algoritmos, o objetivo é estimar, a partir de um conjunto de princípios e regras, quais itens serão potencialmente interessantes para um determinado usuário, antes que ele tenha manifestado ativamente seu interesse pelo conteúdo recomendado.

O modelo da figura 5 representa o funcionamento básico de um sistema de recomendação (MELVILLE, 2010). Na figura, a matriz representa a relação entre usuários e itens, onde as pontuações equivalem à classificação dada ao interesse de um usuário (u) específico por um item determinado. O trabalho dos sistemas de recomendação é, portanto, prever qual a classificação do interesse do usuário nos itens onde ainda não foi possível aferir um grau de classificação (as lacunas em branco na matriz da figura). Em um sistema de recomendação tradicional, os itens com maior classificação de interesse tendem a ser os itens recomendados para o usuário.

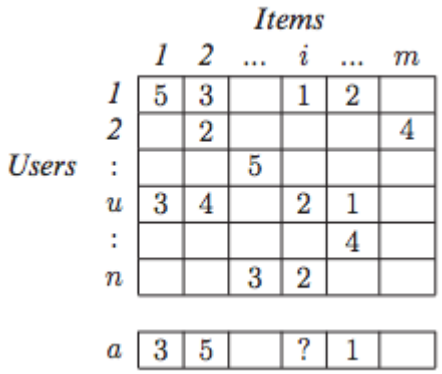


Figura 5 - Matriz de avaliação de usuários em sistemas de recomendação (Melville, 2010)

Ricci (2011) amplia a discussão sobre os objetivos dos sistemas de recomendação, a medida que vai além da já citada ampliação das receitas de produtos digitais e da questão da relevância - outro objetivo óbvio de um sistema de recomendação é sugerir itens relevantes para seu usuário. Embora sejam estes de fato os principais objetivos dos sistema de recomendação, é importante destacar ainda alguns objetivos secundários: a “novidade” – garantir que o item recomendado seja algo que o usuário ainda não tenha visto; “Serendipity” ou “sorte” – recomendar algo que fuja do interesse óbvio mas que seja igualmente atrativo, de maneira a possibilitar a descoberta de novos padrões e verdadeiramente surpreender o usuário; e a “Diversidade” – garantir que os itens recomendados não sejam tão semelhantes entre si visando aumentar a probabilidade de que algo de fato interesse ao usuário.

Para que uma recomendação aconteça, é necessário que sejam definidos critérios de escolha ou seleção de quais itens devem ser priorizados, e de que maneira (ou sob que regras ou lógicas algorítmicas) o sistema deve funcionar. Diakopoulos (2013) afirma que é esse conjunto de regras que permite o funcionamento autônomo do modelo, independente de serem regras estáticas (pré-definidas por programadores) ou dinâmicas (algoritmos que aprendem e se modificam a partir do comportamento dos usuários).

> “A tomada de decisão autônoma é o cerne do poder algorítmico. As decisões algorítmicas podem basear-se em regras sobre o que deve acontecer em um processo, dado o que já aconteceu, ou em cálculos sobre quantidades maciças de dados. As próprias regras podem ser articuladas diretamente pelos programadores, ou ser dinâmicas e flexíveis com base nos dados. Por exemplo, algoritmos de aprendizagem em máquina permitem que outros algoritmos tomem decisões inteligentes com base em padrões aprendidos em dados”. (DIAKOPOULOS, 2013)

Nesse ponto, Mitchell (1997) alerta que o principal fator que determina o sucesso de um sistema automatizado está em suas regras gerais e específicas de projeto. A definição de como o sistema irá aprender, de quais os dados que devem ser levados em conta, de quais são as suas regras de funcionamento, são fatores chave para o sucesso ou fracasso desse sistema:

> "A primeira escolha de projeto que enfrentamos é escolher o tipo de experiência de treinamento a partir da qual nosso sistema aprenderá. O tipo de experiência de treinamento disponível pode ter um impacto significativo no sucesso ou fracasso do aluno." (MITCHELL, 1997)

Atualmente, muitas abordagens algorítmicas estão disponíveis para sistemas de recomendação. Muitos dos algoritmos provêm do campo da "machine learning" (numa tradução literal, aprendizado de máquinas), um subcampo de inteligência artificial que produz algoritmos para aprendizagem, previsão e tomada de decisão. Shalev-Shwartz (2014) define "machine learning" como o processo que permite a detecção automatizada de padrões em uma massa de dados, estando cada vez mais presente em diferentes áreas:

> "O termo 'machine-learning' refere-se à detecção automatizada de padrões em dados. Nas últimas duas décadas, tornou-se uma ferramenta comum em quase qualquer tarefa que requer a extração de informações de grandes conjuntos de dados. Nós estamos rodeados por uma tecnologia baseada em aprendizado de máquina: os mecanismos de pesquisa aprendem como trazer os melhores resultados (ao colocar anúncios rentáveis), softwares anti-spam aprendem a filtrar nossas mensagens de e-mail, e as transações com cartão de crédito são protegidas por um software que aprende a detectar fraudes. Câmeras digitais aprendem a detectar rostos e aplicativos de assistência pessoal inteligente em smartphones aprender a reconhecer comandos de voz. Carros são equipados com sistemas de prevenção de acidentes que são construídos usando algoritmos de aprendizado de máquina". (SHALEV-SHWARZ, 2014)

Smola (2008) reitera o crescimento da importância da área da aprendizagem automatizada dos sistemas a medida que cada vez é necessário lidar com grandes e crescentes massas de dados:

> "Nas últimas duas décadas, o Aprendizado de Máquina tornou-se um dos pilares de tecnologia da informação e com isso, um ponto central, embora geralmente escondido, de nossas vidas. Com a crescente quantidade de dados se tornando disponível há uma boa razão para acreditar que a análise de dados inteligente se tornará ainda mais difundida como um ingrediente necessário para o progresso tecnológico." (SMOLA, 2008)

Neste ponto, cabe ressaltar que, apesar do aprofundamento em aspectos técnicos da aprendizagem automatizada e dos algoritmos não ser foco desta pesquisa, faz-se necessário apresentar alguns aspectos introdutórios da lógica algorítmica no âmbito dos sistemas de recomendação.

Jones (2013) aponta a Correlação de Pearson (“Pearson correlation”) – onde a semelhança entre dois usuários (e seus atributos, como artigos lidos ou itens comprados) pode ser calculada com precisão ao medir a dependência linear entre duas variáveis (ou usuários) como uma função de seus atributos; e os Algoritmos de Agrupamento (“Clustering algorithms”) – identificam semelhanças entre itens, calculando sua distância e agrupando os itens mais próximos entre si como de um mesmo “tipo”.

Diakopoulos (2013) determina ainda que os algoritmos dos sistemas de recomendação poderiam ser divididos em quatro categorias com base nos tipos de decisões que eles tomam. Segundo ele, é possível avaliar e interpretar o poder algorítmico ao pensar nas decisões atômicas de cada algoritmo, a saber: 1) priorização; 2) classificação; 3) associação e 4) filtragem (DIAKOPOULOS, 2013).

1) **Priorização**: A priorização de itens ou conteúdos tem a função de chamar mais a atenção para determinados itens em detrimento de outros. Em geral, é utilizado para definir os critérios de destaque ou ordenação da recomendação.
2) **Classificação**: As decisões de classificação tratam de categorizar um ou mais itens específicos como constituintes de um mesmo tipo ou classe, agrupando esses itens.
3) **Associação**: As decisões de associação dizem respeito a marcação de relações entre os itens, ou seja, o quanto tais itens constituem uma relação de similaridade entre si sob algum tipo de regra ou critério definidos.
4) **Filtragem**: Já a filtragem envolve a inclusão ou exclusão de itens de acordo com várias regras ou critérios definidos. Em muitos casos, os algoritmos de filtragem podem levar em consideração decisões obtidas a partir dos três processos anteriores (priorização, classificação e/ou associação).

Entretanto, apesar de essencialmente a maioria dos sistemas de recomendação funcionar a partir de lógicas algorítmicas, está claro que, por sua

própria definição, cada algoritmo pode se constituir em um conjunto próprio e distinto de etapas e regras. Dessa forma, cada sistema de recomendação poderá mensurar os interesses de seus usuários de forma única, de acordo com o problema que a lógica algorítmica foi projetada para resolver ou os relacionamentos presentes nos dados analisados. Assim, diferentes sistemas de recomendação tendem a apresentar resultados bastante distintos.

Alguns dos principais estudiosos da área vem analisando e categorizando sistemas de recomendação automatizados a partir de características comuns em suas lógicas de funcionamento.

Adomavicius e Tuzhilin (2005) propõem a classificação dos sitemas de recomendação em duas categorias principais, com base em como as recomendações são feitas:

1) **Recomendações baseadas em conteúdo**: o usuário receberá recomendações de itens similares aos que já acessou no passado.
2) **Recomendações colaborativas**: serão recomendados ao usuário itens que pessoas com gostos semelhantes demonstraram ter gostado no passado.

Melville e Sindhwani (2010) determinam que os sistemas de recomendação de conteúdo poderiam ser categorizados em três tipos, baseados em qual será o pilar para a análise dos algoritmos (os usuários, os conteúdos, ou ambos): 1) Filtragem Colaborativa (“Collaborative Filtering”); 2) Recomendação baseada nos conteúdos (“Content-based recommending”) e 3) Abordagem Híbrida (“Hybrid approaches”).

1) **Filtragem Colaborativa:** em sistemas baseados na Filtragem Colaborativa, um usuário recebe recomendações de itens com base em ações passadas de todos os demais usuários coletivamente. Ou seja, neste modelo, o que funcionou bem para a maioria dos usuários sob um definido critério classificatório tenderia a funcionar bem para qualquer novo usuário do sistema.

   Tais sistemas funcionam a partir da coleta de dados de ações dos usuários, de maneira que seja possível classificar os itens daquele determinado

domínio e explorar padrões de comportamento entre vários usuários na determinação de quais desses itens poderá interessar ao usuário ativo.

Para o funcionamento deste modelo, um subconjunto de usuários é escolhido com base em sua semelhança com o usuário ativo, e uma combinação ponderada de suas ações naquele sistema é usada para produzir previsões para este usuário. Em sua maioria, essas abordagens podem ser generalizadas por: Atribuir um peso a todos os usuários em relação à semelhança com o usuário ativo; selecionar um número de usuários que tenham a maior semelhança com o usuário ativo; calcular uma previsão de uma combinação ponderada baseada nos usuários selecionados.

2) **Recomendação baseada nos conteúdos:** em sistemas de recomendação baseada nos conteúdos, o objetivo é recomendar itens que sejam semelhantes em conteúdo a itens que o usuário ativo gostou ou acessou no passado.

   No modelo mais puro, estes sistemas se utilizam apenas da matriz de classificação dos interesses dos usuários. Essas abordagens tratam todos os usuários e itens como unidades atômicas, onde as previsões são feitas sem levar em conta especificações individuais de cada usuário. Melville (2010) alerta sobre a possibilidade de uma recomendação melhor personalizada à medida que se sabe mais informações sobre o usuário ou sobre o item em questão (não somente seu comportamento passado dentro daquele site ou sistema, mas inferências sobre características particulares de cada indivíduo).

   Em um exemplo prático de um sistema de recomendação de filmes, ao saber informações de gênero de um filme e ter informações de que um usuário gostou de "Star Wars” e “Blade Runner", poderia-se inferir a preferência pelo gênero de ficção científica e recomendar um terceiro filme desse estilo.

3) **Abordagem Híbrida:** os métodos chamados “híbridos” combinam elementos da filtragem colaborativa e da recomendação baseada nos conteúdos, de maneira a aproveitar os pontos fortes e evitar os pontos fracos de cada uma delas.

Jones (2013) é outro autor a reiterar a ideia de que a maioria dos sistemas de recomendação pode se enquadrar em duas abordagens básicas, além de um terceiro modelo híbrido: 1) Filtragem Colaborativa (“Collaborative Filtering”); 2)

Recomendação baseada nos conteúdos (“Content-based”) e 3) Abordagem Híbrida (“Hybrid approaches”).

1) **Filtragem Colaborativa**: Recomendação baseada no comportamento anterior do usuário ativo. Tal modelo poderia ser construído exclusivamente a partir do comportamento de um único usuário ou pelo comportamento de outros usuários com características semelhantes (modelo mais comum).
2) **Filtragem baseada em conteúdo**: Recomendação construída com base no comportamento de um usuário. Por exemplo, essa abordagem pode usar informações de navegação histórica. Se um usuário geralmente lê artigos de um mesmo assunto, a filtragem baseada em conteúdo pode usar esse histórico para identificar e recomendar conteúdo semelhante. Este conteúdo poderia ser definido manualmente ou extraído automaticamente com base em outros métodos de similaridade.
3) **Abordagem Híbrida**: combinações de filtragem colaborativa e baseada em conteúdo visando aumentar a eficiência (e a complexidade) dos sistemas de recomendação.

A abordagem de Ricci (2011) sobre categorização é parecida a de Melville e Sindhwani (2010) e Jones (2013), mas segmentando ainda mais e ampliando o espectro total da categorização. Para ele, os sistemas de recomendação de conteúdo poderiam ser categorizados em seis tipos: 1) Recomendação Baseada no Conteúdo (“Content-based”); 2) Filtragem Colaborativa (“collaborative filtering”); 3) Demográfica (“Demographic”); 4) Recomendação por Conhecimento de Domínio (“Knowledge-Based”); 5) Recomendação baseada na Comunidade (“Community-Based”) e 6) Sistemas de Recomendação Híbridos (“Hybrid Recommender Systems”). A seguir são apresentadas e detalhadas as categorias para sistemas de recomendação descritas por Ricci (2011):

1) **Recomendação Baseada no Conteúdo:** O sistema aprende a recomendar itens similares aos que o usuário já manifestou interesse previamente. A similaridade entre os itens é definida por critérios estabelecidos e atribuídos aos itens comparados.

2) **Filtragem Colaborativa**: O sistema recomenda ao usuário conteúdos que interessaram a outros usuários, preferencialmente os de características similares. O foco deixa de ser o conteúdo em si e passa a ser o perfil de quem acessa. A similaridade de gostos entre usuários é calculada com base em avaliações comuns entre eles, acessos a conteúdos parecidos e padrões de comportamento comparativamente próximos.
3) **Demográfica**: O conteúdo recomendado é baseado no perfil demográfico do usuário. É assumido que pessoas com características demográficas semelhantes (ex: mesma faixa etária, renda familiar similar ou moradores de uma mesma região) tendem a ter interesse em conteúdos similares.
4) **Recomendação por Conhecimento de Domínio**: A recomendação acontece a partir de um conhecimento prévio do domínio do conteúdo em questão, independente de uma similaridade clara entre os itens comparados ou do interesse prévio de outros usuários. Por exemplo, o sistema entenderia que alguém que compra espetos para churrasco poderia se interessar pela compra de carvão, por um entendimento prévio do domínio (em um churrasco, os dois itens são utilizados em conjunto), mesmo não havendo uma similaridade de características entre os itens isoladamente.
5) **Recomendação Baseada na Comunidade**: A recomendação é baseada no que amigos ou pessoas relacionadas ao usuário ativo manifestaram interesse prévio. Se um amigo de um usuário consome determinado conteúdo, entende-se que o mesmo tenha grande probabilidade de interessar ao usuário. Para esse tipo de recomendação, é necessário que o sistema tenha acesso a informações sobre pessoas que são amigas do usuário em questão.
6) **Sistemas de Recomendação Híbridos:** Sistemas que utilizam uma combinação de dois ou mais sistemas citados anteriormente. A combinação dos sistemas permite o aproveitamento das vantagens dos sistemas e o descarte de possíveis desvantagens que cada um deles apresente individualmente. Tendem a ser modelos mais complexos, mas comparativamente apresentam os melhores resultados de acerto na oferta de conteúdo relevante.

Em todas as propostas de categorização observa-se uma tentativa de consolidar o conhecimento acerca do funcionamento dos algoritmos do ponto de

vista técnico, de maneira que possa ser aplicada a todos os domínios. Entretanto, a diversidade entre as regras algorítmicas utilizadas para cada domínio específico é um dos grandes desafios para os estudos acerca do tema, em função da impossibilidade de se obter acesso ao código fonte dos sistemas de recomendação de conteúdo utilizados na grande maioria dos sites e produtos digitais - por serem de caráter privado e trazerem vantagens competitivas às respectivas empresas.

Nesse sentido, uma das principais técnicas para análise de algoritmos de sistemas de recomendação já existentes e implementados sem ter acesso ao código fonte é utilização da “Teoria da Engenharia Reversa”, proposta por Diakopoulos (2013). A aplicação desta técnica visa a compreensão das especificações de um sistema através de um exame rigoroso sobre o conhecimento do domínio, observação e dedução para descobrir um modelo de como esse sistema funciona. Neste caso, a avaliação e o entendimento dos algoritmos se dá a partir da observação rigorosa do ‘output’ (o item que é recomendado), e não pela análise do código fonte por trás da lógica de funcionamento desses elementos.

É importante salientar que por ser uma área de estudos bastante nova, tanto a implementação quanto a categorização desses sistemas de recomendação são campos de interesse elevado e abertos a constantes contribuições. Ao mesmo tempo que resolve uma série de questões, a utilização dos sistemas de recomendação automatizados acarreta em alguns problemas já catalogados e descritos na literatura sobre o tema. Adomavicius e Tuzhilin (2005) apontam para a relevância da pesquisa sobre o tema, a abrangência de sua aplicação e a necessidade de trabalhar em caminhos para a solução dos problemas em cada um dos domínios em que a prática da recomendação sistemática for aplicada.

> "O interesse nesta área (sistemas de recomendação) ainda é elevado porque constitui uma área de pesquisa rica em problemas e devido à abundância de aplicações práticas que ajudam os usuários a lidar com a sobrecarga de informações e fornecer recomendações, conteúdo e serviços personalizados". (ADOMAVICIUS E TUZHILIN, 2005)

#### 1.2.1.1 Desafios na utilização de sistemas de recomendação automatizados

Um dos primeiros desafios na utilização de sistemas de recomendação é a sua escalabilidade, ou seja, como mantê-los funcionando à medida que aumenta

cada vez mais a quantidade de dados que devem ser analisados pelos algoritmos. Ricci (2011) alerta para a necessidade de novas abordagens e estudos de avaliação de soluções para recomendação em sistemas com grande quantidade de dados:

> "À medida que a pesquisa sobre técnicas básicas progride e amadurece, torna-se claro que uma questão fundamental para os sistemas de recomendação é determinar como incorporar as técnicas básicas de recomendação em sistemas operacionais reais e como lidar com conjuntos maciços e dinâmicos de dados produzidos pelas interações de usuários com itens (classificações, preferências, comentários, etc.). Uma solução que funciona bem quando testada off-line em conjuntos de dados relativamente pequenos pode tornar-se ineficaz ou mesmo totalmente inaplicável em conjuntos de dados muito grandes. Novas abordagens e estudos de avaliação em grande escala são necessários (RICCI, 2011)

Jones (2013) ressalta ainda o problema da escalabilidade quando na utilização de sistemas de recomendação que funcionem em cenários de "tempo real", ou seja, sem margem de tempo para um demorado processo de tratamento dos dados:

> Um problema endêmico dos sistemas de recomendação em larga escala é a escalabilidade. Os algoritmos tradicionais funcionam bem com quantidades menores de dados, mas quando os conjuntos de dados crescem, os algoritmos tradicionais podem ter dificuldade em manter-se. Embora isso não seja um problema para o processamento off-line, são necessárias abordagens mais especializadas para cenários em tempo real. (JONES, 2013)

Outro fator de dificuldade referenciado na literatura, também de ordem tecnólogica e relacionado a quantidade de dados a disposição é a chamada "data sparsity" (na tradução, "dados esparsos"). Em resumo, o conceito de "data sparsity" diz respeito ao fato dos dados provenientes dos usuários tenderem a ser incompletos e desordenados.Tratando do tema, Guo (2012) afirma que nenhum produto detem o controle sobre as interações de seu usuário, ou sobre quais itens cada usuário vai avaliar ou interagir, e em qual ordem. Dessa maneira, nunca se tem garantias de um padrão de informações comum a todos os itens.

Ricci (2011) também aborda a questão dos dados esparsos como um fator que dificulta a capacidade de "prever" o que o usuário vai querer consumir no futuro, já que as análises são feitas a partir de dados incompletos e muitas vezes pouco consistentes. Por exemplo: em um serviço de filmes, por mais que um usuário goste de filmes de ação, ele não vai avaliar ou assistir a todos eles. Por mais que o sistema de recomendação possa inferir que o usuário gosta de filmes de ação, a recomendação de um filme de ação não avaliado pode não ser bem-sucedida (o usuário pode já ter assistido ao filme, ou não gostar de um ator em especial, etc).

Um terceiro problema de ordem tecnológica na previsibilidade dos algoritmos de sistemas de recomendação é o denominado "Cold Start" (na tradução literal, "início frio"). Este problema diz respeito a “dificuldade de inicializar os sistemas de recomendação para novos usuários ou novos itens”. (GUO, 2012). Como exemplo, um livro recém-publicado em um site de vendas de livros demora um determinado tempo inicial (o “início frio”) para receber avaliações ou interações. Dessa maneira, é necessário um período de tempo até que as ações de usuários possam fornecer os dados necessários para que as recomendações sejam feitas com maior assertividade.

Sobre este problema, Ricci (2016) aponta que o problema do “cold start” traz dificuldades especialmente em sistemas de recomendação que utilizam o modelo da filtragem colaborativa (onde dados do comportamento dos usuários são levados em conta para projetar as recomendações).

> “Um dos principais problemas nos sistemas de recomendação é que o número de informações inicialmente disponíveis é relativamente pequeno. Nesses casos, torna-se mais difícil aplicar modelos tradicionais de filtragem colaborativa”. (RICCI, 2016)

Além dos fatores tecnológicos, a utilização de sistemas de recomendação automatizados suscita em algumas discussões no campo da ética e da moral. Primeiramente, a questão do controle ou da (falta de) privacidade é uma questão ainda em aberto e amplamente discutida pelos autores da área. No momento em que um sistema se utiliza de informações e dados do usuário que muitas vezes não estão explicitamente determinados, poderia-se entender que a privacidade de cada indivíduo não está sendo respeitada.

Ricci (2011) aponta que o impacto na questão da privacidade é evidente, mas sustenta que a defesa da privacidade tenha que ser parte do projeto de soluções para recomendação de conteúdo, de maneira a garantir que dados sigilosos não sejam coletados nem utilizados por terceiros:

> “Os sistemas de recomendação exploram os dados dos usuários para gerar recomendações personalizadas. Na tentativa de criar recomendações cada vez melhores, eles coletam o máximo de dados de usuário possível. Isso terá claramente um impacto negativo na privacidade dos usuários e os usuários podem começar a sentir que o sistema sabe muito sobre suas preferências verdadeiras. Portanto, há uma necessidade de projetar soluções que usarão de forma parcimoniosa e sensata os dados do usuário. Ao mesmo tempo, essas soluções garantirão que o conhecimento sobre os usuários não possa ser acessado gratuitamente por usuários mal-intencionados. (RICCI 2011)

O autor aprofunda que a questão da privacidade deva ser amplamente discutida. Por um lado, ele indica a sensibilidade da questão, mas por outro entende que a limitação no acesso aos dados reais dos usuários certamente impediria o avanço dos algoritmos de recomendação:

> “A privacidade nos sistemas de recomendação baseia-se fortemente em informações e interações dos usuários, o que pode ser implícito ou explícito. Este feedback contém informações significativas sobre os interesses do usuário e pode revelar muito sobre suas opiniões políticas, orientações sexuais e preferências pessoais. Em muitos casos, essas informações podem ser altamente sensíveis, o que leva a problemas de privacidade. Tais preocupações de privacidade são significativas na medida em que impedem a liberação de dados necessários para o avanço dos algoritmos de recomendação. A disponibilidade de dados reais é crucial para avanços algorítmicos”. (RICCI, 2016)

Jones (2013) alerta ainda para a realidade dos algoritmos se aprofundarem tanto na privacidade dos usuários a ponto de se tornarem capazes de identificar padrões de comportamento desconhecidos pelos próprios usuários:

> Finalmente, considerações de proteção de privacidade também são um desafio. Os algoritmos de recomendação podem identificar padrões que os indivíduos podem nem saber. Um exemplo recente é o caso de uma grande empresa que poderia calcular um índice de previsão de gravidez com base em hábitos de compra. Através do uso de anúncios direcionados, um pai ficou surpreso ao saber que sua filha adolescente estava grávida. O preditor da empresa era tão preciso que poderia prever a data de vencimento de uma futura mãe com base em produtos que ela comprou. (JONES, 2013)

Não obstante as discussões acerca da privacidade, alguns autores amplificam a questão ética abordando a impossibilidade de existir uma “neutralidade” nas recomendações dos sistemas automatizados. A existência de um viés (consciente ou inconsciente) nas regras e lógicas algorítmicas que regem o funcionamento de um sistema de recomendação são apontadas como uma questão a ser discutida.

Diakopoulos (2013) questiona a dificuldade em estabelecer se os critérios definidos para um sistema de recomendação são de fato justos ou discriminatórios; se são politizados ou tendenciosos de alguma maneira negativa para os usuários. (DIAKOPOULOS, 2013)

Enquanto muitas vezes a visão editorial de um produto de notícias tende a ser mais facilmente percebida pelos leitores e usuários, a lógica algorítmica transmite uma falsa sensação de neutralidade. Para Diakopoulos (2013), esta visão deve ser prontamente desconsiderada nos estudos e pesquisas sobre o tema, uma vez que

sistemas de recomendação e suas lógicas algorítmicas são originalmente projetados pelo homem:

> "... deve ficar claro que existem várias influências do ser humano incorporadas em algoritmos, tais como escolhas de critérios, dados de treinamento, semântica e interpretação. Qualquer investigação deve, portanto, considerar os algoritmos como objetos de criação humana e levar em consideração a intenção, inclusive a de qualquer grupo ou processo institucional que possa ter influenciado seu design ". (DIAKOPOULOS, 2013)

A impossibilidade de um sistema de recomendação verdadeiramente isento e neutro em sua lógica conduzem a um inevitável viés nos itens e conteúdos recomendados para cada usuário. Sobre isso, Jolly (2014) comenta sobre os algoritmos acabarem direcionando o usuário somente para os mesmos assuntos e conteúdos com os quais o usuário já manifestou algum grau de concordância:

> "Algoritmos tornam muito mais fácil não só para você encontrar o conteúdo que lhe interessa, mas para que um conteúdo que o algoritmo pensa que vai interessar lhe encontre. Ou seja, eles maximizam os cliques excluindo outros tipos de conteúdo, ajudando a reforçar uma visão de mundo existente, diminuindo a chance de um leitor de encontrar conteúdo fora do que eles já conhecem e acreditam. (JOLLY, 2014)

Tal condição leva ao efeito conhecido como “filtro bolha” – termo apresentado em 2011 por Eli Parisier em seu livro “*O filtro invisível*: o que a internet está escondendo de você” (2012). Parisier define o filtro bolha como “uma característica da web de apresentar apenas o que é relevante para o usuário, mas que leva à formação de filtros que impedem os usuários de ter contato espontâneo com informações classificadas como menos relevantes pelos algoritmos”.

Além de todos os desafios e problemas característicos da utilização de sistemas automatizados de recomendação nos mais diversos domínios de conhecimento, existem ainda o que Ricci (2016) chamou de “desafios específicos de domínio”, ou seja, questões que não necessariamente se aplicam a sistemas de recomendação em qualquer área, mas especificamente em um determinado domínio. A seguir, serão discutidos aspectos referentes aos sistemas de recomendação no domínio da notícia e seus respectivos desafios específicos.

#### 1.2.1.2 Recomendação de conteúdo em sites de notícias

Nos primeiros sites de notícias, a recomendação de conteúdo se dava apenas de maneira estática: os mesmos conteúdos eram recomendados igualmente para todos os usuários daquele produto. Neste modelo, segundo Hendrix (2014) a curadoria do conteúdo que é recomendado sempre foi de responsabilidade do corpo editorial. Desconsiderando-se o aspecto mercadológico (recomendação de notícias pagas ou patrocinadas por empresas), toda a recomendação de conteúdo se dava baseada em seu grau de relevância para o público na avaliação do corpo editorial do site.

Entretanto, apesar da tentativa por parte do corpo editorial em manter o controle sobre o que deveria ou não ser acessado pelos usuários, as possibilidades tecnológicas da internet como ambiente de consumo apontavam para outros caminhos. Nesse sentido, Stray (2012) defende a ideia de que nem todos os usuários querem as mesmas histórias, e que o modelo de padronização da recomendação deveria ser abandonado pelos sites de notícias:

> "E se abandonarmos a idéia de que todos vejam as mesmas histórias? Essa foi uma limitação tecnológica pré-Internet, e talvez tenhamos deixado o que era possível tornar-se o que é certo. Quero reconhecer que cada pessoa não só tem interesses únicos, mas também é afetada de maneira exclusiva por eventos maiores e tem uma capacidade única de interagir". (STRAY, 2012)

Da mesma maneira, os modelos de recomendação estática de conteúdo resultavam em baixas taxas de engajamento por parte dos usuários. Mais uma vez, Stray (2010) critica a entrega das mesmas informações para diferentes pessoas, sem levar em conta as individualidades:

> “Entregar a mesma informação para todos e esperar que todos tenham engajamento de forma equivalente não é apenas um insulto à diversidade, é ignorar o fato da nossa civilização ser construída com base na especialização. Diferentes pessoas têm diferentes capacidades, posições e paixões, e assim eles vão precisar de diferentes agendas de notícias. Jornalistas empáticos sempre foram conscientes disso, mas antes não era possível segmentar histórias individuais para pessoas específicas”. (STRAY, 2010)

Aliado a isso, o crescimento exponencial no volume e na variedade de conteúdos publicados nos sites de notícias tornou cada vez mais difícil e menos eficiente a realização dessa curadoria de forma manual ou não-automatizada. Em função disso, a tarefa de apresentar conteúdos verdadeiramente relevantes para o

usuário passa a se apresentar como uma questão de cada vez maior complexidade para a capacidade humana, sem o auxílio de sistemas ou algoritmos:

> Como, então, um editor deve escolher qual parte minúscula do mundo ele deve nos mostrar? Às vezes, há um evento tão grande, tão universal, que exige atenção. Os desastres naturais e as revoluções vêm à mente. Para todas as outras histórias, não acho que haja uma resposta. Não podemos nem concordar sobre quais problemas são importantes. Nenhum conjunto único de manchetes pode representar fielmente tudo o que importa no mundo. (STRAY, 2012)

Assim, a chegada dos sistemas de recomendação automatizados aos sites de notícias, apesar de tardia em comparação a outros domínos, se mostrou inevitável. Conforme apresentado anteriormente, empresas de diferentes áreas, como Amazon, Spotify e Netflix, já faziam uso de sistemas de recomendação automatizada há algum tempo, buscando oferecer conteúdo de maneira precisa para seus usuários e consequentemente ampliar suas receitas. Para se ter uma ideia da eficiência desses mecanismos nos domínios citados, no Netflix cerca de 75% da atividade dos usuários se dá a partir das recomendações feitas por esses sistemas (HENDRIX, 2014).

Neste momento, a questão passa a ser então como implementar um modelo similar de recomendação de conteúdo no âmbito das notícias, de modo a definir o que é ou não relevante para cada usuário e personalizar ao máximo a oferta de notícias. Stray (2012) novamente aponta para a dificuldade em resolver essa questão nos sites de notícias.

> “Se nem todas as pessoas querem ler as mesmas notícias ao mesmo tempo, então a questão se torna: quem deve ver o que? É uma pergunta difícil. É uma questão de escolha editorial, de design de filtros, de que tipo de discussão cívica teremos. É a questão básica que enfrentamos quando abraçamos a capacidade da rede de fornecer informações personalizadas”. (STRAY, 2012)

No âmbito da notícia, um dos primeiros sites a utilizar sistemas de recomendação foi o norte-americano “Google News”, ao implementar um modelo de recomendação baseada nos conteúdos (“content-based”) já no final da primeira década dos anos 2000. Neste modelo, as páginas de conteúdo do site passaram a contar com um elemento que exibia links para outras matérias que fossem da mesma categoria do conteúdo original da página.

Em relação a esses elementos, não há um consenso na literatura quanto a uma possível nomenclatura. Ricci (2011) se refere a eles como “sugestão de itens a

serem usados por um usuário"; Stray (2010) utiliza "caixas de conteúdos recomendados", enquanto Hendrix (2014) trata apenas como "conteúdo recomendado". Por conta disso, será proposto nessa tese o uso da nomenclatura "elementos de recomendação de conteúdo", utilizada em Lohmann (2013). Como elemento de recomendação de conteúdo, entende-se os componentes das páginas de conteúdo de um site (noticioso ou não) a partir dos quais são apresentados conteúdos deste ou de outro site que possam ser diretamente acessados pelo usuário.

Na figura 6, é apresentado um exemplo de elemento de recomendação de conteúdo em páginas de sites de notícias. A figura representa uma matéria do site G1 (g1.com.br), com os elementos de recomendação destacados em amarelo e ampliados para melhor visualização.

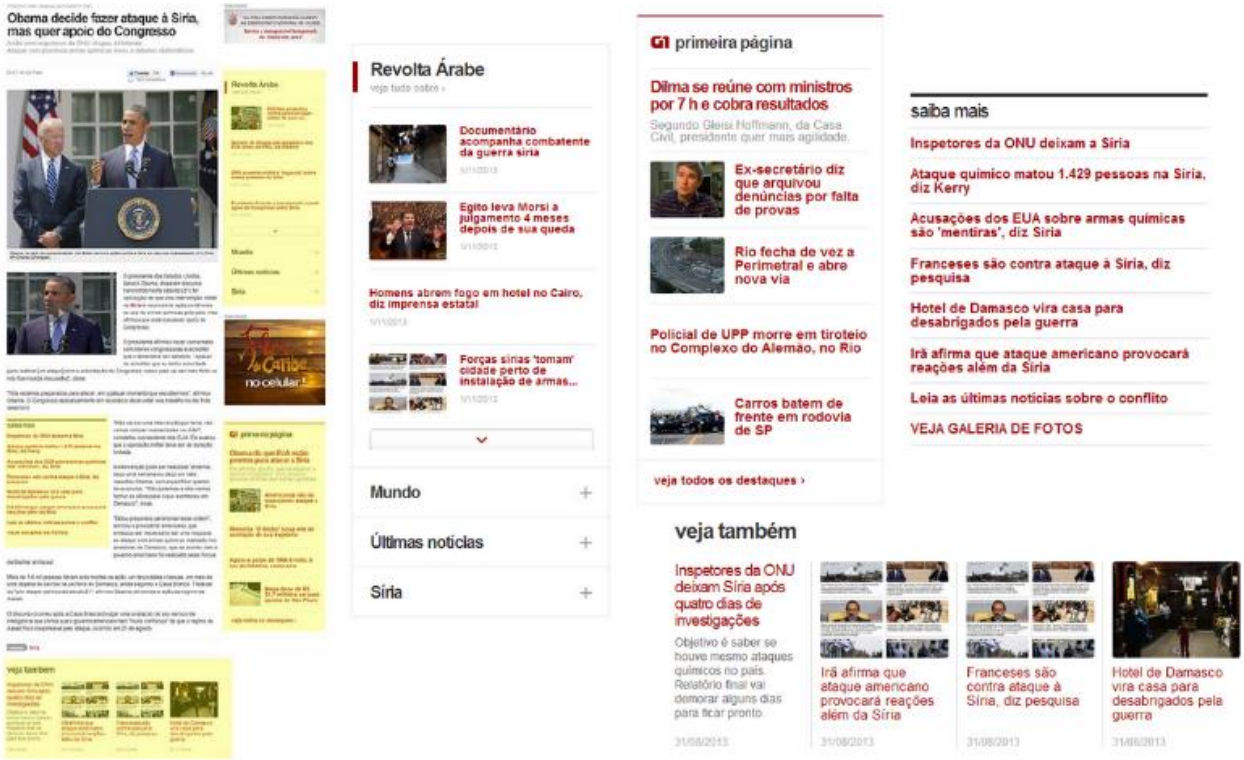


Figura 6 - Exemplo de matéria do site G1 (g1.com.br) e dos elementos para consumo de novos conteúdos.

Entretanto, no âmbito da notícia, apesar da melhora em relação a recomendação editorial manual ou estática, os elementos de recomendação de conteúdo alimentados por sistemas automatizados continuam sem apresentar os

mesmos resultados (do ponto de vista do engajamento dos usuários) observados em outros domínios:

> "Os sistemas de recomendação foram usados com relativo sucesso em produtos como livros e filmes, mas surpreendentemente descobriu-se até o momento pouca aplicação na recomendação de artigos de notícias". (GARCIN et all, 2016)

Tal insucesso até o momento se deve sobretudo aos desafios na utilização de sistemas automatizados de recomendação apresentados no subcapítulo anterior, a saber: escalabilidade; "data sparcity" (dados esparsos); "cold start" (início frio); privacidade/questões éticas e os "filtros bolha". No âmbito de sua implementação em sites de notícia, tais desafios não apenas se fazem presentes como são profundamente amplificados a partir de situações específicas inerentes ao domínio da notícia.

No que diz respeito ao desafio da escalabilidade, conforme mencionado em capítulos anteriores, a quantidade de conteúdo noticioso produzido e/ou consumido é cada vez maior e mais complexo, bem como os caminhos ou fluxos de navegação experimentados por cada usuário. Ao contrário de domínios com uma quantidade delimitada e pouco variável de itens (um site de venda de livros, por exemplo) a produção de novas notícias é constante e diária. Da mesma maneira, o interesse dos usuários por notícias é algo mais fluido e variável do que o gosto por filmes ou livros, como indica Hendrix (2014):

> "O interesse por notícias é variável (a mesma pessoa pode se interessar muito por futebol durante a Copa do Mundo mas no resto do tempo não ter interesse em esportes), ao passo que o gosto por música e filmes tende a ser muito mais homogêneo. Isso faz com que o investimento (financeiro e intelectual) em sistemas de recomendação de notícias tenha que ser elevado, o que dificulta sua implementação". (HENDRIX, 2014)

Sendo assim, torna-se ainda mais complexo no âmbito das notícias tornar escalável um modelo que funcione bem com uma quantidade reduzida de dados e manter sua assertividade elevada em um cenário real, com tamanha quantidade e variedade de dados inerentes ao meio.

Da mesma forma, os dados de consumo de notícias são esparsos (data sparsity), uma vez que ao contrário de outros domínios, nos sites de notícias nem sempre se exige que a navegação do usuário aconteça logada (de tal maneira que seria possível capturar maiores informações deste usuário). De um modo geral,

entende-se que nesse contexto o processo de cadastro muitas vezes funciona como uma barreira que supera o próprio interesse em uma notícia em específico, resultando no abandono do site ou na queda ainda maior no engajamento. Garcin (2016) aponta que com acessos cada vez mais acontecendo via buscadores e redes sociais, temos um cenário onde os usuários chegam praticamente ‘anônimos’ nos sites de notícias e muitas vezes apenas os dados atuais da visita podem ser trabalhados para fazer recomendações.

Hendrix (2014) ressalta que além da atividade “anônima”, outro fator que amplia o desafio dos dados esparsos na recomendação de conteúdo é o fato dos usuários em geral consumirem notícias de múltiplas fontes. Assim, um site específico teria dificuldades em saber quais assuntos já foram acessados em outros sites.

> “Para construir um bom sistema de recomendação é necessário ter uma grande quantidade de dados a respeito dos seus usuários. Ao contrário do contexto das empresas citadas anteriormente, a relação dos usuários com as empresas noticiosas é um pouco diferente. Normalmente a navegação não se dá de maneira logada, e os usuários tendem a consumir notícias de múltiplas fontes, não apenas de um único site”. (HENDRIX, 2014)

Outro desafio que se agrava no domínio das notícias é o chamado “Cold Start” ou “início frio”. Enquanto em outros domínios o fator cronológico pode ter menor relevância, nos sites de notícias são as chamadas “notícias quentes” (as mais recentes) que em geral tendem a ser mais relevantes para o usuário. Entretanto, quanto mais recente é uma notícia, menos se tem dados sobre ela. O desafio se apresenta quando a quantidade de informações necessárias para recomendar um determinado conteúdo só é atingida quando este já passou a ter pouca ou nenhuma relevância para os usuários. Garcin (2016) indica esta como outra das razões que ajudariam a explicar o insucesso dos sistemas de recomendação de notícias atuais, uma vez que técnicas baseadas no comportamento do usuário se apresentam complexas de serem aplicadas.

> “Artigos de notícias devem ser recomendados logo após escritos, fazendo com que haja pouco tempo para coletar dados de acesso e sua popularidade. Assim, técnicas para recomendações baseadas no comportamento do usuário foram consideradas difíceis de aplicar. Muitas vezes, há poucos dados disponíveis sobre o comportamento passado de um usuário dentro daquele site”. (GARCIN et all, 2016)

Enquanto a questão da (falta) de privacidade dos sistemas de recomendação também se mostra presente nos sites de notícias (afinal é necessário rastrear o máximo de informações daquele usuário, ou que tipos de conteúdo ele costuma acessar), o desafio de maior gravidade do ponto de vista ético se dá na ausência de neutralidade das lógicas algorítmicas e no viés que a recomendação de conteúdo apresenta. Diakopoulos (2013) aponta que a lógica editorial pode se tornar ainda mais sutil e passar despercebida por meio dos algoritmos de recomendação, ao mostrar indícios de que uma simples decisão de exibir uma maior quantidade de notícias de um determinado tema poderia influenciar a decisão de um processo eleitoral:

> "Os jornais sempre conseguiram ter uma voz editorial e possivelmente afetar os padrões em uma eleição, por exemplo, com base em seu viés editorial. Mas o que estamos vendo agora é a capacidade de escalar uma população de uma maneira muito mais poderosa. O Facebook recentemente fez um estudo que descobriu que simplesmente mostrar mais notícias de um determinado lado do que de outro no fluxo de notícias afetaria decisões de votação" (DIAKOPOULOS, 2013)

Segundo Jolly (2014), tal efeito, denominado "filtro bolha" e já citado no subcapítulo anterior, vem se agravando de maneira perigosa a partir do avanço na utilização de sistemas de recomendação em sites de notícias:

> "O filtro bolha é o mundo criado pela mudança de "curadores humanos" - como editores de jornais que determinavam pelo grau de importância as notícias da primeira página – para algoritmos que apresentam o conteúdo que eles acreditam ser mais provável de interessar ao usuário. Este novo universo digital é um lugar aconchegante, povoado por nossas pessoas e idéias favoritas. Mas, finalmente, é perigoso". (JOLLY, 2014)

Como se observa, tais desafios característicos dos sistemas de recomendação em geral são amplificados pela complexidade do comportamento dos usuários no consumo de conteúdo de notícias, auxiliando a compreender em parte as dificuldades para definir critérios para a recomendação. Por conta disso, os modelos convencionais de "recomendação baseada em conteúdo" ou mesmo os modelos híbridos que funcionam bem em outras áreas vem mostrando resultados ruins no âmbito das notícias, o que é representado pela queda no engajamento de leitores e usuários.

> "Ao contrário do senso comum, identificamos em nossas pesquisas que recomendações baseadas em conteúdo ou modelos híbridos tem um desempenho surpreendentemente fraco em sites de notícias. Nós explicamos isso pelo fato de

> que os usuários não gostam de ler várias histórias sobre o mesmo tópico: um usuário lê em média 7 assuntos diferentes a cada 10 notícias". (GARCIN et all, 2016)

Por tudo isso, o estudo a respeito dos sistemas de recomendação em sites de notícias e dos elementos de recomendação de conteúdo constituem um desafio para diversas áreas de conhecimento, tais como computação, sistemas de informação, comunicação e design. No momento em que os modelos de recomendação automatizada não apresentam o resultado esperado, é necessário propor novas soluções para o problema na tentativa de estimular o usuário a se manter engajado dentro do produto noticioso a partir do consumo estimulado de novos conteúdos.

Figura 7 – Quadro resumo dos principais problemas e desafios no design de soluções para a recomendação de conteúdo de notícias, com a inclusão de um 6º elemento, o fator contextual, que será apresentado a seguir.

### 1.2.3 Sistemas de recomendação contextual

Além dos desafios comuns a grande parte dos domínios na implementação de sistemas de recomendação automatizados, Ricci (2016) reforça a necessidade de se aprofundar, para cada domínio, no que o autor chama de "desafios específicos". Para Ricci (2016), em diferentes domínios, o contexto em que cada recomendação acontece é fator chave para o desenvolvimento de sistemas automatizados mais

assertivos. Ricci apresenta, portanto, o conceito de “sistemas de recomendação contextual”.

> “Desafios específicos de domínio em sistemas de recomendação: em diferentes domínios, o contexto da recomendação desempenha um papel crítico. Portanto, a noção de sistemas de recomendação contextual foi desenvolvida para abordar a informação adicional que surge nesses domínios. Essa noção é usada com diferentes modificações para vários tipos de dados, como dados temporais, dados de localização ou dados sociais”. (RICCI, 2016)

Tal noção de contexto pode ser relacionada ao que David (2014) apresenta como “produto contextual”, no qual se tenta antecipar os desejos do usuário a partir do entendimento de determinadas condições contextuais:

> **“**Um produto contextual entende a história completa em torno de uma experiência humana, para trazer aos usuários exatamente o que eles querem, com uma interação mínima**.** Produtos contextuais são sobre prever o futuro. Os engenheiros de software trabalham para ensinar-lhes a analisar determinadas condições do usuário, o ambiente local e as circunstâncias externas, para prever o conjunto de ações que o usuário deseja realizar”. (DAVID, 2014)

Para se projetar produtos levando em conta o contexto, David (2014) aponta para a necessidade de subdividir o tema em “3 esferas do contexto”: 1) contexto do usuário; 2) contexto ambiental; 3) contexto mundial, a saber:

1) O “contexto do usuário” diz respeito às diferenças interpessoais, ou seja, tudo aquilo que cada usuário gosta, odeia, possui (por exemplo, os dispositivos), instala (por exemplo, os aplicativos), etc. Esta esfera do contexto trata ainda do “estado de espírito” de cada usuário (entediado, atrasado, faminto, etc.), seus hábitos e seu estado atual de movimento (se está em pé, caminhando, correndo, etc). Variações em um ou mais desses aspectos significaria diferentes soluções para atender melhor a cada usuário. Como exemplos, uma pessoa com fome tenderia a ser mais impactada por um conteúdo sobre comida. Da mesma forma, uma pessoa utilizando o smartphone enquanto caminha pela rua tende a se comportar de maneira diferente em sua navegação a um usuário em seu notebook de casa.
2) O “contexto ambiental” diz respeito a fatores como: condições climáticas, dia da semana, a localização, o tipo de local em que o usuário se encontra (se em casa, trabalho, loja, ponto de ônibus, etc.), ou qualquer outro aspecto físico externo que possa exercer influência em seu comportamento. Da

mesma maneira que o contexto do usuário, variações no contexto ambiental afetam a maneira como um usuário deve se relacionar com um produto ou experiência de uso. Como exemplos, em períodos frios o usuário se sentiria mais interessado em conteúdos sobre roupas de inverno, ou usuários em horário de trabalho tenderiam a evitar conteúdos associados a momentos de lazer.

3) O “contexto mundial” trata de algo que está acontecendo em outro lugar e que possa estar relacionado ao usuário, como eventos esportivos, eventos de notícias, clima, atrasos de voos, trânsito, horários de entrega de encomendas, programas de TV que todos estão assistindo ou o assunto do momento nas redes sociais. Da mesma forma que os contextos anteriores, o contexto mundial interfere no interesse e no comportamento dos usuários. Como exemplo, um usuário tende a se interessar mais por política próximo ao período eleitoral, ou sobre como é o trânsito em Buenos Aires quando está com viagem marcada para o país.

Ressaltando a relevância dos aspectos contextuais na recomendação de conteúdo, em especial em produtos de notícias, Stray (2010) destaca que oferecer o conteúdo correto na hora certa (contexto) seria mais importante do que produzir bom conteúdo:

> “Eu não quero o produto com o melhor conteúdo em geral, eu quero o produto que vai me oferecer o melhor conteúdo a cada vez que eu acessá-lo. (STRAY, 2010a)

Rampton (2014) reforça a necessidade de que cada vez mais o conteúdo deva chegar de maneira contextual e personalizada para os usuários:

> “Normalmente, nossa navegação por essa infinidade de sites e páginas acontecia através de buscas, sugestões ou compartilhamentos de pessoas em nossas redes. Isso exigia ou que você soubesse pelo que estava procurando, ou que você confiasse que as outras pessoas iriam encontrar o conteúdo para você. Entretanto, com o passar dos anos e o advento de novas plataformas, nossa forma de descobrir conteúdos se modificou. Ao invés de um modelo baseado em uma ação ativa do usuário em buscar por algo específico, hoje a descoberta de conteúdos é automatizada, personalizada e muito mais contextual do que antes, e continua em processo de evolução”. (RAMPTON, 2014)

Por conta disso, Ricci (2016) aprofunda a questão da influência do contexto ao apontar alguns dos principais fatores contextuais que interferem na eficácia dos sistemas de recomendação.

> “Os sistemas de recomendação baseados em contexto ou com reconhecimento de contexto levam em consideração vários tipos de informações contextuais ao fazer recomendações. Tais informações contextuais podem incluir tempo, localização ou dados sociais. Por exemplo, os tipos de roupas recomendados por um revendedor podem depender da estação e da localização do cliente. Outro exemplo é o caso em que um determinado tipo de festival ou feriado afeta a atividade do cliente subjacente. Em geral, observou-se que o uso de tais informações contextuais pode melhorar consideravelmente a eficácia do processo de recomendação. Os sistemas de recomendação baseados em contexto são incrivelmente poderosos porque as ideias subjacentes são relevantes para uma ampla variedade de configurações específicas de domínio”. (RICCI, 2016)

Dessa maneira, Ricci (2016) apresenta os três principais tipos de sistemas de recomendação contextuais que vem sendo utilizados na prática em diversos domínios: 1) sistemas de recomendação sensíveis ao tempo; 2) sistemas de recomendação baseados em localização; 3) sistemas de recomendação social.

1) Sistemas de recomendação sensíveis ao tempo (“Time-Sensitive Recommender Systems”): Em muitos domínios, as recomendações para um item podem evoluir com o tempo. Por exemplo, no domínio da notícia, um conteúdo sobre um novo lançamento tecnológico pode ser de extrema relevância nos primeiros dias, mas a medida que novos lançamentos surgem ela vai perdendo importância. Enquanto isso, uma notícia sobre os benefícios da vitamina C poderia ter o mesmo nível de relevância por um período de tempo mais longo. Nesses casos, seria extremamente importante incorporar o conhecimento temporal/cronológico no processo de recomendação.
2) Sistemas de recomendação baseados em localização (“Location-Based Recommender Systems ”): Com a crescente popularização dos dispositivos móveis com GPS, os usuários geralmente estão interessados em recomendações baseadas em localização. Por exemplo, um usuário viajando pode querer informações sobre o trânsito local, sobre restaurantes naquela região, ou sobre as condições climáticas. Após voltar da viagem, tais informações passam a ter pouca ou nenhuma relevância para o mesmo usuário.

3) Sistemas de recomendação social (“Social Recommender Systems”): Os sistemas de recomendação social são baseados na combinação de diferentes aspectos das redes sociais: conexões pessoais, hashtags, assuntos do momento, núemro de curtidas e comentários, etc. Por exemplo, se um conteúdo está sendo amplamente abordado em uma rede social, ele passa a ter uma provável relevância para ser recomendado para outros usuários.

Entretanto, apesar de apresentar melhores resultados em sua aplicação, os sistemas de recomendação que levam em conta questões contextuais dos usuários ainda são pouco utilizados (ou sequer o são) nos sites de notícias brasileiros. A inclusão de fatores contextuais nos critérios de recomendação de conteúdo noticioso apresenta-se como um dos principais caminhos para os sites de notícias brasileiros na tentativa de ampliar suas métricas de engajamento.

# 2 **METODOLOGIA DE PESQUISA**

Após o levantamento teórico/conceitual a partir de referencial bibliográfico acerca da temática da recomendação de conteúdo (no âmbito geral e mais especificamente no domínio dos sites de notícias), são apresentados os procedimentos metodológicos utilizados para o levantamento sobre o estado atual da recomendação de conteúdo em sites de notícias brasileiros, a fim de propor uma categorização para estes sistemas de recomendação.

Para a validação das hipóteses apresentadas anteriormente, são apresentados ainda os procedimentos para análise de métricas de sites de notícias, bem como para a pesquisa com usuários.

## 2.1 **Levantamento de dados para categorização da recomendação de conteúdo nos sites de notícias brasileiros**

Para analisar o estado atual da utilização de sistemas de recomendação de conteúdo nos sites de notícias brasileiros, foram selecionados para análise 8 sites de

notícias - os de maior audiência (fonte: Alexa 2016) - a saber: G1, O Globo, Terra, Uol, Folha de São Paulo, R7, Estadão e Yahoo Notícias. Dessa maneira, entende-se que apesar de não atingir a totalidade dos sites de notícias do Brasil, a pesquisa abrangeu uma quantidade de amostras suficiente para a identificação de grande parte dos padrões de recomendação de conteúdo vigentes nos sites atuais.

Por conta da impossibilidade de ter acesso aos códigos fonte dos sistemas de recomendação de conteúdo utilizados pelos sites de notícia - por serem de caráter privado e trazerem vantagens competitivas às respectivas empresas, optou-se por utilizar a técnica da "Engenharia Reversa", proposta por Diakopoulos (2013) e específica para analise de algoritmos.

Na técnica, a avaliação dos sistemas de recomendação se dá a partir da observação rigorosa daquilo que está visível para todos que acessam os sites de notícias - os conteúdos que são recomendados nas páginas de matéria dos sites.

A partir da observação, aprofunda-se no entendimento das regras funcionais dos elementos analisados (que conteúdos estão sendo recomendados; critérios para ordenação e escolha desses conteúdos; em que momentos eles são recomendados, etc).

Para estruturar e padronizar o resultado da análise, utilizou-se como referência o conceito das quatro decisões atômicas algorítmicas proposto por Diakopoulos (2013) e descrito na fundamentação teórica desta tese, a saber: 1) priorização; 2) classificação; 3) associação e 4) filtragem.

A partir dos padrões identificados nesta abordagem, propõe-se a categorização dos principais critérios para recomendação de conteúdo identificados nos sites de notícias.

## 2.2 **Análise de métricas de navegação: site O Globo**

Para a validação das duas primeiras hipóteses apresentadas nesta tese, realizou-se a análise das métricas de um mesmo produto de notícias brasileiro em específico, o site do jornal O Globo (www.oglobo.com.br). Na realização da análise, utilizou-se o produto "Google Analytics" (analytics.google.com), ferramenta gratuita e principal referência para monitoramento e análise de métricas de sites.

Todas as métricas analisadas são referentes apenas a versão web/desktop do site, não contemplando a versão para dispositivos móveis (por questões técnicas, a versão para dispositivos móveis não estava configurada para gerar métricas na ferramenta citada). Dessa maneira, os resultados apresentados poderão até trazer indícios a respeito do comportamento dos usuários nos demais dispositivos, mas apenas podem ser considerados conclusivos no âmbito do acesso por plataformas web/desktop.

Para a validação da primeira hipótese, será feito um comparativo das chamadas "métricas de engajamento" de duas versões do site do jornal O Globo em períodos distintos: a primeira versão analisada é anterior ao redesenho das páginas de matéria do site, ocorrido em 31/05/2014; já a segunda versão é a nova página de matéria, lançada após a referida data. Neste processo de redesenho, entre outras mudanças no design das páginas, foram feitas alterações nos critérios e nos elementos de recomendação de conteúdo das páginas.

O comparativo apresentado trará as métricas de engajamento das versões antiga e nova, de maneira a avaliar se alterações de design realizadas nas páginas de matéria são efetivamente capazes de influenciar tais métricas (quer positiva, quer negativamente).

Foram selecionadas especificamente as três métricas diretamente ligadas ao engajamento dos usuários:

1) "Taxa de rejeição": Percentual de usuários que abandonam o site após acessar uma única página;
2) "Páginas por sessão": Média do número de páginas visitadas em cada sessão de um mesmo usuário;
3) "Duração média da navegação por sessão": Tempo total que o usuário permanece no site em uma mesma sessão.

O recorte de tempo selecionado para a captura das métricas foi de um período de dois meses: para a versão antiga da página (anterior ao redesenho) foi selecionado o período de 15.03.2014 até 15.05.2014; para a nova versão da página de matéria (após o redesenho), foi selecionado o período equivalente, porém exatamente um ano depois (15.03.2015 até 15.05.2015). A definição por um período de dois meses foi considerada suficiente para sublimar possíveis fatores pontuais

que viessem a distorcer as métricas analisadas (ao comparar as métricas da página de matéria antiga em diferentes períodos de dois meses de duração, a margem de oscilação ficou abaixo de 3% em todos os casos). A escolha por utilizar exatamente o mesmo período do ano foi o de evitar possíveis influências sazonais de distorção nas métricas (eventos como Copa do Mundo ou datas comemorativas como Natal, Carnaval, etc).

Em ambos os cenários, foram filtrados somente os acessos iniciados tendo como página de destino um conteúdo, ou seja, foram descartadas as sessões iniciadas a partir de todas as homepages do site.

É importante ressaltar que o procedimento metodológico proposto tem por objetivo apenas a validação da hipótese de que a alteração no design da página de matéria em um site de notícias (incluindo mudanças nos elementos de recomendação de conteúdo - seja no layout, seja na usabilidade, seja no critério da oferta) influenciariam nas métricas de engajamento deste site, o que será confirmado em caso de variação dessas métricas nos dois cenários propostos. Não há o intuito de, nesse momento, afirmar que a nova versão seja melhor ou pior do que a versão antiga, nem de determinar com precisão quais dos fatores alterados promoveu esse impacto.

Em relação a validação da segunda hipótese, a análise será dividida e apresentada em duas partes. Na primeira parte, objetivou-se verificar se acessos que tem um conteúdo como “página de destino” resultam em um engajamento pior do que os acessos que tem as homepages como “página de destino” em sites de notícias. Tal confirmação serviria para corroborar a ideia de que as páginas de conteúdo atuais (no caso específico, do site do jornal O Globo) são pouco estimulantes no que diz respeito a manter o usuário engajado na utilização do site de notícias.

Para isso, foi feito um novo comparativo, desta vez entre os acessos iniciados a partir das homepages e os acessos iniciados por páginas de matérias da versão nova do site (após o redesenho citado anteriormente, ocorrido em 31/05/2014). Para o levantamento das métricas, foi selecionado também um período de 2 meses (15.03.2015 até 15.05.2015), com o mesmo objetivo de minimizar distorções pontuais. Com isso, foi possível estabelecer uma comparação entre acessos iniciados a partir de grupos de páginas de destino diferentes (conteúdos e

homepages), e suas respectivas diferenças do ponto de vista do engajamento dos usuários.

Para continuar a validação da segunda hipótese, será avaliado se a “origem do tráfego” (ou seja, de qual site ou serviço o usuário foi direcionado no momento do primeiro acesso ao site de notícias) teria influência no engajamento dos usuários. Para esta parte do estudo, foram selecionadas e analisadas as métricas de engajamento do site sob o recorte da variável “origem do tráfego” do site do jornal O Globo. A seguir, os resultados serão comparados com a análise proposta anteriormente, em relação a variável “páginas de destino”. Tal análise teve por objetivo avaliar se a origem do tráfego seria um fator de relevância nas métricas de engajamento dos usuários, ou se estes usuários vindos de origens externas como redes sociais ou buscadores teriam baixo engajamento em função da página de destino ser, em sua maioria, uma página de conteúdo (conforme validado na análise anterior).

Dessa forma, é possível avaliar o quanto o site ou a ferramenta utilizada no momento imediatamente anterior ao acesso (ou seja, a origem do tráfego) impacta na sequência da experiência de uso no site de notícias, bem como o quanto é a própria experiência oferecida pelo site (página de destino) o fator que promove diferentes índices de engajamento.

Para essa etapa do estudo, inicialmente foi feito um levantamento e análise das métricas de utilização do site d’O Globo em um período de 2 meses (15.03.2015 até 15.05.2015). Tal período novamente foi definido visando evitar possíveis distorções/variações específicas. As métricas selecionadas para a análise foram novamente as três definidas nas duas etapas anteriores: “taxa de rejeição”, “páginas por sessão” e “duração média por sessão”.

Após a definição do recorte de tempo e a seleção das métricas, as variáveis “origem do tráfego” e “página de destino” foram utilizadas na segmentação em múltiplos cenários que permitissem a avaliação de seu impacto nas métricas de engajamento. Como “origem do tráfego”, foram selecionadas as 4 origens mais representativas no contexto do site d’O Globo: 1) Globo.com; 2) Google; 3) Facebook e 4) Tráfego Direto. Como “página de destino”, os resultados foram agrupados em 2 tipos: homepages ou páginas de conteúdo (matérias).

## 2.3 **Pesquisa com usuários: questionário**

Para a validação da terceira e última hipótese apresentada nesta tese, que trata da influência de questões contextuais sobre o engajamento dos usuários com sites de notícias, serão apresentados os resultados de uma técnica de pesquisa estruturada em formato de questionário realizada com usuários de sites de notícias.

No questionário, serão inicialmente solicitadas algumas poucas informações pessoais somente para identificação do respondente. Na sequência, serão apresentadas cinco frases escritas em padrão de título ou “manchete” de matéria jornalística, sobre cinco temas/assuntos distintos. Para cada uma dessas manchetes será solicitado que o usuário avalie o seu grau de interesse em consumir esse tipo de conteúdo (em escala de 1 a 5, sendo 1 = nenhum interesse e 5 = muito interesse). A seguir, para cada uma das manchetes serão propostos três cenários possíveis referentes ao contexto do respondente: “Em Casa”, “No Trabalho” e “Em uma viagem”. Será solicitado que o respondente avalie em quais dos três cenários ele teria maior probabilidade de consumir cada uma das cinco notícias (deixando claro que ele pode marcar quantas opções quiser, inclusive “nenhuma”).

O questionário será aplicado de maneira online via redes sociais, sem um recorte específico de público. O objetivo da pesquisa é avaliar, coletiva e individualmente cada uma das respostas e verificar se, na avaliação geral, existiriam diferenças na probabilidade de consumir cada uma das notícias em função da mudança do contexto proposto. Além disso, dentre o grupo de respondentes, será avaliado qual o percentual de pessoas que admitem a influência do aspecto contextual (momento em que acessa uma notícia) em seus hábitos de consumo de conteúdos. Para isso, as respostas serão avaliadas a partir de três aspectos:

1) Para as cinco notícias apresentadas, será avaliado coletivamente o percentual obtido para cada um dos três cenários contextuais propostos: “em casa”, “no trabalho” e “em uma viagem”. Caso alguma das notícias apresente ao final da pesquisa um percentual equivalente ou muito parecido entre eles, será considerado que esta notícia é pouco ou nada impactada pela mudança de contexto.

2) Respondentes que individualmente assinalarem para cada uma das notícias que: a) consumiriam em todos os contextos; ou b) não consumiriam em nenhum contexto serão considerados “imunes” às variações de contexto. Todos os demais respondentes que tiverem qualquer alteração nesse padrão serão considerados impactados pela questão contextual proposta.
3) Respondentes que individualmente assinalarem possuir um alto nível de interesse (5 na escala) em uma notícia e que assinalarem que consumiriam seu conteúdo nos três cenários contextuais propostos serão considerados “imunes” à questão contextual em caso de notícias de grande interesse. Todos os demais respondentes que assinalarem que não consumiriam ao menos uma das notícias de alto nível de interesse nos três contextos apresentados, serão considerados impactados pela questão contextual proposta.

A partir dos resultados obtidos na pesquisa, serão propostos cenários teóricos para a recomendação de conteúdo aplicada a alguns dos participantes dessa pesquisa. Para isso, serão selecionados 3 participantes de diferentes áreas de atuação. Os cenários serão propostos a partir do cruzamento das informações de nível de interesse e de contextos com maior possibilidade para o consumo de cada notícia. O objetivo é identificar cenários em que a recomendação poderia ampliar o nível de engajamento de usuários reais a partir da inclusão de uma variável contextual.

# 3 **LEVANTAMENTO DO ESTADO ATUAL DA RECOMENDAÇÃO DE CONTEÚDO EM SITES DE NOTÍCIAS BRASILEIROS E PROPOSTA DE CATEGORIZAÇÃO**

Após o entendimento a respeito do funcionamento dos sistemas de recomendação de conteúdo baseados em algoritmos do ponto de vista técnico e do levantamento teórico acerca deste tema, faz-se necessário entender na prática como esses sistemas vêm sendo utilizados no domínio do jornalismo e dos sites de notícias.

Para analisar o estado atual da utilização de sistemas de recomendação de conteúdo nos sites de notícias brasileiros, conforme apresentado no item 2.1 do capítulo que trata dos procedimentos metodológicos da pesquisa, foram selecionados e analisados os seguintes 8 sites: G1, O Globo, Terra, Uol, Folha de São Paulo, R7, Estadão e Yahoo Notícias.

Conforme já mencionado no capítulo de fundamentação teórica, uma das maiores dificuldades no estudo acerca deste tema é a impossibilidade de acessar os códigos fonte dos sistemas de recomendação de conteúdo utilizados pelos sites de notícia - por serem de caráter privado e trazerem vantagens competitivas às respectivas empresas. Por conta disso, optou-se por utilizar a “Teoria da Engenharia Reversa”, uma técnica específica para analise de algoritmos proposta por Diakopoulos (2013).

Neste caso, a avaliação dos sistemas de recomendação se deu a partir da observação rigorosa do ‘output’, ou seja, daquilo que está visível para todos que acessam os sites de notícias - os conteúdos que são recomendados nas páginas de matéria dos sites. Sendo assim, para cada um dos sites citados, foram observadas as páginas de matéria e os respectivos elementos de recomendação de conteúdo presentes em cada uma delas.

Com os elementos identificados, aprofundou-se no entendimento de suas regras funcionais, avaliando cada um dos conteúdos oferecidos nesses elementos, do ponto de vista da relação existente entre os conteúdos e os usuários, o site e a empresa de notícias. Foram observados quais os conteúdos que estão sendo recomendados, qual o critério de ordenação e escolha desses conteúdos, em que momentos eles são recomendados, etc. A partir daí, foi possível aproximar-se de suas regras de funcionamento.

De maneira a estruturar a análise dos elementos de recomendação de conteúdo dos sites de notícias selecionados, utilizou-se como referência o conceito das quatro decisões atômicas algorítmicas proposto por Diakopoulos (2013) e descrito na fundamentação teórica desta tese, a saber: 1) priorização; 2) classificação; 3) associação e 4) filtragem. Esta abordagem permitiu a padronização nos critérios de observação dos elementos de recomendação.

Dessa forma, a partir do entendimento a respeito do atual momento dos sites de notícias no Brasil e dos padrões identificados nesta abordagem, será proposta uma categorização dos principais critérios para recomendação de conteúdo

identificados nesses sites. Tal contribuição servirá como referência para a indicação de novos critérios para um modelo de recomendação que leve mais em conta os fatores contextuais, na tentativa de melhorar o engajamento dos usuários com os sites de notícias.

## 3.1 **Estado atual da recomendação de conteúdo em sites de notícias brasileiros**

A seguir, será apresentado o resultado consolidado do trabalho de análise das páginas de matérias dos 8 sites de notícias brasileiros selecionados, conforme definido acima. É importante ressaltar que a análise se deu ao longo do ano de 2016 e, portanto, muitos destes sites já passaram por novos processos de reformulação, apresentando diferenças em relação às versões aqui apresentadas. Por conta disso, essa análise deve ser considerada um retrato dos sites de notícias brasileiros nesse período, e o acompanhamento da questão e a análise frequente acerca dos elementos de recomendação se fazem necessários na busca por novos modelos que possam vir a surgir na recomendação de conteúdo em páginas de notícias.

Nos subcapítulos a seguir serão apresentados os principais modelos de elementos de recomendação de conteúdo identificados especificamente nas páginas de matérias. No âmbito dos sites de notícias, a matéria é sem dúvida o principal e mais relevante formato de conteúdo. Além disso, do ponto de vista da recomendação de conteúdo, as páginas de matéria são as que mais se distinguem entre cada um dos sites analisados.

Nas demais páginas de conteúdos como fotos, vídeos ou galerias, foram identificados critérios de recomendação muito parecidos entre si, quase sempre sugerindo outros conteúdos de mesmo tipo (ex: em uma galeria de fotos, recomendação de novas galerias de fotos de outros temas). Já em páginas de conteúdos do tipo “coluna”, o padrão em todos os sites analisados é a recomendação de outras colunas do mesmo autor, independente da temática relacionada a coluna inicialmente acessada.

Sendo assim, segue abaixo a descrição dos elementos de recomendação identificados nos principais sites de notícias do Brasil.

### 3.1.1 G1

A figura 8 apresenta um exemplo de página de matéria do site G1 (g1.com.br). Nele, é possível identificar 3 elementos distintos para recomendação de conteúdo: “Rio de Janeiro”, “Saiba Mais” e “Primeira Página”.

Figura 8 – Screenshot de uma página de matéria do site G1 (Autor: Augusto Lohmann)

A exceção do elemento “Primeira Página”, que recomenda de maneira editorial os conteúdos em destaque na homepage do site (independente do contexto do usuário), os outros 2 elementos identificados oferecem conteúdos diretamente relacionados à matéria que está sendo consumida pelo usuário. Nos 2 elementos, os conteúdos recomendados são sempre relacionados a mesma editoria da matéria principal (Rio de Janeiro).

### 3.1.2 O Globo

A página de matéria do site do jornal O Globo (oglobo.com.br) conta com 6 elementos de recomendação de conteúdo destacados, a saber: "Brasil" (editoria da matéria), "Veja Também", "Anterior e Próxima", "Mais Populares", "Em destaque agora no Globo" e "Mais lidas".

Figura 9: Página de matéria do site do jornal o Globo, com elementos de recomendação de conteúdo destacados.

Sobre os critérios utilizados para a recomendação, o elemento “Brasil” é na verdade um elemento que assume o nome da editoria da matéria que foi acessada, recomendando mais notícias dessa mesma editoria. O elemento “Veja Também” funciona de maneira similar em relação aos critérios para a recomendação.

Já o elemento “Mais Lidas” traz a recomendação dos conteúdos mais consumidos de todo o site dentro de um determinado período de tempo definido pelo corpo editorial do site, da mesma maneira que o elemento “Mais Populares”.

“Em destaque agora no Globo” é o elemento que traz as recomendações editoriais, a partir do conteúdo que está em destaque na homepage do site.

Por último, o elemento “Anterior e Próxima” atribui a lógica cronologia às recomendações, trazendo notícias publicadas imediatamente antes e depois daquela que está sendo acessada.

### 3.1.3 Terra

A figura 10 representa a página de matéria do site Terra (terra.com.br). Nesse caso, é possível identificar 3 principais blocos com elementos de oferta de conteúdos: “Saiba Mais”, “Tudo sobre Brasil” e “Sugerido por Outbrain”.

Figura 10 - Screenshot de uma página de matéria do site Terra (Autor: Augusto Lohmann)

O elemento “Saiba Mais” recomenda conteúdos diretamente relacionados à matéria principal que está sendo acessada pelo usuário. Da mesma forma, o elemento “Tudo sobre Brasil” traz recomendações diretamente relacionadas ao tema da matéria principal, que está cadastrado justamente na editoria “Brasil. É

interessante notar que, apesar dos 2 elementos trazerem recomendações relacionadas ao conteúdo da matéria que está sendo acessada, os critérios para esse relacionamento variam entre si. No elemento “Saiba Mais”, as notícias não apenas são do Brasil como também tratam especificamente do tema da violência, exatamente como na matéria principal. Já no elemento “Tudo sobre Brasil”, os critérios são menos detalhados, e são recomendadas notícias da editoria Brasil, independente da temática da violência.

O último elemento de recomendação da página, “Sugerido por Outbrain”, faz menção ao uso da plataforma Outbrain (http://www.outbrain.com), plataforma especializada na recomendação de conteúdos a partir de um modelo híbrido de algoritmos, com ênfase em conteúdos patrocinados. O trecho abaixo, retirado do próprio site da empresa, apresenta a complexidade do sistema utilizado para uma melhor oferta de conteúdos relevantes:

> Alguns dos algoritmos da Outbrain analisam o conteúdo a ser oferecido, determinando quais páginas podem ser mais atraentes. Outros estudam as ações da pessoa no site para exibir, a ela, links mais adequados a seus interesses.
> São mais de 50 algoritmos. Alguns deles competem entre si, já que cada um encontra uma resposta diferente sobre qual conteúdo deve ser recomendado. Outros decidem isso com base no tema da página em que a pessoa está, nos links em que ela clicou e em outros dados.

Nesse caso, ao observar o resultado prático do sistema da Outbrain no site avaliado, pode-se perceber que as notícias recomendadas são pouco relacionadas diretamente ao conteúdo que está sendo consumido, por utilizarem-se de outros critérios na composição do grau de relevância da oferta para o usuário que não somente a relação com o tema da notícia.

### 3.1.4 UOL

A figura 11 apresenta um exemplo de matéria do site Uol Notícias. Nesse caso, foram identificados 3 elementos para recomendação de conteúdos: “Veja Também”, “Recomendado para você” e um terceiro elemento sem título, a direita da página.

Figura 11 - Screenshot de uma página de matéria do site UOL (Autor: Augusto Lohmann)

O elemento “Veja Também” oferece conteúdos com temática bastante similar ao do conteúdo da matéria que está sendo acessada (violência, tiroteios, etc).

Já a seção “Recomendado para você” utiliza o mesmo sistema de recomendação da empresa Outbrain (identificado anteriormente no site do Terra). Esse sistema, como já explicado, é baseado no modelo híbrido de recomendação de conteúdos, e tenta apresentar opções com alto índice de personalização, ao mesmo

tempo que prioriza a oferta de conteúdos patrocinados sem necessariamente nenhuma relação com o conteúdo que está sendo consumido.

O elemento da coluna da direita, sem título, apresenta sempre os mesmos itens, independentemente da matéria que esteja sendo acessada. Nesse caso, foi identificado como um elemento cujo conteúdo é determinado pela equipe editorial do site.

### 3.1.5 Folha de São Paulo

Na figura 12, uma página de matéria do site Folha de São Paulo (folha.com.br), é possível identificar 5 elementos de recomendação de conteúdo: “Leia também”, “Mais Lidas” (dentro da editoria), “+ Livraria”, “Livraria da Folha” e “Folha Shop”.

Figura 12 - Screenshot de uma página de matéria do site Folha de São Paulo (Autor: Augusto Lohmann)

O elemento “Leia também” apresenta mais conteúdos de maior destaque editorial sobre o mesmo assunto da matéria principal. Já o elemento “Mais Lidas” da editoria seleciona os conteúdos mais acessados sobre aquele assunto dentro de um

período de tempo. Esse é um elemento que exemplifica um modelo híbrido que reúne mais de um critério de recomendação em um mesmo elemento.

Os demais elementos da página são todos voltados para a venda de produtos, recomendado conteúdos patrocinados de outros sites, sem nenhuma relação com o conteúdo noticioso do próprio site.

### 3.1.6 R7

A figura 13 apresenta um exemplo de matéria do site R7 Notícias. Nessa página, foi possível identificar dois elementos principais para oferta de conteúdo recomendado: “Notícias Relacionadas” e “Veja Também”.

Figura 13 - Screenshot de uma página de matéria do site R7 Notícias (Autor: Augusto Lohmann)

No elemento “Notícias Relacionadas”, a recomendação é feita a partir de conteúdos que tratam do mesmo tema da violência (mesmo assunto), com noticias que se limitam a esse aspecto. Já no elemento “Veja Também”, o critério utilizado para relacionamento é a editoria “Brasil”, e nesse caso é possível observar a recomendação de assuntos mais diversos e menos similares ao conteúdo originalmente acessado na matéria principal.

### 3.1.7 Estadão

A figura 14 apresenta um exemplo de matéria do site de notícias Estadão (estadao.com.br). Nessa página, foi possível identificar dois elementos principais para oferta de conteúdo recomendado: “Agora no Estadão”, “Veja Também”, “+ Internacional” (editoria da notícia) e “Galerias Estadão”.

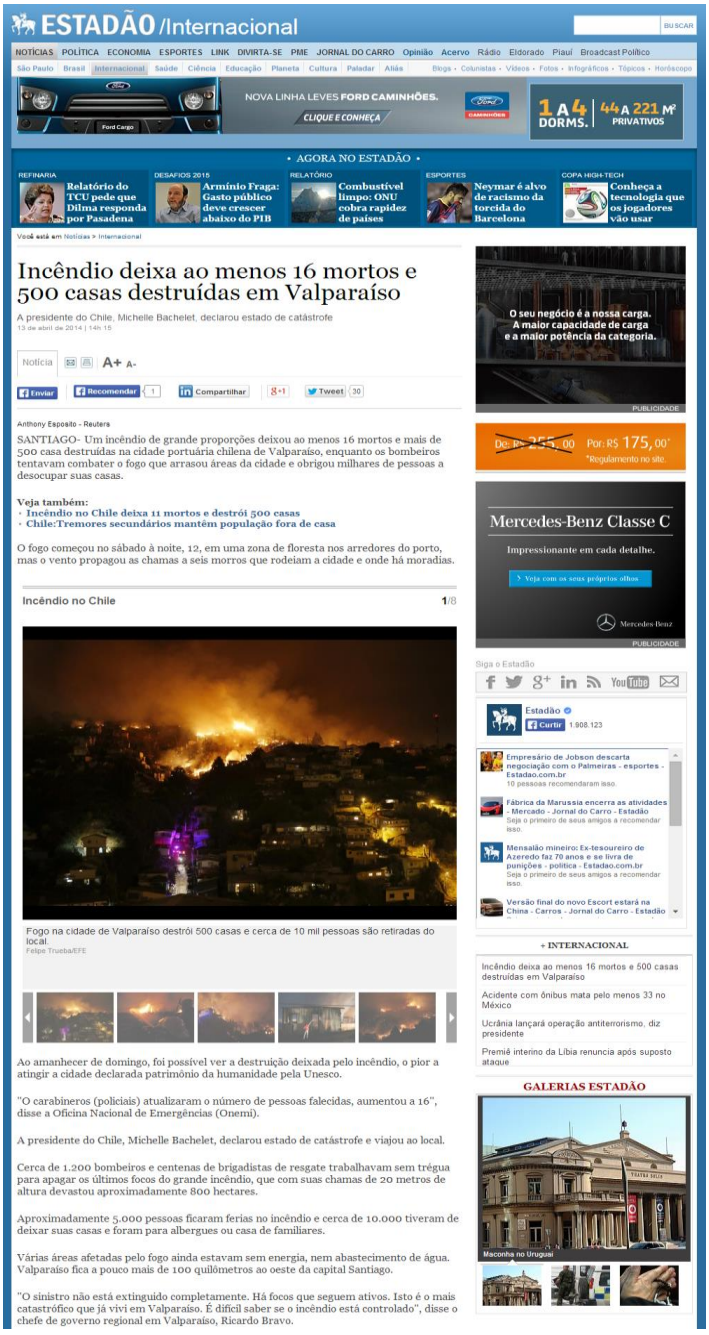


Figura 14 - Screenshot de uma página de matéria do site Estadão Notícias (Autor: Augusto Lohmann)

O elemento “Agora no Estadão” traz as notícias mais recentes que estão em destaque na homepage do site. Já os elementos “Veja Também” e “+ Internacional” recomendam notícias relacionadas, da mesma editoria da notícia principal (porém aparecem em momentos diferentes da experiência de consumo). Por último, o elemento “Galerias Estadão” recomenda outras galerias de fotos do próprio site (neste caso, a matéria selecionada para análise continha uma galeria de fotos em meio ao conteúdo textual).

### 3.1.8 Yahoo Notícias

A figura 15 apresenta um exemplo de matéria do site Yahoo Notícias. Nesse exemplo, foi possível identificar 3 principais elementos para recomendação de conteúdo: “Destaques em Brasil”, “Leia Também” e “Notícias que deixaram”.

Figura 15 - Screenshot de uma página de matéria do site Yahoo Notícias (Autor: Augusto Lohmann)

O elemento “Destaques em Brasil” oferece aos usuários conteúdo relacionado ao assunto da matéria que está sendo acessada.

Já o elemento “Leia Também” traz a recomendação de conteúdos considerados relevantes editorialmente para o site, mesmo que sem nenhuma relação direta com o conteúdo acessado pelo usuário. Nesse caso, é mais um exemplo de um elemento baseado em critérios editoriais.

Por último, o elemento “Notícias que deixaram” recomenda notícias que deixaram outros usuários com um determinado estado de humor (bom humor, neutro, mau humor, etc). Este elemento não traz relação com o assunto ou temática

do conteúdo. Sendo assim, o critério tem relação direta com a maneira com que a comunidade de usuários do site avaliou um determinado conteúdo acessado. Este critério de recomendação foi encontrado apenas no site do Yahoo Notícias, não sendo utilizado por nenhum dos outros sites analisados.

## 3.2 **Proposta de categorização para a recomendação de conteúdo nos sites de notícias brasileiros**

Após a identificação e análise do funcionamento dos principais elementos de recomendação das páginas de conteúdo dos 8 sites selecionados, procurou-se reconhecer os padrões nos critérios para a oferta de conteúdo. A partir desses padrões, é proposta uma segmentação desses critérios em categorias que retratem o estado atual da utilização de sistemas de recomendação de conteúdos em sites de notícias brasileiros.

É importante ressaltar também que as categorias que serão propostas não foram necessariamente identificadas em sua aplicação pura e isolada nos elementos de recomendação. Em grande parte dos casos analisados, os elementos apresentam combinações dessas categorias integradas, de maneira a oferecer o conteúdo de maneira cada vez mais especializada e próxima do que se acredita ser a melhor experiência de uso em cada contexto. O isolamento desses critérios se deu apenas para um melhor entendimento de cada um deles sob o aspecto acadêmico e para a apresentação didática dos conceitos associados.

Tal categorização servirá também como referência para a compreensão de problemas que levam ao baixo engajamento dos usuários nos sites de notícias, em comparação com outros domínios onde a recomendação sistemática de conteúdos apresenta melhores resultados no ponto de vista do engajamento do público.

Em um primeiro momento, identificou-se a possibilidade de segmentar a recomendação de conteúdo em 4 eixos principais, conforme Lohmann (2013), a saber: empresa, conteúdo, usuário e comunidade.

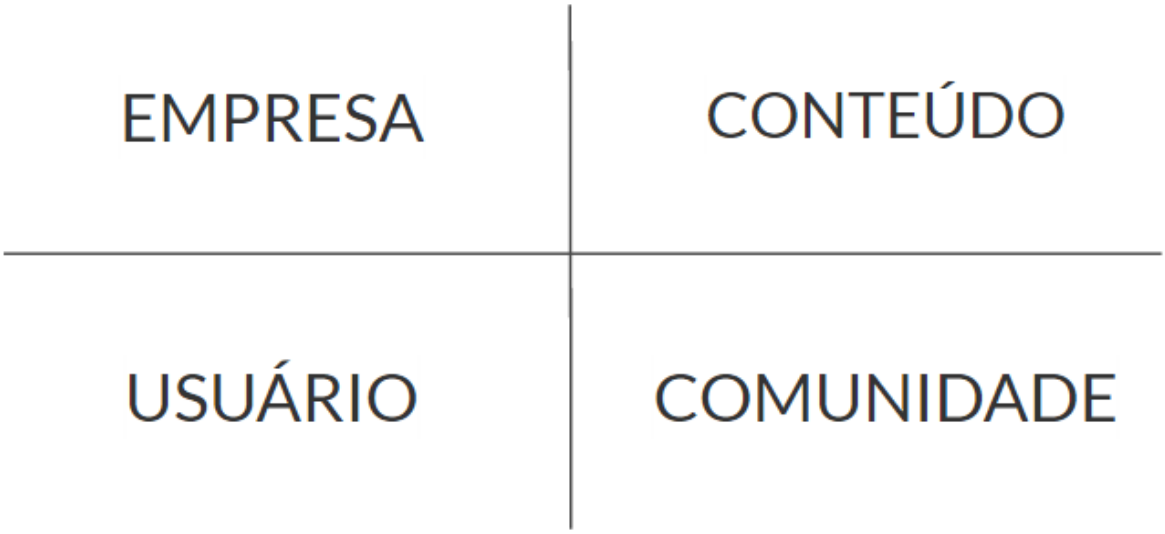


Figura 16: Eixos da recomendação de conteúdo em páginas de matéria: Empresa, Conteúdo, Usuário e Comunidade

O eixo “Empresa” é o que contempla a recomendação de conteúdo cujos critérios são definidos pela empresa noticiosa. Nos elementos do eixo “empresa”, a recomendação de conteúdo é baseada principalmente naquilo que a empresa jornalística decide que deve ser informado para o usuário. Essa escolha pode ser tanto por critérios editoriais (aquilo que a empresa julga importante) quanto por critérios mercadológicos (conteúdo patrocinado que gera dinheiro para a empresa). Assim, os conteúdos oferecidos não têm necessariamente uma relação direta com a notícia que está sendo acessada pelo usuário, ou com algum interesse previamente manifestado por este ou pelo restante da comunidade de usuários do site. Neste eixo foram encontrados elementos de recomendação categorizados como: “Editorial” e “Patrocinado”, que serão apresentados e definidos nos próximos subcapítulos.

Já o eixo “Conteúdo” contempla toda a recomendação de conteúdo cujos critérios são herdados do próprio conteúdo noticioso que está sendo acessado naquele momento. Ou seja, o acesso do usuário a um determinado conteúdo forneceria as informações para a recomendação de conteúdos relacionados por diversos critérios. Nesse contexto, o conteúdo recomendado pode não ser relevante jornalisticamente na visão da empresa noticiosa, nem necessariamente ter relação com características da navegação prévia do usuário, mas é sempre algo que tem semelhanças com o conteúdo principal que está sendo acessado. Neste eixo foram encontrados elementos de recomendação categorizados como: “Assunto”, “Autoria”,

“Cronologia”, “Geolocalização” e “Tipo de Conteúdo”, que serão apresentados e definidos nos próximos subcapítulos.

No terceiro eixo “Usuário”, a recomendação de conteúdo é baseada em características ou atitudes do próprio usuário que acessa o site, a partir de informações que ele fornece ou por sua própria experiência prévia de navegação. Nesse caso, o conteúdo recomendado pode não ser relevante jornalisticamente na visão da empresa, pode não ter relação com o conteúdo que estava sendo acessado, mas tem relação com o que o usuário normalmente tem o hábito de acessar (“Pessoal”) ou com o comportamento de alguém que faça parte da rede de pessoas ligadas a esse usuário (“Relacionamento Social”), que serão apresentados e definidos nos próximos subcapítulos.

O 4º e último eixo “Comunidade” trata da recomendação de conteúdo baseada em características de grupos de usuários que compõem a comunidade geral que acessa o site. Nesse contexto, o conteúdo recomendado pode não ser relevante jornalisticamente, pode não ter relação com o conteúdo acessado no momento, pode não ter relação com a navegação de um usuário específico ou de seus amigos, mas tem relação com a navegação de um grupo ou até de toda a comunidade de usuários do site. Neste eixo foram encontrados elementos de recomendação categorizados como: “Emocional” e “Popularidade”, que serão apresentados e definidos nos próximos subcapítulos.

Sendo assim, dentro dos 4 eixos apresentados, a pesquisa realizada possibilitou identificar e categorizar os critérios para os sistemas de recomendação nas páginas de conteúdo dos sites de notícias brasileiros em 11 categorias, conforme apresentado na figura 17, a saber: 1) Editorial; 2) Patrocinado; 3) Assunto; 4) Autoria; 5) Cronologia; 6) Geolocalização; 7) Tipo de conteúdo; 8) Pessoal; 9) Relacionamento Social; 10) Emocional; 11) Popularidade.

Figura 17: Categorias de recomendação de conteúdo no âmbito do jornalismo relacionadas aos 4 eixos: empresa, conteúdo, usuário e comunidade

Conforme descrito anteriormente, para estruturar e padronizar a definição de cada uma das categorias, foi utilizado como referência o conceito das quatro decisões atômicas algorítmicas proposto por Diakopoulos (priorização; classificação; associação e filtragem). A seguir, portanto, será detalhada cada uma das 11 categorias identificadas na recomendação de conteúdo em sites de notícias brasileiros.

### 3.2.1 Editorial

Na categorização editorial, a recomendação de conteúdo não segue nenhuma lógica relacionada a características dos usuários ou do conteúdo noticioso em si. A organização e oferta dos conteúdos é baseada naquilo que o corpo editorial da empresa considera um fato de relevância noticiosa. Este critério ainda é bastante comum e foi encontrado na maioria dos sites de notícias brasileiros, sendo talvez o que mais se assemelha aos critérios de oferta de conteúdo convencionais herdado do jornalismo impresso.

No exemplo da figura 18, o elemento de recomendação de conteúdo do site G1 dá destaque a notícias que estão na *homepage* do site - que tende a ser o conteúdo considerado de maior relevância pelo corpo editorial. Nesse contexto, o

que se acredita é que o número de cliques e interações será elevado na medida em que a curadoria do conteúdo oferece o que aconteceu de mais “importante”, e que, portanto, ampliaria o engajamento do leitor/usuário.

Figura 18: Box de notícias mais relevantes selecionadas pela equipe editorial do site. Fonte: G1 – Bem Estar

Sendo assim, foi possível identificar o conjunto de regras dos elementos de recomendação de conteúdo do tipo “editorial” a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

- **Priorização**: Os elementos de recomendação do tipo “editorial” priorizam conteúdos selecionados com base na avaliação de sua importância relativa por parte do corpo editorial do site de notícias. Foram encontradas duas soluções para a priorização desses conteúdos: cronológica (exibe com maior destaque com base na

data e horário de publicação de cada notícia) ou relevância (exibe com maior destaque com base no critério de escolha editorial).

**- Classificação**: Nos elementos de recomendação editorial, a classificação das notícias se dá em função do alto nível de relevância atribuído a cada uma delas por parte do corpo editorial do site. Notícias consideradas de grande relevância editorial são classificadas e agrupadas em um mesmo elemento de recomendação.

**- Associação**: Não foi identificada nenhuma relação definida de associação entre os conteúdos recomendados nos elementos do tipo “editorial”.

**- Filtragem**: Com relação a filtragem, são recomendados somente os conteúdos identificados de alguma forma como relevantes por parte do corpo editorial. Todos os demais conteúdos são excluídos da recomendação.

### 3.2.2 Patrocinado

A recomendação de conteúdos patrocinada não segue necessariamente nenhuma lógica relacionada a características dos usuários ou do conteúdo noticioso em si. Neste modelo, alguns (ou todos) os espaços para recomendação são disponibilizados para conteúdos pagos por empresas terceiras para obterem destaque e atingirem o maior número de usuários dentro do site de notícias.

Este tipo de recomendação foi identificado em muitos dos sites brasileiros. O resultado tende a ser positivo para a empresa de notícias do ponto de vista do retorno financeiro, mas prejudicial do ponto de vista do engajamento, já que diminue muito a probabilidade de um conteúdo recomendado ser do interesse do usuário em questão. Na maioria das vezes, o conteúdo recomendado sequer é do site acessado, mas de sites externos. A interação com esses elementos, nesses casos, representa o abandono do site por parte do usuário.

No exemplo da figura 19, o elemento de recomendação de conteúdo do site Terra dá destaque a notícias patrocinadas. O elemento mescla 2 conteúdos do próprio site do Terra com 2 conteúdos externos e patrocinados. Nesse contexto, é priorizado o retorno financeiro imediato em detrimento do maior engajamento de longo prazo.

Figura 19: Elemento de recomendação com a presença de conteúdo patrocinado por terceiros em matéria do site Terra (terra.com).

Sendo assim, foi possível identificar o conjunto de regras dos elementos de recomendação de conteúdo do tipo “patrocinado” a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

**- Priorização**: Os elementos de recomendação do tipo “patrocinado” priorizam conteúdos com base nos valores pagos pela empresa patrocinadora. Quanto mais a empresa paga, em geral mais destaque aquele determinado conteúdo obtém em um elemento de recomendação.

**- Classificação**: Nesses elementos não foram identificados critérios de classificação.

**- Associação**: Não foi identificada nenhuma relação definida de associação entre os conteúdos recomendados nos elementos do tipo “patrocinado”.

**- Filtragem**: Com relação a filtragem, são recomendados somente os conteúdos patrocinados. Todo o restante do conteúdo é excluído do elemento.

### 3.2.3 Assunto

Nessa categoria, a recomendação de conteúdo se dá pelo assunto da notícia que está sendo acessada, ou seja, notícias do mesmo assunto e/ou temática são sugeridos. No exemplo da figura 20, na leitura de um conteúdo sobre “Ciência e Saúde”, são recomendadas notícias sobre “Ciência” e “Saúde”.

Figura 20: Box de mais notícias de um mesmo assunto/tema (Ciência e Saúde). Fonte: G1 – Bem Estar

A recomendação por assunto é bastante comum em sites noticiosos, e mostra-se de grande relevância em contextos onde o assunto do conteúdo é fator de relevância para motivar o engajamento do usuário.

Sendo assim, foi possível identificar o conjunto de regras dos elementos de recomendação de conteúdo do tipo “assunto” a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

**- Priorização**: Os elementos de recomendação do tipo “assunto” em geral priorizam notícias pelo critério cronológico, ou seja, atribuem um melhor posicionamento na ordenação com base na data e horário de publicação de cada notícia.

**- Classificação**: Nos elementos de recomendação por assunto, a classificação das notícias se dá em função da categoria, tema ou editoria atribuída a cada uma delas. Notícias que tratem de um mesmo “assunto” (por exemplo, esportes, política, economia, educação, etc) são classificadas e agrupadas em um mesmo elemento de recomendação.

**- Associação**: Não foi identificada nenhuma relação definida de associação entre os conteúdos recomendados nos elementos do tipo “assunto”.

**- Filtragem**: Com relação a filtragem, são recomendados somente os conteúdos classificados como de um mesmo “assunto”. Todos os demais conteúdos são excluídos da recomendação.

### 3.2.4 Autoria

A recomendação de conteúdo por “autoria” é mais um critério que tem relação direta com uma característica do conteúdo que está sendo acessado. A autoria normalmente é utilizada quando se entende que o autor do conteúdo é tão ou mais relevante do que o próprio conteúdo em si.

Elementos de recomendação por autoria são bastante utilizado em sites noticiosos para oferecer conteúdos de jornalistas, colunistas, colaboradores e/ou blogueiros que, por si só, sejam capazes de atrair o interesse dos usuários, independente do assunto ou editoria do conteúdo apresentado.

No exemplo da figura 21, do site O Globo, o conteúdo é recomendado por sua autoria, de tal forma que o usuário possa consumir mais conteúdos do mesmo autor da matéria originalmente acessada.

Figura 21: Box de notícias agrupadas por autoria do conteúdo. Fonte. O Globo

Sendo assim, fez-se possível a identificação do conjunto de regras dos elementos de recomendação de conteúdo do tipo “autoria” a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

**- Priorização**: Os elementos de recomendação por “autoria” encontrados priorizam notícias pelo critério cronológico, ou seja, atribuem um melhor posicionamento na ordenação com base na data e horário de publicação de cada notícia.

**- Classificação**: Nos elementos de recomendação por autoria, a classificação das notícias se dá em função do autor de cada notícia, independente do assunto ou editoria do conteúdo em questão. Todas as notícias produzidas por um mesmo autor são classificadas como de um mesmo tipo.

**- Associação**: Não foi identificada nenhuma relação definida de associação entre os conteúdos recomendados nos elementos do tipo por “autoria”.

**- Filtragem**: Com relação a filtragem, são recomendados somente os conteúdos classificados como escritos por um mesmo “autor”. Todos os demais conteúdos são excluídos da recomendação.

### 3.2.5 Cronologia

A recomendação por “cronologia” é bastante comum no contexto do jornalismo online, e está associada a uma característica inerente ao próprio conteúdo (sua data de publicação). A simples priorização da oferta de conteúdos recomendados em função da data de sua publicação está presente em quase todos os tipos de elementos de recomendação, já que no contexto das notícias esta informação é bastante significativa. A recência da notícia é sem dúvida um dos critérios mais relevantes na formação do interesse do usuário em relação a um conteúdo noticioso, motivando uma possível interação por parte desse usuário.

Entretanto, existem elementos onde a classificação dos conteúdos recomendados se dá única e exclusivamente em função de sua cronologia, tendo como critério somente a data de publicação/atualização de cada conteúdo, independente do assunto ou tema.

Como exemplo, elementos do tipo “Últimas Notícias” como o da figura 22 e áreas de plantão com todas as notícias em ordem cronológica são os principais exemplos da utilização dessa categoria de recomedação de conteúdo.

Figura 22: Elemento de recomendação “Últimas Notícias”. Fonte: G1

Sendo assim, fez-se possível a identificação do conjunto de regras dos elementos de recomendação de conteúdo do tipo “cronologia” a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

- **Priorização**: Os elementos de recomendação do tipo “cronologia” priorizam notícias pelo critério cronológico, ou seja, atribuem um melhor posicionamento na ordenação com base em notícias cujas datas e horários de publicação são mais recentes.

- **Classificação**: Nos elementos de recomendação por “cronologia” não foi identificado nenhum critério de classificação de conteúdo. Todas as notícias, de

qualquer assunto ou editoria podem ser exibidas de acordo com a priorização descrita acima.

**- Associação**: Não foi identificada nenhuma relação definida de associação entre os conteúdos recomendados nos elementos do tipo por “cronologia”.

**- Filtragem**: Em elementos desse tipo não há uma filtragem de conteúdos. Todos os conteúdos podem ser recomendados em função de sua priorização descrita acima.

### 3.2.6 Geolocalização

A localização geográfica ou geolocalização é um importante critério para a recomendação de notícias, e está relacionado a uma característica do próprio conteúdo que se está consumindo. A localização - seja do usuário, seja do conteúdo da notícia - tende a ser um critério de bastante relevância no interesse dos usuários pelo consumo de uma notícia (por exemplo, um assalto que acontece na cidade do usuário tende a ser de mais relevância do que um assalto similar ocorrido em outro estado ou país).

Foram encontrados diversos exemplos de elementos de recomendação oferecendo conteúdo organizado por geolocalização (como na figura 23, em uma área de conteúdo com notícias ocorridas em um determinado estado e/ou região). Tende a ser utilizado quando a geolocalização da notícia se torna um fator de escolha tão ou mais relevante do que o próprio conteúdo ou o assunto da notícia.

Figura 23: Box de Notícias filtradas pela região geográfica (RJ). Fonte: G1 – RJ

Nesta categoria, foi possível identificar o conjunto de regras dos elementos de recomendação de conteúdo do tipo “geolocalização” a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

- **Priorização**: Os elementos de recomendação por “geolocalização” encontrados nos sites de notícias brasileiros priorizam o conteúdo sugerido pelo critério cronológico, ou seja, dão maior destaque com base na data e horário de publicação de cada notícia.

- **Classificação**: Nos elementos de recomendação por geolocalização, a classificação das notícias se dá em função da localização atribuída ao conteúdo, independente de seu assunto ou editoria. Todas as notícias de uma mesma localidade (cidade, estado ou país) são classificadas como de um mesmo tipo.

- **Associação**: Nestes elementos a associação se dá quando a localização do usuário é identificada e conteúdos daquela mesma localidade são recomendados, mesmo que o conteúdo originalmente acessado não tenha nenhuma relação com a geolocalização do usuário.

**- Filtragem**: Com relação a filtragem, são recomendados somente os conteúdos classificados como pertencentes a uma localidade específica. Todos os conteúdos de outras localidades ou sem localidade atribuída são excluídos da recomendação.

### 3.2.7 Tipo de conteúdo

A recomendação por tipo de conteúdo se faz presente quando se entende que o usuário deseja continuar sua navegação por conteúdos do mesmo tipo (artigo, notícia, foto, vídeo, galeria, etc). Por exemplo, ao consumir um vídeo sobre um acidente automobilístico, o usuário teria mais interesse em continuar navegando por outros vídeos de quaisquer assuntos, e não por conteúdos de outros tipos sobre o assunto “acidentes automobilísticos”.

Na figura 24, um exemplo de elemento de recomendação do site do Globo Online recomendando outras galerias de fotos a um usuário que acessa uma galeria de fotos de uma partida de futebol. Neste exemplo, não há nenhuma relação entre o assunto ou editoria da galeria acessada e as recomendações em questão.

Figura 24. Box com oferta de mais galerias em um ambiente de consumo de galeria de fotos. Fonte: O Globo Online

Tal critério é bastante utilizado especialmente para o consumo de fotos, vídeos, infográficos e outros tipos de conteúdo diferentes da matéria jornalística convencional.

Nesta categoria, foi possível identificar o conjunto de regras dos elementos de recomendação por “tipo de conteúdo” a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

**- Priorização**: Os elementos de recomendação por “tipo de conteúdo” encontrados priorizam conteúdos pelo critério cronológico, ou seja, atribuem uma posição de maior destaque na ordenação com base nos mais recentemente publicados.

**- Classificação**: A classificação nestes elementos se dá em função o tipo do conteúdo: artigo, notícia, foto, vídeo, galeria, etc. Os conteúdos de um mesmo tipo são agrupados e passam a ser classificados como de uma mesma categoria.

**- Associação**: Não foi identificada nenhuma relação definida de associação entre os conteúdos recomendados nos elementos do tipo de conteúdo.

**- Filtragem**: Com relação a filtragem, são recomendados somente os conteúdos classificados como de um mesmo “tipo”. Por exemplo, em uma galeria de fotos, recomenda-se outras galerias de fotos. Em um vídeo, recomenda-se outros vídeos. Neste caso, todos os demais conteúdos são excluídos do elemento de recomendação.

### 3.2.8 Pessoal

A recomendação pessoal se dá de acordo com determinados padrões de comportamento e de características do próprio usuário que está exercendo a navegação. No exemplo da figura 25, o elemento permite que o usuário personalize o conteúdo que será oferecido (no caso, a escolha do time faz com que notícias sobre seu time sejam apresentadas).

Figura 25: Box para escolha do time para personalização do conteúdo oferecido. Fonte: Globo.com

Este modelo em particular apresenta-se como uma solução bastante funcional para a recomendação de conteúdo, na medida em que tende a apresentar somente o que é de interesse do usuário (seja esse interesse manifesto, a partir de uma interação consciente, ou mesmo baseado em métricas inferidas a partir de toda a experiência de navegação desse usuário pelo site noticioso). Entretanto, esse tipo de solução ainda é usado de forma bastante limitada e incipiente pelos sites de notícias brasileiros e foram pouco encontrados na pesquisa realizada.

Para estes elementos, foi possível identificar o conjunto de regras a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

**- Priorização**: Os elementos de recomendação “pessoal” encontrados nos sites de notícias brasileiros priorizam o conteúdo sugerido pelo critério cronológico, ou seja, atribuem um melhor posicionamento na ordenação aos conteúdos mais recentes, com base na data e horário de publicação de cada notícia.

**- Classificação**: Nos elementos de recomendação pessoal não foram identificados critérios de classificação de conteúdo. Todas as notícias, de qualquer assunto ou editoria podem ser exibidas de acordo com a priorização e a associação deste conteúdo.

**- Associação**: Nestes elementos a associação se dá em geral por meio de informações pessoais de cada usuário, que podem ser obtidas por meio de uma interação ativa com o site ou por quaisquer outras fontes de dados que o site tenha acesso (histórico de navegação do usuário neste ou em outros produtos, gostos pessoais, características, comportamentos, etc). A partir dessas informações, o elemento de recomendação sugere conteúdos que podem não ter relação nenhuma com a notícia que está sendo acessada naquele momento, mas que se entende que poderão interessar ao perfil daquele usuário.

**- Filtragem**: Não foi identificado um critério para a filtragem de determinados tipos de conteúdo. Todas as notícias poderiam vir a ser incluídas nos elementos de recomendação do tipo “pessoal”.

### 3.2.9 Relacionamento Social

Alguns sites de notícias brasileiros recomendam conteúdo a partir do relacionamento social do usuário, ou seja, pelo padrão de consumo de outros usuários que estejam no círculo social do usuário que está acessando a página de conteúdo naquele momento. No exemplo da figura 26, o elemento “Seus amigos leram...” recomenda para o usuário conteúdos que foram acessados por seus amigos da rede social *Facebook*.

Figura 26: Box de notícias lidas pelos amigos do usuário no Facebook. Fonte: G1

Nesse contexto, a relevância se dá na medida em que se acredita haver interesses em comum entre o usuário e seus relacionamentos em redes sociais, e dessa maneira, o conteúdo recomendado tende a atrair o usuário e ampliar seu engajamento. Na recomendação de conteúdo por relacionamento social, se um amigo/conhecido do usuário em uma rede social se interessa por uma determinada notícia, acredita-se que a chance desse mesmo usuário manifestar interesse pela mesma notícia seja maior.

Nestes elementos, foi possível identificar o conjunto de regras a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

- **Priorização**: Os elementos de recomendação por “relacionamento social” encontrados nos sites de notícias brasileiros priorizam o conteúdo sugerido pelo critério cronológico, ou seja, atribuem um melhor posicionamento na ordenação aos conteúdos mais recentes, com base na data e horário de publicação de cada notícia.
- **Classificação**: Nos elementos de recomendação por relacionamento pessoal não foram identificados critérios de classificação de conteúdo. Todas as notícias, de

qualquer assunto ou editoria podem ser exibidas de acordo com a priorização e a associação deste conteúdo.

**- Associação**: Nestes elementos a associação se dá por meio do relacionamento social do usuário. Em geral, foram encontrados exemplos utilizando a rede social Facebook ou o email do Gmail como fonte de informações para esse tipo de associação. A partir dessas informações, o elemento de recomendação sugere conteúdos que podem não ter relação nenhuma com a notícia que está sendo acessada naquele momento, mas que interessaram a amigos daquele usuário e consequentemente poderão vir a interessá-lo.

**- Filtragem**: Não foi identificado um critério para a filtragem de determinados tipos de conteúdo. Todas as notícias poderiam vir a ser incluídas nos elementos de recomendação do tipo “relacionamento social”.

### 3.2.10 Emocional

A recomendação emocional foi pouco identificada nos sites jornalísticos acessados durante a realização da pesquisa. A partir da possibilidade oferecida aos usuários de avaliar as notícias do site pelo critério emocional (o usuário decide se, ao ler aquele conteúdo, ficou feliz, ficou triste, ficou de mau humor, etc), essa informação passa a ser utilizada para estruturar e organizar o conteúdo do site.

No exemplo da figura 27, do site Yahoo Notícias, ao interagir com a página indicando que o conteúdo acessado provocou no usuário uma sensação de mau humor, o elemento “Notícias que deixaram” recomenda outras notícias que deixaram outros usuários de mau humor, independente do assunto ou da temática do conteúdo. Sendo assim, o critério tem relação direta com a maneira com que a comunidade de usuários do site avaliou um determinado conteúdo acessado.

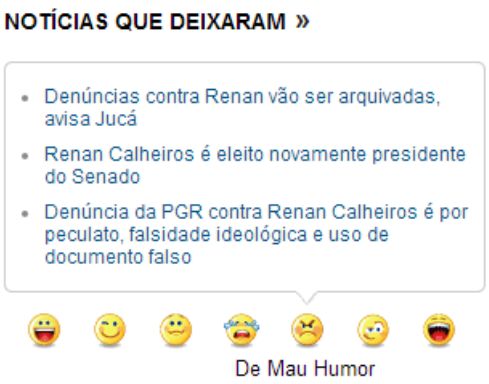


Figura 27. Box com oferta de mais notícias que provocaram a mesma reação emocional em outros usuários.
Fonte: Yahoo Notícias

Nesta categoria, foi possível identificar o conjunto de regras a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

**- Priorização**: Os elementos de recomendação "emocional" encontrados nos sites de notícias brasileiros priorizam o conteúdo sugerido pelo critério cronológico, ou seja, atribuem um melhor posicionamento na ordenação aos conteúdos mais recentes, com base na data e horário de publicação de cada notícia.

**- Classificação**: Nos elementos de recomendação "emocional" os conteúdos são classificados de acordo com a avaliação dos demais usuários. Todas as notícias, de qualquer assunto ou editoria podem ser exibidas categorizadas pelos critérios emocionais estabelecidos pelo site.

**- Associação**: Não foi identificada nenhuma relação definida de associação entre os conteúdos recomendados.

**- Filtragem**: Não foi identificado um critério para a filtragem de determinados tipos de conteúdo. Todas as notícias poderiam vir a ser incluídas nos elementos de recomendação do tipo "emocional".

### 3.2.11 Popularidade

A recomendação de conteúdo em função de sua popularidade, ou seja, por uma maior quantidade de acessos, interações ou visualizações, foi encontrada na

grande maioria dos sites brasileiros pesquisados. Este critério leva em consideração que, se um determinado tipo de conteúdo se mostra relevante para a maior parte da comunidade de usuários de um site, ele tem mais chances de ser relevante para um determinado usuário individualmente.

No exemplo da figura 28, elementos como boxes de “Mais Lidas” oferecem conteúdos mais acessados, mais lidos, mais comentados, ou de maneira geral, os conteúdos que resultaram em uma maior quantidade de interações pela comunidade de usuários, levando em conta justamente a possível relevância desse tipo de conteúdo, conforme destacado no parágrafo anterior.

Figura 28: Box de Mais Lidas. Fonte: G1

Nesta categoria, foi possível identificar o conjunto de regras a partir da utilização do modelo proposto por Diakopoulos (2013), no qual são descritas as quatro decisões atômicas inerentes aos algoritmos (priorização; classificação; associação e filtragem), conforme a seguir:

**- Priorização**: Os elementos de recomendação por “popularidade” encontrados nos sites de notícias brasileiros priorizam o conteúdo sugerido em função de suas métricas. Quanto maiores os índices da métrica definida como parâmetro, mais relevância a notícia recebe no elemento de recomendação.

**- Classificação**: Nestes elementos, pode ou não haver uma classificação do conteúdo recomendado. Foram encontrados exemplos onde as notícias “mais lidas” eram categorizadas por editorias (por exemplo, “mais lidas de esportes”), mas em outros casos, todas as notícias publicadas no site poderiam ser recomendadas desde que atendessem aos padrões de priorização definidos.

**- Associação**: Não foi identificada nenhuma relação definida de associação entre os conteúdos recomendados.

**- Filtragem**: Em todos os elementos de recomendação por “popularidade” identificados, verificou-se a presença de filtros em função da data de publicação das notícias. Nestes elementos, é fundamental que se limite a oferta somente a conteúdos publicados nas últimas horas (ou no máximo nas últimas 24 horas), para que um mesmo conteúdo bem ranqueado não se mantenha diariamente no topo das recomendações. Essa filtragem é o que garante a “renovação” do conteúdo recomendado.

### 3.2.12 Considerações

A possibilidade de mapear e categorizar os critérios que caracterizam a recomendação de conteúdo, bem como os diferentes usos para cada um desses critérios, são de extrema relevância para o avanço dos sistemas de recomendação atuais. Deve-se ressaltar que tal trabalho de categorização não deve ser considerado algo fechado e definitivo, mas aberto a novas possibilidades de organização de conteúdo que venham a ser utilizadas pelos sites noticiosos.

As onze categorias propostas (1) Editorial; 2) Patrocinado; 3) Assunto; 4) Autoria; 5) Cronologia; 6) Geolocalização; 7) Tipo de conteúdo; 8) Pessoal; 9) Relacionamento Social; 10) Emocional; 11) Popularidade) são um retrato do momento atual, a partir da análise de sites de notícias do Brasil. Tais categorias representam bem a organização da oferta de conteúdo em todos os elementos de navegação que são hoje oferecidos pelos sites pesquisados. Por este motivo, é fundamental destacar que o acompanhamento da questão e a análise frequente acerca dos elementos de navegação se fazem necessários na busca por novas categorias além das previamente mapeadas nessa pesquisa.

# 4 APLICAÇÃO DAS TÉCNICAS DE PESQUISA PARA VALIDAÇÃO DE HIPÓTESES

A seguir, conforme descrito no capítulo de procedimentos metodológicos, após a análise do estado atual da recomendação de conteúdo em sites de notícias brasileiros, serão detalhados os processos de aplicação das técnicas de análise de métricas e de pesquisa com usuários, bem como a avaliação de seus resultados. A análise de métricas do site do jornal o Globo tem por objetivo a validação das hipóteses 1 e 2 propostas nessa tese. Já a pesquisa com usuários visa levantar subsídios para a validação da hipótese 3 dessa tese.

## 4.1 **Análise de Métricas do site do jornal O Globo**

Neste item, serão detalhados os resultados da aplicação da técnica de "análise de métricas" no site do jornal O Globo em três diferentes casos.

O primeiro caso traz um comparativo entre as páginas de matéria (antiga e nova) do site do jornal O Globo. Este comparativo tem por objetivo validar a hipótese 1, de que mudanças no design de páginas de conteúdo (incluindo layout, usabilidade e critérios dos elementos de recomendação de conteúdo) impactam nas métricas de engajamento do usuário com o produto digital de notícias.

No segundo caso são comparadas as métricas de engajamento nos acessos iniciados pelas homepages *versus* os acessos iniciados pelo conteúdo (páginas de matéria). O objetivo deste comparativo é validar se os acessos iniciados diretamente pelo conteúdo como "página de destino" têm métricas de engajamento menores do que acessos iniciados pelas homepages dos sites de notícias (hipótese 2).

Já o terceiro caso faz um comparativo entre as métricas: "origem do tráfego" e "página de destino". É avaliado o impacto de cada uma delas no maior ou menor engajamento dos usuários. Neste caso, objetiva-se validar se a variável "origem do tráfego" é fator relevante na variação das métricas de engajamento dos sites de notícias (hipótese 2).

### 4.1.1 Comparativo entre páginas de matéria (antiga e nova) do site d'O GLOBO

Como já foi descrito no capítulo de procedimentos metodológicos, no processo de redesenho da página de matéria do site do jornal O Globo foram alterados diversos aspectos de layout e estrutura da página como um todo. Entretanto, neste momento serão detalhadas somente as alterações que envolvem os critérios para a recomendação de conteúdo, que é o foco dessa tese.

A figura 29 apresenta um exemplo da versão antiga da página de matéria do site do jornal O Globo (anterior ao redesenho ocorrido em 31/05/2014), com os elementos de recomendação de conteúdo sinalizados em amarelo e isoladamente destacados à direita da imagem. Nessa versão, 5 elementos podem ser identificados: "Agora em Destaque", "Mais Lidas", "Você pode estar interessado", "Blogs e Colunistas" e "Mais de Economia".

Figura 29: Página de matéria do site do jornal o Globo (versão antiga), com elementos de recomendação de conteúdo destacados.

Quando observados os critérios para a recomendação dos conteúdos, o elemento “Você pode estar interessado em” sugere notícias que estão diretamente relacionadas ao interesse previamente manifestado pelo usuário ao longo da navegação pelo site. Já o elemento "Mais de Economia" oferece conteúdos da mesma editoria da notícia acessada (economia). O elemento “Agora em Destaque” apresenta notícias da homepage do site, destacadas editorialmente. No elemento “mais lidas”, as notícias mais acessadas pelos demais usuários do site naquele dia são recomendadas. Por último, o elemento "Blogs e Colunistas" apresenta links para

conteúdos diversos sem nenhuma relação com o usuário, com o conteúdo ou com a navegação dos demais usuários do site.

Já a figura 30 apresenta um exemplo da página de matéria na nova versão do site do jornal O Globo, novamente com os elementos de recomendação de conteúdo sinalizados em amarelo e destacados à direita da imagem. Nessa versão, o número de elementos de recomendação foi ampliado para 6, bem como seu posicionamento e layout. Os elementos de recomendação destacados são: "Brasil", "Veja Também", "Anterior e Próxima", "Mais Populares", "Em destaque agora no Globo" e "Mais lidas".

Figura 30: Página de matéria do site do jornal o Globo (versão nova), com elementos de recomendação de conteúdo destacados.

Sobre os critérios utilizados para a recomendação, na comparação com a versão anterior da página, imediatamente percebe-se que o elemento "Blogs e Colunistas" foi retirado, assim como o elemento “Você pode estar interessado em”. Os elementos que recomendam mais notícias da mesma editoria acessada foram mantidos (“Brasil”), com a inclusão de um segundo elemento com o mesmo critério

de recomendação, chamado de “Veja Também”. Os elementos “Mais Lidas” e “Em destaque agora no Globo” se mantém sob os mesmos critérios da versão anterior. Além destes, são incluídos mais dois novos elementos de recomendação de conteúdo: “Mais Populares” (que utiliza o critério de sugerir notícias muito acessadas pela maioria dos usuários) e o elemento “Anterior e Próxima”, que atribui um elemento de cronologia trazendo notícias publicadas imediatamente antes e depois daquela que está sendo acessada.

Detalhadas as diferenças entre as duas versões de página de matéria no que diz respeito aos critérios utilizados para a recomendação de conteúdo, será apresentado um comparativo entre as métricas de engajamento de ambas as páginas, conforme procedimentos metodológicos definidos no capítulo anterior. Como resultados dessa análise comparativa, observou-se que a nova versão apresentou melhora nas três métricas de engajamento definidas, a saber: “taxa de rejeição”, “páginas por sessão” e “duração média da navegação por sessão” (figura 31).

Figura 31: Comparação das métricas de acesso ao site do globo a partir da página de matéria como página de destino (Versão antiga X Versão atual).

Como pode ser visto no gráfico acima, a taxa de rejeição do site (*Bounce Rate*) caiu de 80,08% para 75,94% na comparação com a versão antiga da página de matéria, o que significou uma melhora de 5,18%. Da mesma forma, o número de páginas por sessão (*pages/session*) para os usuários que acessam o site a partir da nova página de matéria aumentou de 1,32 para 1,42, resultando em 7,48% de melhora. Já em relação a duração média da navegação por sessão (*Avg. Session Duration*), foi possível observar uma melhora significativa de 14,54% (duração média aumentou de 01:41 minutos para 01:55 minutos por usuário).

A partir do resultado da análise comparativa feita sob condições semelhantes (mesmo período de tempo e mesma época do ano), em um recorte de tempo consideravelmente longo e com um grande número de acessos constitui uma situação onde é possível considerar que houve uma alteração consistente nas métricas de engajamento do site (no caso específico, uma melhora nas três métricas avaliadas). Dessa maneira, os números ajudam a corroborar a hipótese de que mudanças no design das páginas de matéria, incluindo os critérios por trás dos elementos de recomendação de conteúdo são fatores chave no engajamento dos usuários com o site (para melhor ou pior). Tais modificações (sejam no layout, na usabilidade ou nos critérios para a oferta de conteúdo dos elementos de recomendação da página de matéria) foram capazes de afetar o comportamento do usuário, o que se refletiu na melhora entre as métricas de engajamento do site.

A melhora identificada nas três métricas avaliadas trazem indícios de uma maior e melhor utilização dos elementos de recomendação de conteúdo na nova versão de página de matéria, contribuindo para que os usuários tenham se mantido por mais tempo no site de notícias, acessando um número maior de páginas por visita. Entretanto, esta última questão não pode ser considerada conclusiva somente a partir dos levantamentos feitos para este estudo.

### 4.1.2 Comparativo entre as métricas de engajamento nos acessos iniciados pelas homepages *versus* os acessos iniciados pelas páginas de matéria

No segundo caso, conforme descrito no capítulo de procedimentos metodológicos, são comparadas as métricas de engajamento dos usuários que acessaram o site do jornal o Globo tendo como “página de destino” a homepage do site com as métricas de usuários que iniciaram seu acesso diretamente por páginas de conteúdo diversas. Em ambos os casos, a versão do site avaliada é a nova versão (pós redesenho).

As figuras abaixo representam a interface da ferramenta “Google Analytics” com as métricas de engajamento para quem acessou o site do Globo tendo com página de destino a homepage (figura 32) ou uma matéria (figura 33). Optou-se por

manter o idioma inglês original da ferramenta para padronizar a compreensão dos termos.

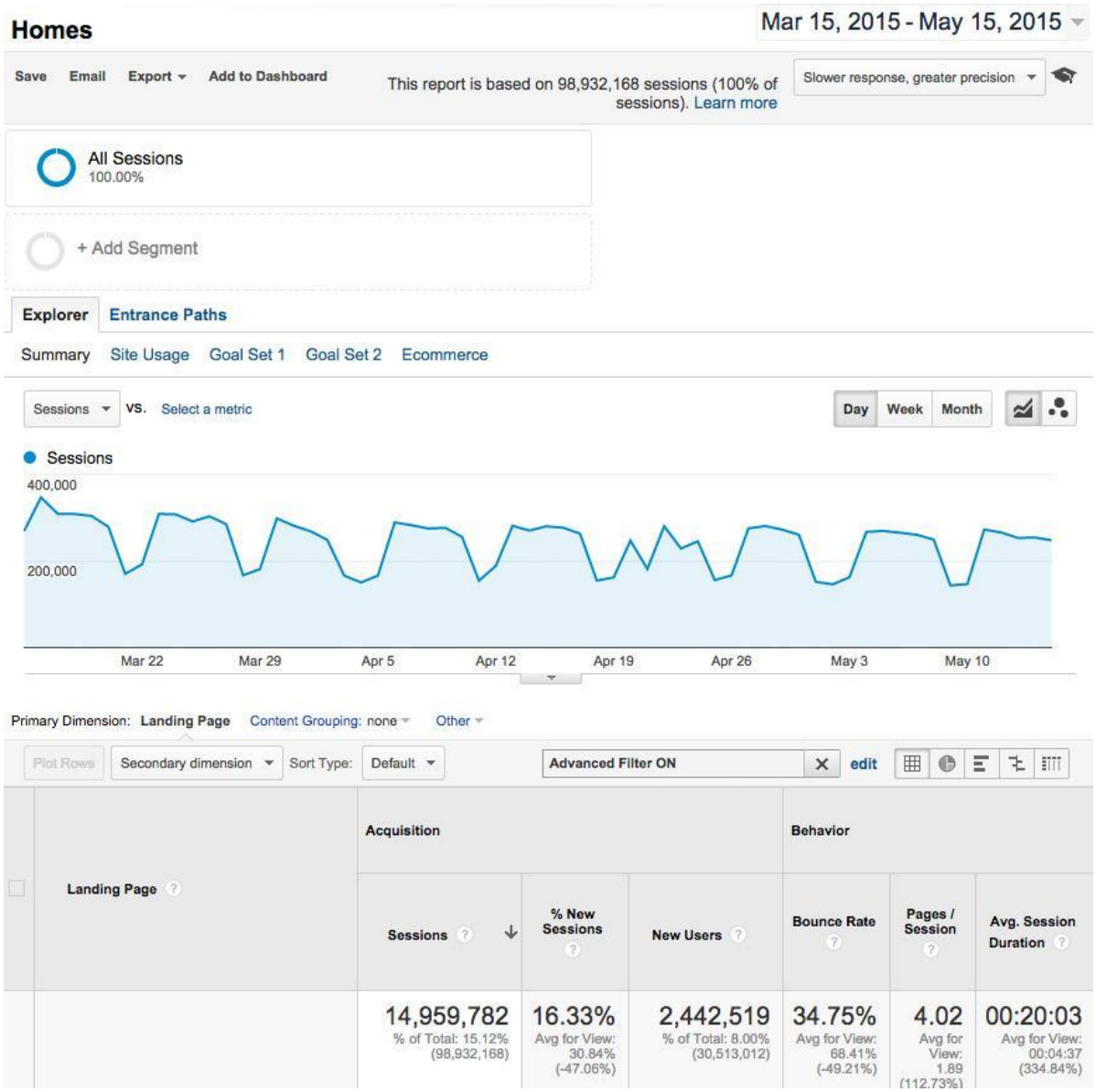


Figura 32: Métricas do site do Globo entre acessos tendo como página de destino a homepage.

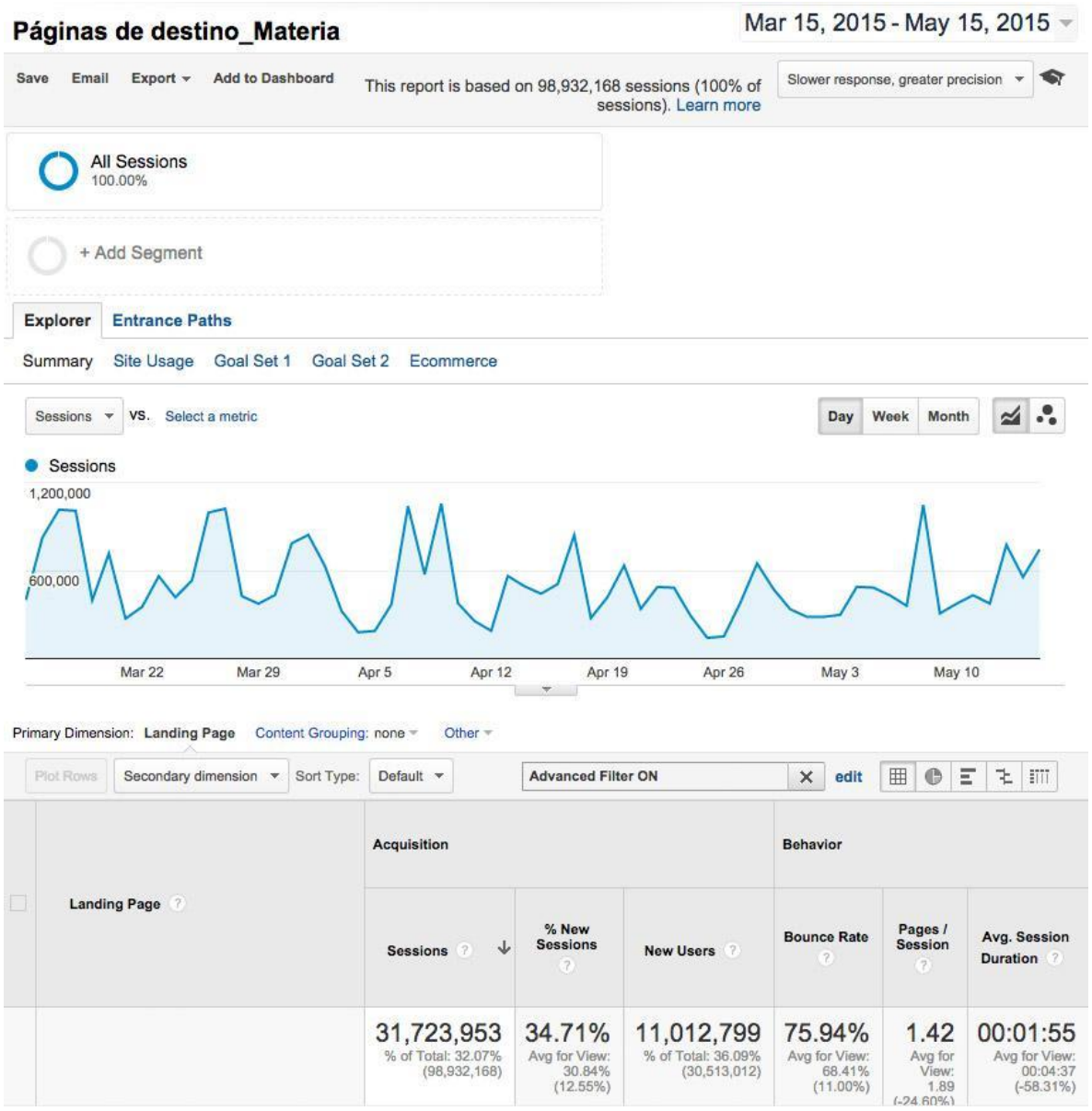


Figura 33: Métricas do site do Globo entre acessos tendo matérias como página de destino.

Apesar da melhora consistente nas métricas da “nova” página de matéria do Globo (apresentada no subcapítulo anterior), observa-se que quando a comparação é com os acessos iniciados a partir das homepages do site, o engajamento médio dos usuários continua relativamente baixo em acessos originados a partir das páginas de matéria, como pode ser visto no comparativo da figura 34.

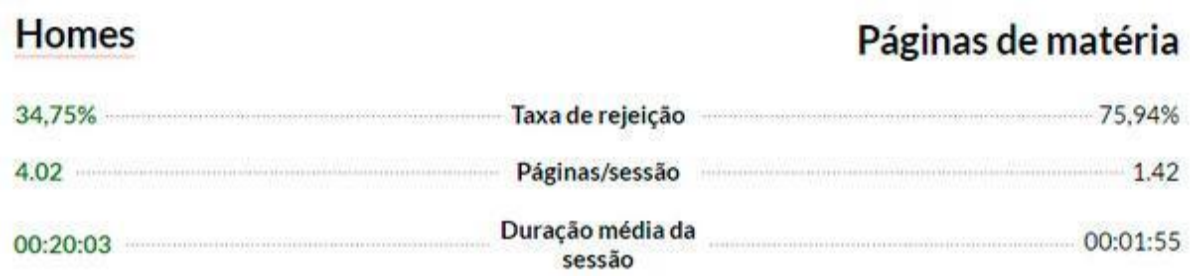


Figura 34: Consolidação dos dados comparativos entre acessos a partir das homepages versus acessos a partir de matérias em um mesmo período de tempo.

Nos resultados consolidados dos gráficos das figuras 32 e 33 que são apresentados na figura 34, observa-se que a qualidade do engajamento dos usuários que iniciam o acesso ao site a partir das páginas de matéria é substancialmente menor em relação a quem acessa a partir das homepages. Quando a própria matéria ou conteúdo é a *página de destino*, a taxa de rejeição dos usuários é de 75,94% (contra 34,75%), ou seja, cerca de 2,2 vezes maior. Já o número de páginas por sessão de quem acessa a partir das páginas de matéria é quase 3 vezes menor. Em relação a duração média por sessão, a diferença é ainda mais acentuada, sendo quase 10 vezes menor o tempo de navegação de quem acessa o site a partir de uma matéria.

A partir destes resultados, no contexto da análise comparativa proposta, foi possível validar que os acessos iniciados diretamente pelas páginas de conteúdo como “página de destino” apresentam métricas de engajamento piores do que os acessos iniciados pelas homepages (hipótese 2). É importante destacar que a validação se dá de forma parcial, pois foi verificada apenas em cima de um caso específico (site do jornal o Globo).

### 4.1.3 Análise do impacto no engajamento de usuários em sites de notícias a partir das variáveis “Origem do Tráfego” e “Página de destino”

Nesta terceira etapa da análise de métricas, conforme descrito no capítulo de procedimentos metodológicos, será avaliado o impacto das variáveis “origem do tráfego” e “página de destino” nas métricas de engajamento do site do jornal o Globo.

Inicialmente, foram identificadas as principais “origens de tráfego” do site, apresentadas no gráfico da figura 35:

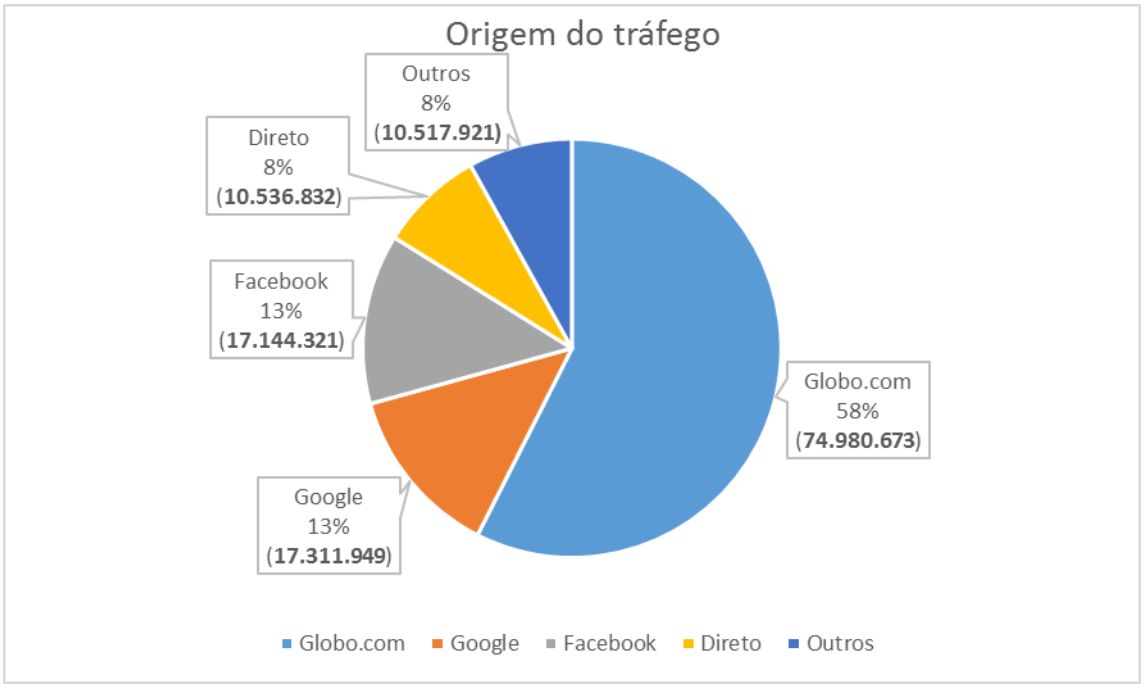


Figura 35 – Origem do tráfego no site d’O Globo (número de sessões).
Fonte: "Elaborado pelo autor, com base na pesquisa realizada".

Conforme apresentado, a maior parte dos acessos ao site do Globo (58%) tem origem no site Globo.com (que apesar da nomenclatura semelhante, é um outro produto, com experiências completamente diferentes, sendo tecnicamente, portanto, uma origem externa). A percepção do usuário em relação aos sites serem de uma mesma empresa é irrelevante tecnicamente para a questão avaliada, uma vez que do total de acessos tendo como origem a Globo.com, somente 5% tem como página de destino uma homepage; os 95% restantes tem como destino uma página de matéria. Por conta disso, o comportamento de quem acessa o site do jornal O Globo a partir da origem “Globo.com” é similar aos acessos que tem por origem o facebook, como veremos a seguir.

A segunda maior origem de tráfego do site do jornal o Globo é o buscador “Google” (google.com), que aparece como origem para 13% do tráfego total. Destes, 42% tem como destino homepages do site d’O Globo, enquanto que os 58% restantes acessam diretamente via páginas de matéria.

A terceira maior origem de tráfego é a rede social Facebook (facebook.com), com os mesmos 13% do segundo colocado. Do total desses acessos, somente 7% tem como destino homepages; o restante (93%) se dá diretamente tendo como página de destino uma matéria ou conteúdo específico.

Em quarto colocado está o chamado “tráfego direto”, que em sua maioria representa os acessos onde o usuário digitou o endereço da página diretamente no navegador. Este segmento representou 8% dos acessos. Do total, 60% deles foram destinados para as homepages, enquanto os 40% restantes foram destinados a páginas de matéria.

Os resultados dessa análise sobre as páginas de destino por origem de tráfego detalhada acima foram consolidados em uma tabela para facilitar a visualização (tabela 1):

**Tabela 1 – % de sessões por tipo de página de destino e por origem do tráfego**

| **Origem do tráfego** | **Página de destino: Homepage** | **Página de destino: Conteúdos** |
|---|---|---|
| Globo.com | 5% | 95% |
| Google | 42% | 58% |
| Facebook | 7% | 93% |
| Direto | 60% | 40% |

Fonte: Métricas do Google analytics do site oglobo.com.br, no período de 3 meses (15/10/2015 a 15/01/2016).

A seguir, a tabela 2 apresenta uma análise inicial das 3 métricas selecionadas (taxa de rejeição, páginas por sessão e duração média da sessão), segmentada pelas mesmas 4 opções de “origem do tráfego” (direto, Facebook, Google e Globo.com) no site d’O Globo.

**Tabela 2 – Métricas por origem do tráfego**

| Origem do Tráfego | Taxa de rejeição | Páginas p/ sessão | Duração média da sessão |
|---|---|---|---|
| Direto | 51,76% | 2,87 | 11:45 |
| Facebook | 78,43% | 1,44 | 2:35 |
| Globo.com | 66,09% | 1,59 | 2:34 |
| Google | 52,58% | 2,80 | 9:05 |

Fonte: Métricas do Google analytics do site oglobo.com.br, no período de 3 meses (15/10/2015 a 15/01/2016).

Como é possível observar, os acessos via tráfego direto e via Google são os que apresentam os melhores resultados (menor taxa de rejeição, maior quantidade de páginas por sessão e mais tempo de navegação por sessão). Já os acessos originados por Globo.com e Facebook apresentam as piores taxas de engajamento, com uma taxa de rejeição que chega a ser quase 30% maior em valores absolutos, com a metade do número de páginas por sessão e com um tempo de navegação de cerca de 20% do registrado nos acessos via tráfego direto.

A partir desse resultado preliminar, seria plausível concluir, por exemplo, que o Facebook como origem do tráfego impacta no (baixo) engajamento de um site de notícias (conforme tabela 2).

Entretanto, de acordo com o que foi apresentado na tabela 1, 95% dos acessos originados pelo Facebook tem como destino uma página de conteúdo. Ao mesmo tempo, o acesso via tráfego direto (que apresenta as melhores métricas de engajamento) normalmente tem como destino uma homepage (60%).

Desta maneira, a primeira parte da análise não foi considerada conclusiva para a validação de que a “origem do tráfego” seria necessariamente um fator de impacto no engajamento do site do jornal o Globo.

Por conta disso, fez-se necessário um aprofundamento no estudo das métricas a fim de compreender se as diferentes origens de tráfego são realmente a causa de um maior ou menor engajamento por parte dos usuários, ou se a influência nas métricas se dá por conta da página de destino (homepage ou matéria), conforme apresentado no subcapítulo anterior (4.1.2).

Dando sequência a análise, o próximo passo foi detalhar as métricas de engajamento em cada uma das 4 origens de tráfego, mas segmentando pela variável “página de destino”. Conforme os procedimentos metodológicos apresentados, foram definidos dois parâmetros: “homepage” (quando a página de destino é uma homepage) e “conteúdo” (quando a página de destino é um conteúdo).

A tabela 3 abaixo apresenta as métricas: “taxa de rejeição”, páginas por sessão” e “duração média da sessão”. Foram selecionados somente os acessos identificados como tráfego direto, segmentado também pela variável “página de destino” (por homepage X conteúdos).

**Tabela 3 – Tráfego Direto (site desktop O Globo).**

| | Página de destino: Homepage | Página de destino: Conteúdos |
|---|---|---|
| Taxa de Rejeição | 39,62% | 61,09% |
| Páginas por Sessão | 3,59 | 1,70 |
| Duração Média da Sessão | 18:10 | 2:35 |

Fonte: Métricas do Google analytics do site oglobo.com.br, no período de 3 meses (15/10/2015 a 15/01/2016).

De acordo com a tabela 3, com a mesma origem de tráfego (tráfego direto), as métricas de engajamento são consistentemente melhores quando a página de destino é uma homepage e não uma matéria do site. Em relação a "taxa de rejeição", em valores absolutos ela é quase 22% menor quando o acesso se dá via homepages. Já o número de páginas por seção é cerca de duas vezes maior (3,59 contra 1,70 dos acessos por páginas de matéria). Em relação a "duração média por seção", a diferença encontrada é ainda mais acentuada: enquanto os acessos que se iniciam pela homepage têm mais de 18 minutos de duração, os acessos iniciados pelas páginas de matéria tem apenas 2:35 minutos de duração.

A tabela 4 apresenta as métricas "taxa de rejeição", páginas por sessão" e "duração média da sessão" quando o acesso se origina via Facebook, segmentando novamente pela variável "página de destino" (por *homes* X conteúdos).

**Tabela 4 – Tráfego via Facebook (site desktop O Globo).**

| | Página de destino: Homepage | Página de destino: Conteúdos |
|---|---|---|
| Taxa de Rejeição | 36,64% | 63,47% |
| Páginas por Sessão | 3,92 | 1,32 |
| Duração Média da Sessão | 24:13 | 1:23 |

Fonte: Métricas do Google analytics do site oglobo.com.br, no período de 3 meses (15/10/2015 a 15/01/2016).

De acordo com a tabela 4, com a mesma origem de tráfego (Facebook), as métricas de engajamento são consistentemente melhores quando a página de destino é uma homepage e não uma matéria do site. As proporções são similares ao identificado na tabela 3.

A tabela 5 apresenta as métricas: "taxa de rejeição", páginas por sessão" e "duração média da sessão" quando o acesso se origina via Globo.com, segmentando pela variável "página de destino" (por *homes* X conteúdos).

**Tabela 5 – Tráfego via Globo.com (site desktop O Globo).**

| | Página de destino: *Homes* | Página de destino: Conteúdos |
|---|---|---|
| Taxa de Rejeição | 36,02% | 55,99% |
| Páginas por Sessão | 4,03 | 1,41 |
| Duração Média da Sessão | 20:38 | 1:43 |

Fonte: Métricas do Google analytics do site oglobo.com.br, no período de 3 meses (15/10/2015 a 15/01/2016).

De acordo com a tabela 5, com a mesma origem de tráfego (Globo.com), as métricas de engajamento novamente são consistentemente melhores quando a página de destino é uma homepage e não uma matéria do site.

Já a tabela 6 apresenta as métricas “taxa de rejeição”, páginas por sessão” e “duração média da sessão” quando o acesso se origina via Google, segmentando pela variável “página de destino” (por homepage ou por conteúdos).

**Tabela 6 – Tráfego via Google (site desktop O Globo).**

| | **Página de destino: Homepage** | **Página de destino: Conteúdos** |
|---|---|---|
| Taxa de Rejeição | 36,02% | 67,47% |
| Páginas por Sessão | 3,89 | 1,51 |
| Duração Média da Sessão | 18:11 | 1:35 |

Fonte: Métricas do Google analytics do site oglobo.com.br, no período de 3 meses (15/10/2015 a 15/01/2016).

De acordo com a tabela 6, com a mesma origem de tráfego (Google), as métricas de engajamento novamente são consistentemente melhores quando a página de destino é uma homepage e não uma matéria do site.

Para deixar clara essa consistência entre métricas de engajamento sempre melhores nos acessos que tem como página de destino uma homepage (em comparação com os acessos via páginas de matéria) independente da “origem do tráfego”, a seguir serão apresentados os gráficos que consolidam esses resultados para cada uma das métricas de engajamento.

O gráfico da figura 36 consolida a métrica “taxa de rejeição” em um comparativo homepage X conteúdos, por cada uma das 4 origens de tráfego. Como pode ser observado, independente da origem do tráfego, a taxa de rejeição de quem inicia seu acesso por uma homepage apresenta percentuais bastante reduzidos

(com pequenas variações entre cada uma das origens analisadas). Da mesma forma, quando o acesso se inicia por páginas de conteúdo, a taxa de rejeição é consistentemente elevada (ou seja, muitos usuários abandonam rapidamente o site), mesmo que novamente com pequenas oscilações.

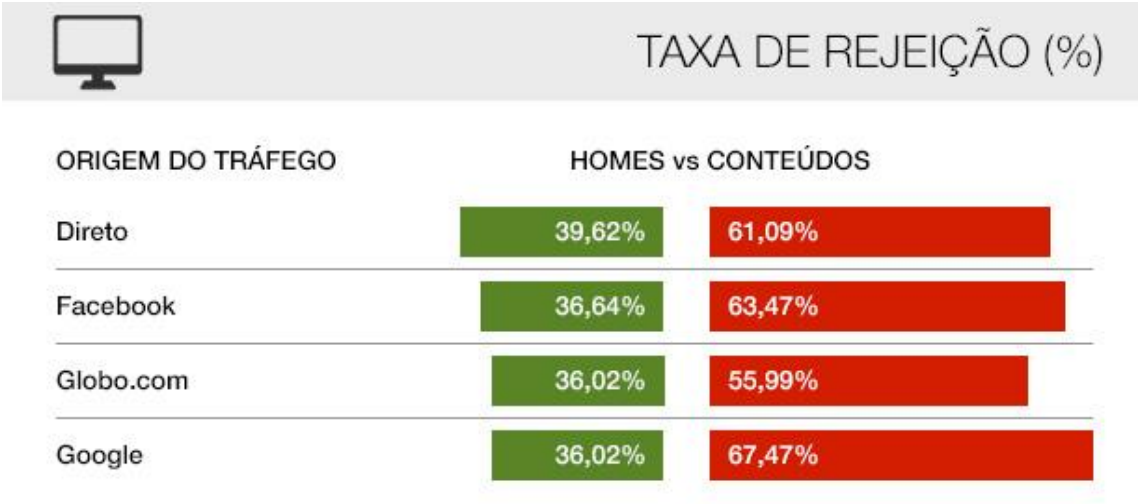


Figura 36 – Métrica “Taxa de rejeição” do site d’O Globo por acessos iniciados via homepages e por acessos iniciados por páginas de conteúdo, segmentados pela “origem do tráfego”.
Fonte: "Elaborado pelo autor, com base na pesquisa realizada".

Nesse caso, é possível observar que mesmo acessando via tráfego direto (onde, a princípio, o engajamento seria sempre maior), a taxa de rejeição é muito elevada quando um conteúdo é a página de destino. Da mesma forma, quando o acesso se dá pela origem “Facebook” (onde o engajamento seria sempre menor), a taxa de rejeição é muito baixa sempre que a página de destino do usuário é uma homepage, e não um conteúdo.

Já o gráfico da figura 37 consolida a métrica “páginas por sessão” em um comparativo homepages X conteúdos, por cada uma das 4 origens de tráfego. A exemplo da figura 4, novamente observa-se que independente da origem, o número de páginas por sessão de quem inicia seu acesso por uma homepage é bastante elevado em relação a quem acessa por páginas de matéria.

Figura 37 – Métrica “páginas por sessão” do site d’O Globo por acessos iniciados via homepages e por acessos iniciados por páginas de conteúdo, segmentados pela “origem do tráfego”.
Fonte: "Elaborado pelo autor, com base na pesquisa realizada".

A exemplo do que foi observado anteriormente, a origem do tráfego traz algumas pequenas variações, mas é a diferença na página de destino que provoca o maior impacto na métrica.

No gráfico apresentado na figura 38, a métrica “duração média da sessão” é apresentada em um comparativo homepages X conteúdos, por cada uma das 4 origens de tráfego. Mais uma vez, independente da origem via tráfego direto, Facebook, Globo.com ou Google, a duração média da sessão de quem inicia seu acesso via homepages é consistentemente maior. Quando a página de destino é uma matéria, a duração média da sessão é em média 15 vezes menor.

Figura 38 – Métrica “Duração média da sessão” do site d’O Globo por acessos iniciados via homepages e por acessos iniciados por páginas de conteúdo, segmentados pela “origem do tráfego”.
Fonte: "Elaborado pelo autor, com base na pesquisa realizada".

Como ressaltado na análise das 2 métricas anteriores, a origem do acesso tem baixa influência no maior ou menor engajamento dos usuários. Nas 3 métricas analisadas, percebe-se que é a página de destino (homepage ou matéria) o maior fator de impacto nos padrões de engajamento identificados.

Com base nessa análise, apesar de ser verdadeira a afirmação de que origens de tráfego diferentes apresentam métricas de engajamento diferentes, e que o acesso via tráfego direto apresenta métricas muito melhores em relação às demais origens de tráfego, a causa desse engajamento não está somente na “origem do tráfego”. Ao segmentar os acessos pela variável “página de destino”, consistentemente observa-se que uma sessão iniciada por homepages apresentou métricas melhores do que uma sessão iniciada por conteúdos, independente da origem.

Desta forma, ao final das análises foi possível validar parcialmente a segunda hipótese proposta nessa tese no caso específico do site do jornal O Globo. Com relação a variável “página de destino”, observou-se que de fato as métricas de engajamento foram melhores em acessos iniciados pelas homepages em comparação aos acessos iniciados por páginas de conteúdo. Entretanto, em relação a variável “origem do tráfego”, observou-se que seu impacto nas métricas de engajamento dos sites de notícias foi relativamente baixo. Apesar de ter demonstrado haver alguma influência da “origem do tráfego”, a variável “pagina de destino” se mostrou mais impactante em um maior ou menor engajamento do usuário com o site do jornal O Globo. É importante destacar que, por dificuldades técnicas e limitação de recursos, não foi possivel fazer uma análise mais abrangente contemplando outras ferramentas e sites de notícias, mas de forma experimental foi possível validar parcialmente a hipótese no caso em questão.

## 4.2 **Pesquisa com Usuários**

Conforme indicado no capítulo sobre os procedimentos metodológicos, neste item será detalhada a pesquisa realizada com usuários acerca da influência do contexto no consumo de notícias. A seguir, serão analisados também os resultados encontrados.

### 4.2.1 Questionário de avaliação da influência do contexto no consumo de notícias

Nesta etapa de pesquisa, objetivou-se validar a hipótese 3 dessa tese, que aponta para a influência do contexto no consumo de notícias. Para isso, foi realizada uma pesquisa exploratória com usuários em formato de questionário. Nesse momento, cabe destacar que todos os resultados obtidos a partir da aplicação desse tipo de técnica devem considerar a possibilidade do usuário responder algo diferente do que ele de fato faria em uma situação real.

Do ponto de vista técnico, para a criação do questionário foi utilizada a ferramenta de uso gratuito “Google Forms”, solução do Google projetada justamente para este tipo de abordagem. O questionário foi aplicado de maneira aberta e online pelas redes sociais. Ao todo, 32 participantes responderam às perguntas propostas no questionário. Tal número de participantes foi considerado satisfatório, a medida que não se objetivava identificar padrões comportamentais, mas sim a proposição de cenários acerca das respostas individuais de cada participante.

Com relação a estrutura, o questionário iniciava-se com 2 perguntas para identificação do perfil de cada respondente: “Endereço de email” e “Principal profissão que exerce atualmente”. Ao todo, entre os 32 respondentes foram identificadas 27 profissões diferentes, o que aponta para uma boa diversidade de perfis. Se a profissão em si é pouco relevante para a sequência da pesquisa, a comprovação da diversidade de perfis ajuda a dirimir possíveis direcionamentos nas respostas.

Na sequência, foram apresentadas 5 manchetes de notícias e/ou artigos replicadas de diferentes sites jornalísticos. O texto utilizado para esta apresentação foi: “A partir de agora, serão apresentadas 5 manchetes de notícias e/ou artigos de sites da internet. Serão feitas perguntas sobre o seu interesse em cada uma delas.”

Cada uma das notícias foi apresentada individualmente, seguida de duas perguntas sobre ela, conforme a seguir: 1) a primeira dizia respeito ao interesse do usuário por aquela notícia – “Você tem interesse por essa notícia? Avalie em uma escala de 1 a 5, sendo 1 = “nenhum interesse” e 5 = “estou louco para ler essa notícia””; 2) já a segunda pergunta introduzia o elemento contextual na decisão por consumir ou não aquela notícia. Dessa maneira, a pergunta feita foi: “Em quais dos seguintes momentos você teria grandes possibilidades de consumir essa notícia? (marque quantas opções quiser)”. As opções de resposta oferecidas foram: a) “em casa”; b) “no trabalho”; e c) “em uma viagem”.

Em relação as 5 notícias selecionadas, optou-se por utilizar temas de editorias distintas entre si, para evitar que o resultado da pesquisa pudesse ser interpretado como limitado a apenas um determinado assunto. As manchetes selecionadas dentro de cada editoria foram escolhidas de forma livre pelo autor.

Sendo assim, as notícias foram: 1) “Nudez ou pornografia: quais são os limites para a arte?” (editoria de cultura); 2) “Com a queda da taxa Selic, qual a melhor maneira de investir o seu dinheiro?” (editoria de economia); 3) “Novo escândalo de corrupção pode ser o maior da história” (editoria de política); 4) “Ranking indica as 10 melhores cidades do mundo para quem quer viajar” (editoria de viagem); e 5) “Dicas para melhorar a qualidade do seu sono” (editoria de saúde).

### 4.2.2 Resultados consolidados do questionário

A seguir, serão apresentados os resultados consolidados da pesquisa para cada uma das notícias:

1) **"Nudez ou pornografia: quais são os limites para a arte?"**: Em relação a 1ª notícia apresentada ("Nudez ou pornografia: quais são os limites para a arte?"), a figura 39 apresenta os resultados referentes ao nível de interesse manifestado pela notícia. Pelo gráfico, observa-se na média um interesse intermediário por parte dos respondentes. Entretanto, um número relativamente grande de participantes declarou não ter nenhum interesse na notícia (28%).

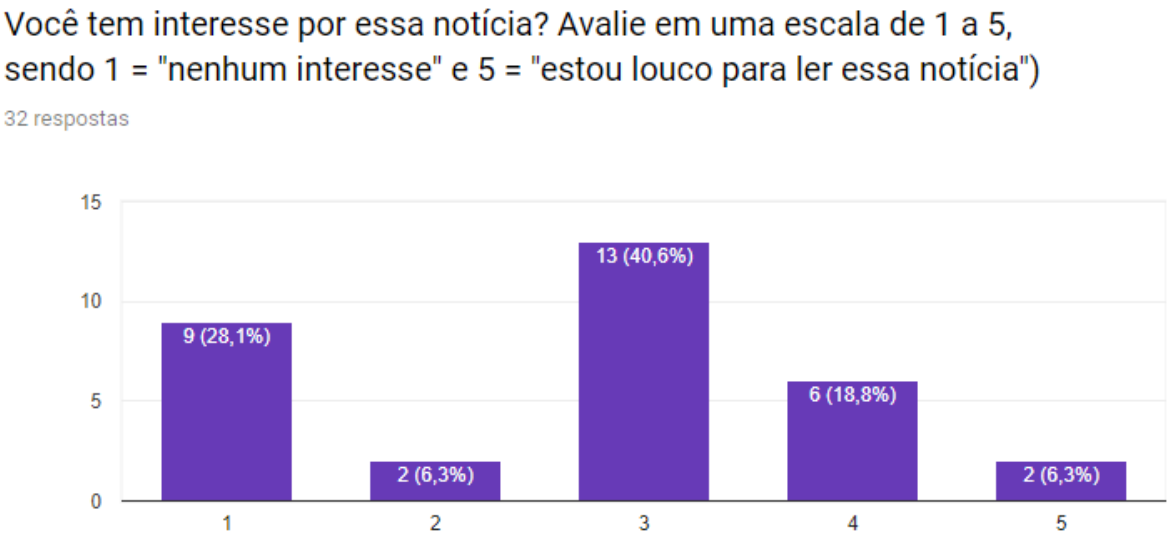


Figura 39: Média do interesse dos participantes da pesquisa pela 1ª notícia apresentada.

Sobre a proposição dos momentos em que o usuário julgou ter maior possibilidade de consumir essa notícia, a opção “em casa” foi amplamente mais votada (80%) em comparação com as opções “no trabalho” ou “em uma viagem” que apresentaram somente 17,2% de possibilidades de consumo da mesma notícia, conforme figura 40.

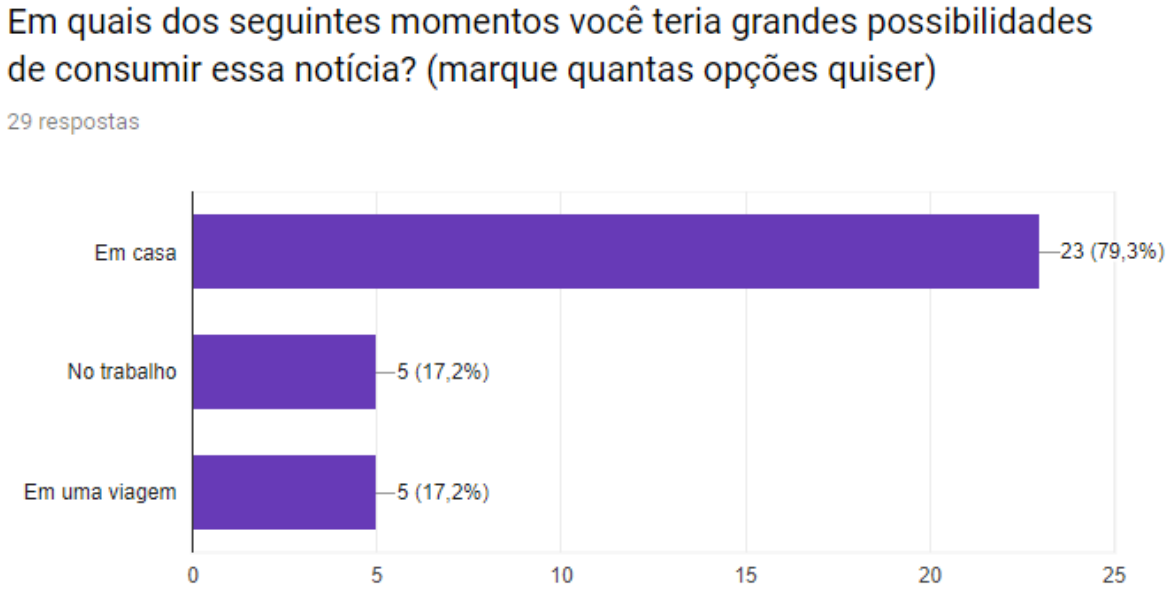


Figura 40: Percentual dos momentos escolhidos como de grande possibilidade para o consumo da notícia.

Neste sentido, pelos critérios estabelecidos para a análise, ficou evidenciado que em relação a essa notícia em específico, o contexto exerceu grande influência nas possibilidades de um usuário consumir essa notícia. Dessa maneira, em um elemento de recomendação de conteúdo, a sugestão desta notícia quando o usuário está “em casa” poderia resultar em uma assertividade quase 5 vezes maior do que nos demais contextos apresentados.

2) “**Com a queda da taxa selic, qual a melhor maneira de investir o seu dinheiro?”**: Em relação a 2ª notícia apresentada, a figura 41 apresenta os resultados referentes ao nível de interesse demonstrado pelos respondentes. Neste caso, observa-se na média um interesse também intermediário, porém com um representativo percentual de avaliações indicando grande interesse na notícia. (Cerca de 40% dos participantes responderam com notas 4 e 5 na avaliação do interesse pela notícia).

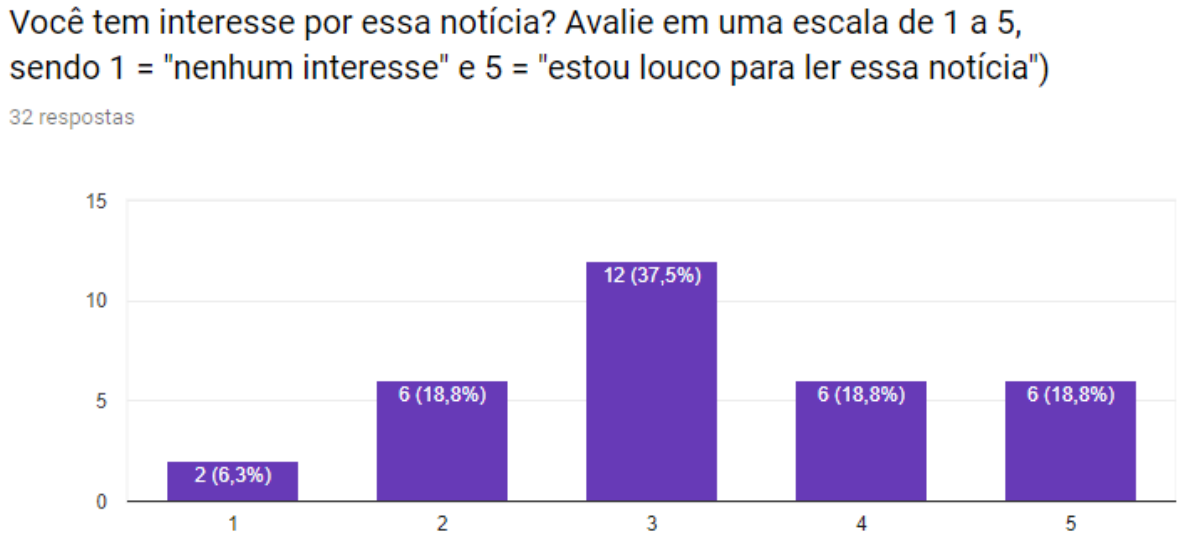


Figura 41: Média do interesse dos participantes da pesquisa pela 2ª notícia apresentada.

Sobre os momentos em que o usuário julgou ter maior possibilidade de consumir essa notícia, a opção “em casa” foi novamente a mais escolhida, com cerca de 78% das respostas. Entretanto, a opção “no trabalho” chegou também a quase 70% das respostas, enquanto a opção “em uma viagem” foi escolhida somente por 18,8% dos participantes, conforme figura 42.

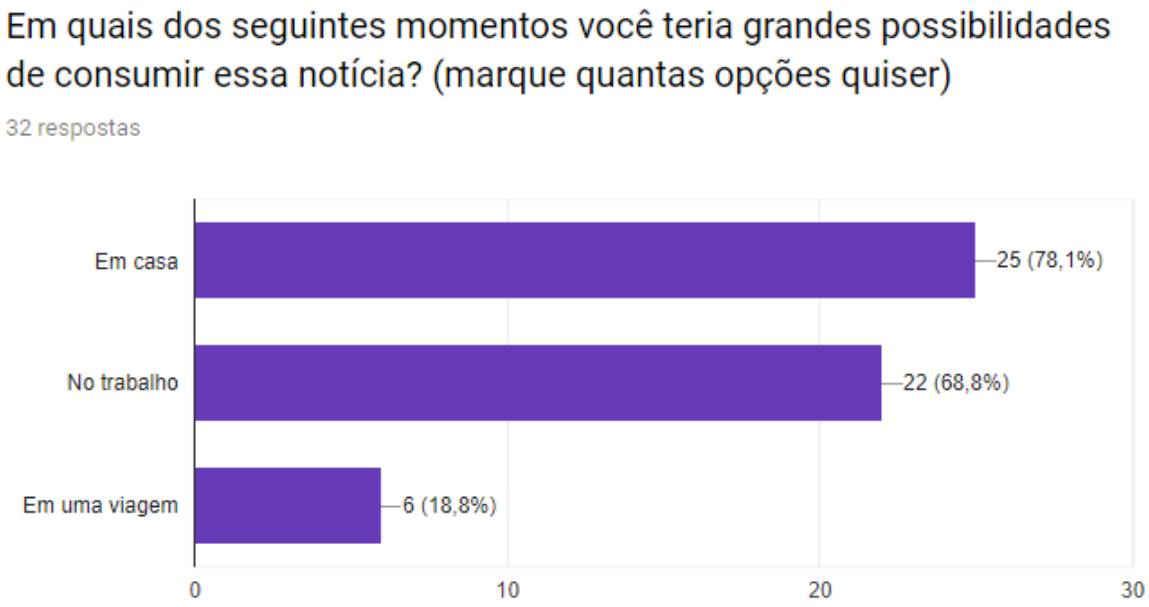


Figura 42: Percentual dos momentos escolhidos como de grande possibilidade para o consumo da notícia.

Pelos critérios estabelecidos no capítulo de procedimentos metodológicos para esta análise, novamente pode-se perceber que o contexto é fator de impacto na possibilidade de um usuário consumir essa notícia. Neste caso, em um elemento de recomendação de conteúdo, a sugestão dessa notícia traria bons resultados quando o usuário estivesse “em casa” ou “no trabalho”, mas baixa assertividade quando estivesse “em uma viagem”.

3) **"Novo escândalo de corrupção pode ser o maior da história**": Na análise das respostas referentes a 3ª notícia apresentada, a figura 43 apresenta os resultados do nível de interesse da média de participantes da pesquisa. Para esta notícia, foi possível observar um interesse acima da média das duas anteriores, com a maioria dos respondentes indicando as pontuações máximas oferecidas (cerca de 63% dos participantes responderam com notas 4 e 5 na avaliação do interesse pela notícia).

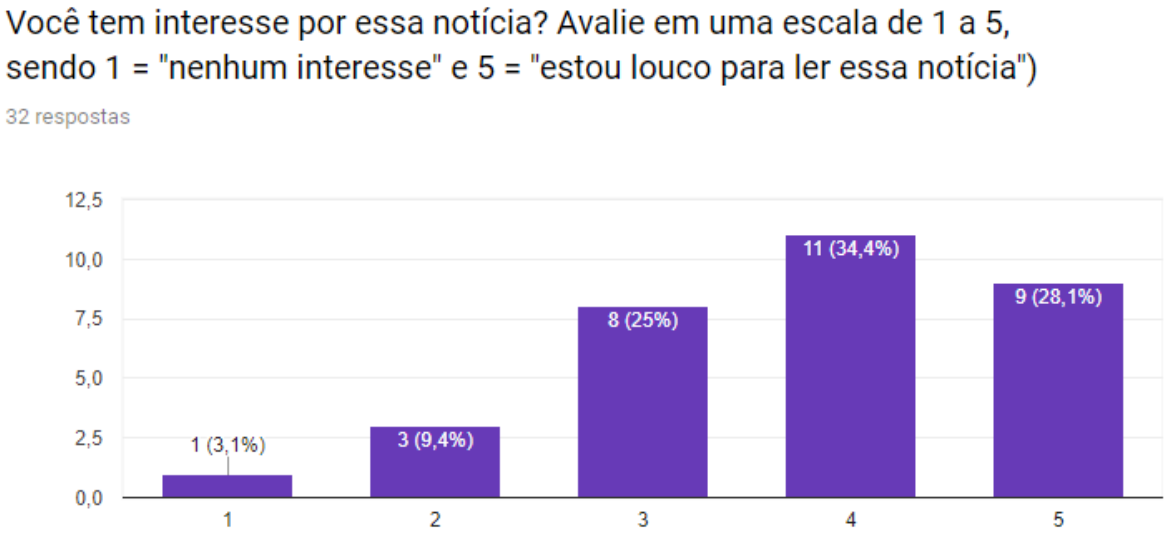


Figura 43: Média do interesse dos participantes da pesquisa pela 3ª notícia apresentada.

Em relação aos momentos que o usuário julgou mais propícios para o consumo dessa notícia, a figura 44 mostra que a opção “em casa” foi amplamente escolhida com cerca de 81% das respostas, mas novamente seguida de perto pela opção “no trabalho” com quase 72% de respostas. A opção “em uma viagem” permaneceu com um percentual relativamente baixo de escolhas (28,1%), porém superior ao apresentado nas duas notícias anteriores.

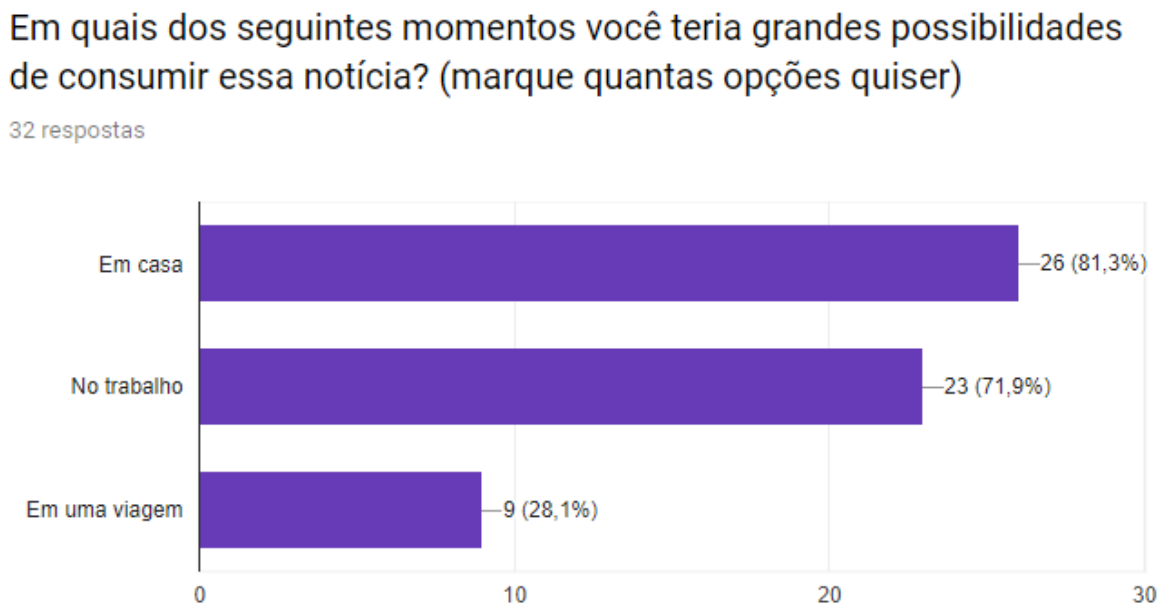


Figura 44: Percentual dos momentos escolhidos como de grande possibilidade para o consumo da notícia.

Mais uma vez, a diferença percentual entre as três opções de resposta propostas ficou bastante acentuada, evidenciado que cada um dos contextos apresentados resulta em diferentes possibilidades de consumir ou não uma notícia. Em um elemento de recomendação de conteúdo, essa notícia teria excelente assertividade nos usuários que estivessem “em casa” ou “no trabalho”, mas ainda apresentaria um baixo índice de engajamento para usuários que estivessem “em uma viagem”.

4) **"Ranking indica as 10 melhores cidades do mundo para quem quer viajar":** Para esta notícia, foi possível observar na média o maior nível de interesse entre as 5 notícias apresentadas (nenhum usuário marcou a opção 1 como resposta e apenas e quase 35% indicaram ter interesse máximo na notícia).

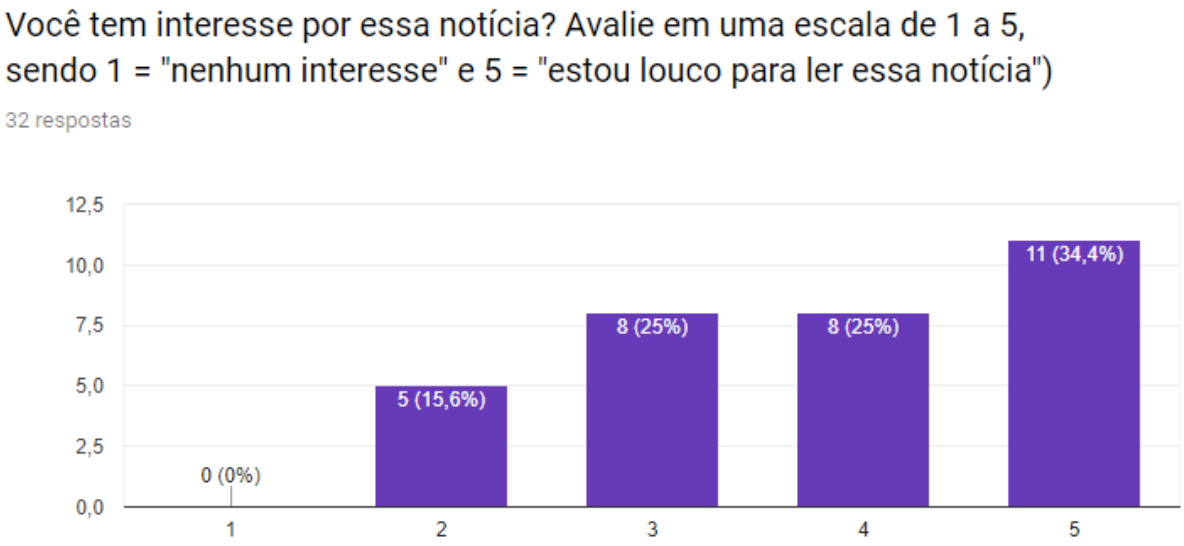


Figura 45: Média do interesse dos participantes da pesquisa pela 4ª notícia apresentada.

Sobre os momentos com grandes possibilidades para o consumo da notícia, novamente a opção “em casa” apresentou uma diferença acentuada em relação as demais, com quase 91% das respostas. A opção “no trabalho” obteve cerca de 53%, enquanto a opção “em uma viagem” permaneceu com o pior percentual entre as três opções contextuais (43,8%). Entretanto, entre todas as notícias, essa foi a que

apresentou o maior percentual de respostas indicando que consumiriam o conteúdo em uma viagem (figura 46).

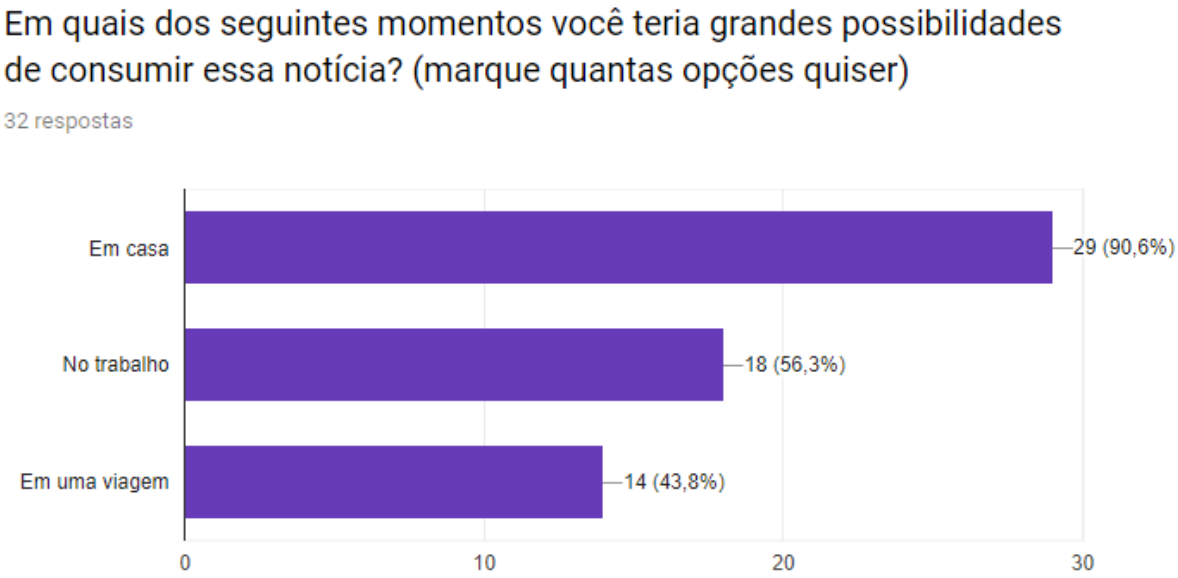


Figura 46: Percentual dos momentos escolhidos como de grande possibilidade para o consumo da notícia.

A exemplo das demais notícias, novamente ficou demonstrado que o contexto é um fator de grande impacto na possibilidade de um usuário consumir uma determinada notícia. Nos elementos de recomendação de sites de notícias, pode-se dizer que este conteúdo iria performar muito bem para usuários no contexto de uso “em casa”, e razoavelmente nos demais contextos apresentados.

5) **"Dicas para melhorar a qualidade do seu sono":** Em relação a 5ª e última notícia apresentada, a figura 47 apresenta os resultados a respeito do interesse demonstrado pelos participantes. Neste caso, observa-se na média um interesse intermediário, com a maioria das avaliações oscilando entre 3 e 4 (cerca de 53% dos participantes responderam desta maneira).

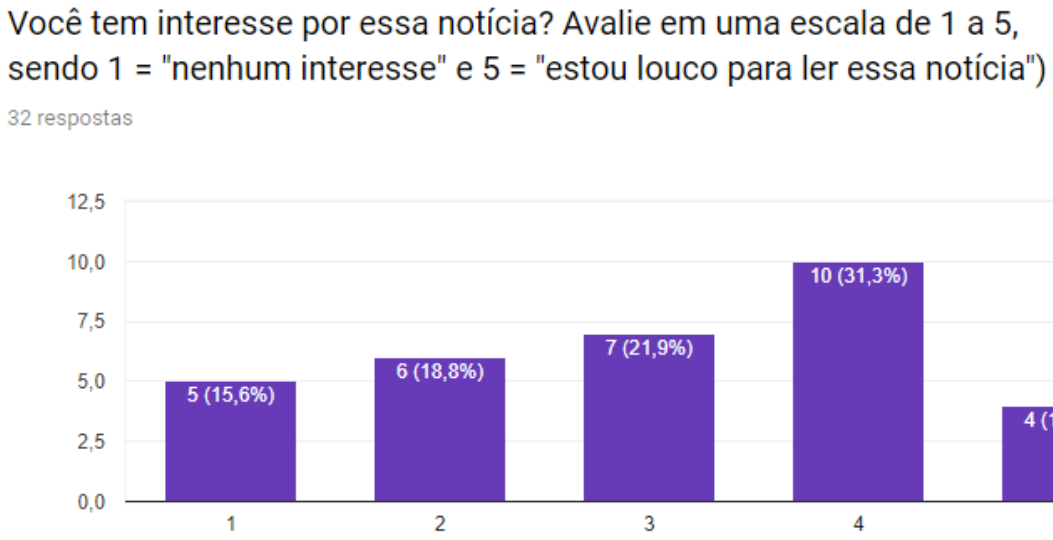


Figura 47: Média do interesse dos participantes da pesquisa pela 5ª notícia apresentada.

A opção “em casa” foi amplamente mais votada com cerca de 90% dos votos como um momento de grandes possibilidades para o consumo da notícia. As opções “no trabalho” (com 41,4%) ou “em uma viagem” (com 24,1%) ficaram bem abaixo na preferência dos participantes, conforme figura 48.

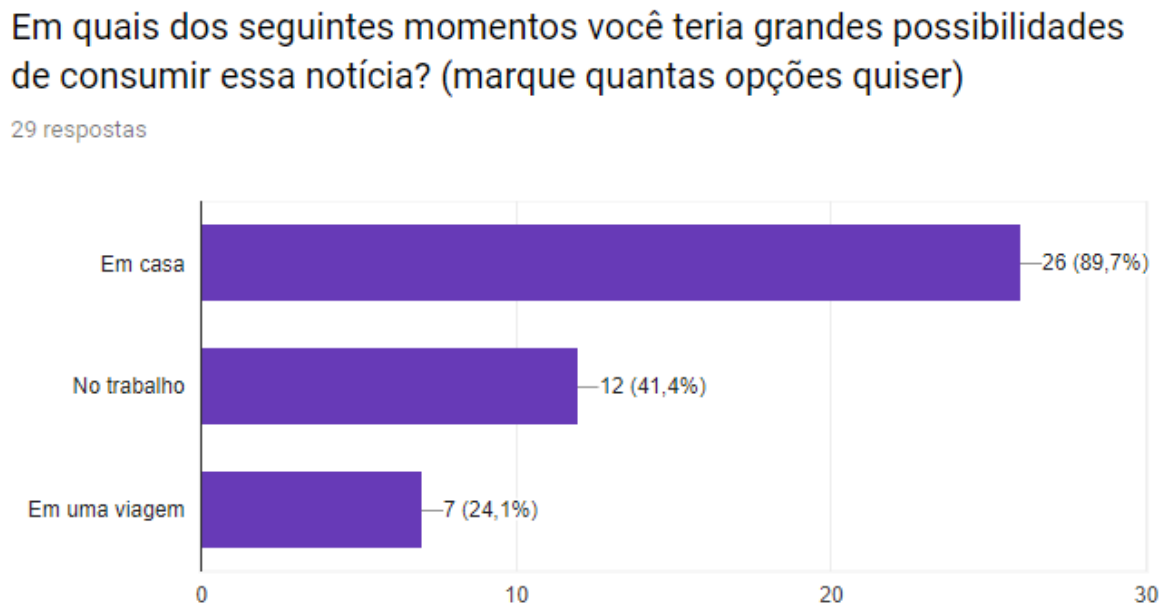


Figura 48: Percentual dos momentos escolhidos como de grande possibilidade para o consumo da notícia.

Assim, conforme os critérios estabelecidos, a diferença entre o percentual de escolha dos 3 momentos apresentados deixou evidente que o contexto exerceu grande influência nas possibilidades de um usuário consumir essa notícia. A

exemplo das demais notícias utilizadas nessa pesquisa, a recomendação deste conteúdo teria uma probabilidade maior de ser assertiva sempre que os usuários estivessem no contexto “em casa”.

Dessa maneira, de acordo com o que foi proposto para a análise dos resultados desse questionário no capítulo de procedimentos metodológicos, a avaliação do percentual dos 3 cenários contextuais propostos (“em casa”, “no trabalho” e “em uma viagem”) foi conclusiva. Em nenhuma das notícias houve uma equiparação ou uma similaridade entre os resultados consolidados, independente do interesse manifestado pelos participantes por aquele determinado assunto. Por conta disso, pode-se afirmar que para 100% das notícias a mudança de contexto impacta nas possibilidades de que ela seja consumida naquele momento.

Entretanto, observou-se também indícios para que se estabeleça uma relação entre o interesse manifestado pela notícia e um aumento nas possibilidades de consumo da mesma. Quanto menor o interesse geral pela notícia, menores as chances de que ela venha a ser consumida em qualquer contexto. Quanto maior o interesse, maior a chance de que a pessoa consuma a notícia em qualquer contexto.

Além disso, foi possível concluir que independente da editoria ou assunto da notícia, os resultados indicam que as pessoas em geral estariam mais propensas a consumir notícias “em casa”, seguido por “no trabalho” em segundo lugar e “em uma viagem” em terceiro lugar. Em 100% dos casos esse padrão se repetiu, independente do interesse maior ou menor pela notícia.

### 4.2.3 Análise individual das respostas

De maneira a complementar a análise dos resultados obtidos na pesquisa, cada uma das respostas dos participantes foi avaliada individualmente. Como primeiro objetivo desta avaliação, procurou-se identificar o percentual de participantes que poderiam ser considerados “imunes” a influência do contexto, conforme os procedimentos metodológicos propostos no capítulo 2.

Entre os 32 respondentes, apenas 2 deles demonstraram não sofrer nenhuma influência do contexto em seus hábitos de consumo de notícias, ao assinalarem que consumiriam as notícias de seu interesse em todos os contextos propostos, e que

não leriam uma notícia de baixo interesse em nenhum dos contextos propostos. Todos os demais participantes apresentaram oscilações nas possibilidades de consumir ao menos uma notícia em um contexto específico, e, portanto, foram considerados impactados pelas variações contextuais. Sendo assim, pode-se afirmar que para cerca de 94% dos participantes, independente do interesse previamente manifestado, o contexto é fator de grande impacto na possibilidade de consumir ou não uma determinada notícia.

Na sequência, como segundo objetivo da avaliação, efetuou-se um recorte apenas das notícias avaliadas com o grau “5” na escala de interesse (ou seja, as notícias mais interessantes para os participantes), para avaliar o percentual de participantes que, apenas para essas notícias, teriam afirmado que tenderiam a consumí-las em qualquer contexto. Ainda assim, somente 6 dos 32 participantes assinalaram que consumiriam uma notícia de alto nível de interesse em qualquer um dos três contextos oferecidos na pesquisa. Ou seja, para todos os demais (26 pessoas ou cerca de 81% dos participantes), mesmo uma notícia de máximo grau de interesse não seria consumida se fosse recomendada em um momento “errado” (conforme exemplo da figura 49).

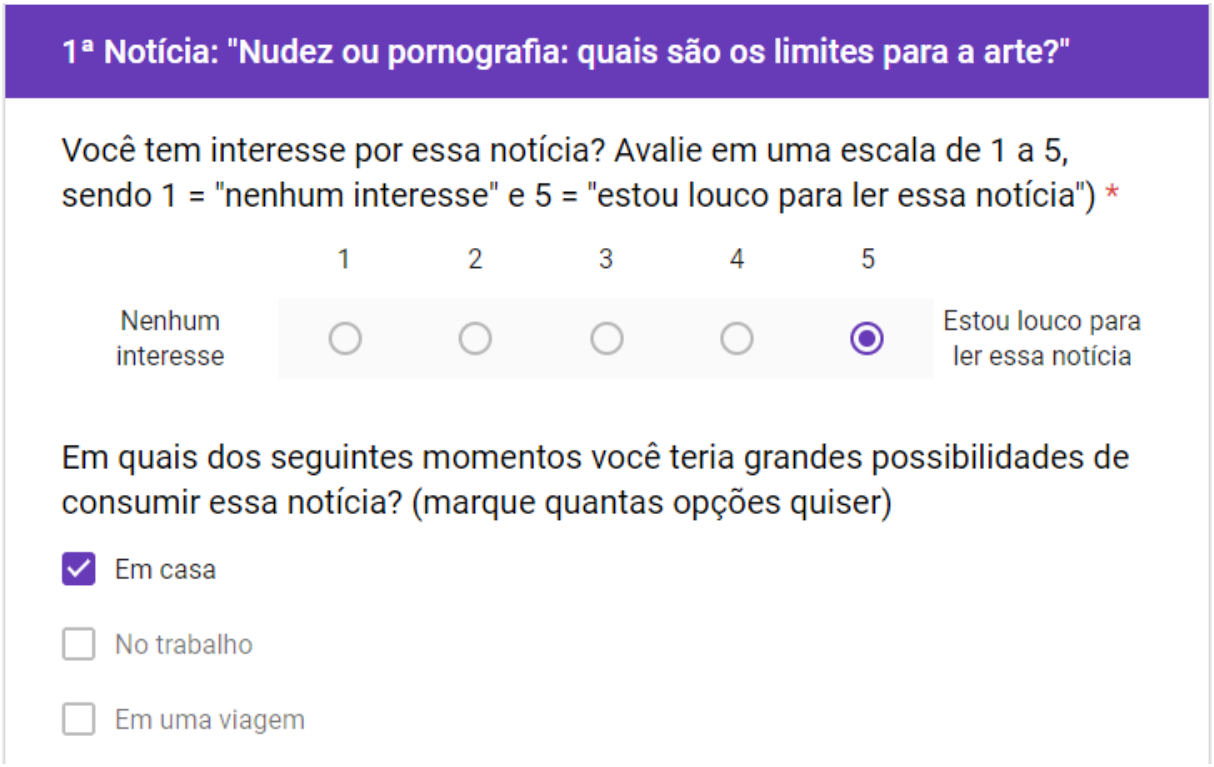


Figura 49: Exemplo de participante da pesquisa indicando que, mesmo com interesse máximo em uma notícia, não iria consumí-la em qualquer contexto.

Fazendo uma referência ao capítulo de fundamentação teórica, tal estatística identificada vai ao encontro do que cita Stray (2010), de que mais importante do que produzir bom conteúdo é oferecê-lo para o usuário na hora certa.

Os resultados dessa pesquisa ajudam a fornecer mais indícios da relevância de se levar em consideração o contexto no momento da recomendação de conteúdos em um site de notícias. Uma recomendação mais “contextual” seria sem dúvidas um fator preponderante para aumentar a assertividade desse modelo, estimulando um maior engajamento dos usuários com o produto de notícias.

Entretanto, conforme apresentado no capítulo 3, o levantamento feito com os principais sites de notícias brasileiros deixou claro que a questão do contexto ainda vem sendo pouco explorada pelos sistemas de recomendação de conteúdo de sites de notícias brasileiros.

# 5 **PROPOSIÇÃO TEÓRICA DE CENÁRIOS PARA A RECOMENDAÇÃO CONTEXTUAL DE CONTEÚDO NOTICIOSO**

A partir do levantamento a respeito do estado atual da recomendação de conteúdo em sites de notícias brasileiros proposto no capítulo 3, observa-se que é praticamente inexistente a presença de fatores que levem em conta o contexto de cada usuário no momento do acesso ao conteúdo notícioso (o que ele está fazendo, como está se sentindo, etc). Com exceção dos critérios de “geolocalização” (que podem considerar a localização atual do usuário para recomendar conteúdos daquela região) e “pessoal” (onde um nível simples de personalização baseada na interação ativa do usuário vem sendo aos poucos implementado), os demais critérios para a recomendação funcionam a partir da compreensão de perfis de usuários com interesses “estáticos”. Ou seja, a recomendação atual subentende que tais usuários não passariam por mudanças em seus interesses ao longo do tempo, ou não sofreriam influência de fatores contextuais em seus hábitos de consumo (levando-se em conta o desprezo aos fatores contextuais na recomendação).

Conforme conceitos apresentados no capítulo de referencial teórico, compreende-se que um dos grandes problemas dos sites de notícias não está

apenas no excesso de conteúdo, e sim na filtragem deste conteúdo. Em geral, a filtragem algorítmica nesses sites é muito baseada apenas na relação do usuário (ou de toda a comunidade de usuários) com o site em questão, como se essa relação por si só definisse com exatidão quem é aquele usuário ou o que ele quer ler/consumir a qualquer momento do dia.

A validação das hipóteses propostas nesta tese, em especial a que trata da influência exercida pelo contexto de cada usuário (pessoal, local ou mundial) na probabilidade do mesmo consumir uma determinada notícia (independente do seu nível de interesse) demonstrou que a inclusão de fatores contextuais é um elemento chave na obtenção de melhores resultados na recomendação de conteúdo jornalístico. Desconsiderar a influência do contexto (ou considerá-lo apenas de forma incipiente) é um dos fatores determinantes para os baixos resultados apresentados no uso de sistemas de recomendação automatizados em sites de notícias.

Cada vez mais existe a compreensão de que não basta estabelecer um perfil estático e definitivo de um usuário no momento da recomendação de conteúdo (e aí estende-se qualquer conteúdo, não apenas o jornalístico). Conforme citado no capítulo de referencial teórico desta tese, Stray (2010) destaca que oferecer o conteúdo correto no momento certo é mais importante até do que produzir bom conteúdo. Ou seja, se o objetivo é ampliar o engajamento e a assertividade da recomendação de conteúdo, não adianta simplesmente saber o que o usuário normalmente lê (ou compra). É preciso inferir o que esse usuário que consumir agora (e o “agora” envolve não apenas a questão temporal, mas todos os fatores contextuais envolvidos).

Em outros domínios e áreas de atuação que não o jornalismo, algumas empresas vêm sendo pioneiras na utilização de informações contextuais para recomendar conteúdo de maneira mais assertiva.

Desde 2012, a empresa “Google” vem experimentando e aprimorando uma ferramenta chamada “Google Now” (https://www.google.com/intl/pt-BR/landing/now). Com seu slogan “A informação certa na hora certa”, o Google Now se apresenta como um assistente pessoal “inteligente”, que a partir das mais diversas informações obtidas de cada usuário, é capaz de responder perguntas, fazer recomendações e agir antes mesmo do usuário se manifestar ou interagir ativamente à procura de algo. O mecanismo de inteligência artificial do Google Now é capaz de prever com grande assertividade o que um determinado usuário precisa naquele exato

momento, “antes mesmo de perguntar” (é dessa maneira que a ferramenta se descreve no site oficial).

Para exemplificar como a ferramenta funciona, a partir das informações de uma passagem aérea adquirida pelo usuário (cujas informações são enviadas para o email Google no ato da compra), o Google Now consegue recomendar conteúdos a respeito da cidade/país de destino, informações sobre o câmbio, o trânsito no dia da viagem, sugestões do melhor caminho a seguir, avisar de um possível atraso do vôo, etc. O mesmo tipo de recomendação se aplica a diversos outros aspectos, como: informações de trânsito e transporte público (avisos sobre o tempo de espera até o próximo ônibus, ou qual o melhor trem a pegar); clima; próximos compromissos; notícias; reservas em restaurantes; atrações próximas da localização do usuário; etc. Em resumo, toda informação coletada pela plataforma Google sobre um determinado usuário (e-mails, buscas, navegação, gps do dispositivo móvel, etc) é utilizada para antecipar comportamentos daquele usuário para recomendar o conteúdo certo no contexto exato.

A IBM, outra grande empresa da área de tecnologia, lançou em 2011 o projeto chamado IBM Watson (https://www.ibm.com/watson/about/), um sistema para o processamento avançado, recuperação de informação, representação de conhecimento, raciocínio automatizado e tecnologias de aprendizado de máquinas. Dentre uma infinidade de possibilidades, a plataforma vem sendo bastante utilizada em áreas como saúde, educação e transporte. Entretanto, em seu próprio site oficial, a IBM destaca a recomendação como um dos recursos oferecidos pela plataforma, destacando a importância do fator contextual: “Com o Watson, você pode fazer recomendações mais seguras e direcionadas, desenhando soluções a partir de um amplo conjunto de informações e entendendo nuances do contexto atual”.

Entre os sites de comércio eletrônico (e-commerce), a Amazon (www.amazon.com) vem aprimorando cada vez mais seus algoritmos de recomendação a partir de informações contextuais de seus usuários. Em 2017, os resultados de testes que levaram em conta fatores contextuais simples como “a hora do dia” ou o “tipo de dispositivo do usuário” apresentaram melhora de 21% na taxa de conversão (onde o usuário efetivamente exerce uma compra) após uma semana de funcionamento do novo algoritmo (https://blog.acolyer.org/2017/09/27/an-efficient-bandit-algorithm-for-real-time-multivariate-optimization).

Dessa maneira, em função dos bons resultados que vem sendo apresentados em outros domínios quando informações contextuais são utilizadas entre os fatores de recomendação de conteúdos, faz-se necessário avançar nos estudos acerca da implementação desses mesmos fatores nos sistemas de recomendação de sites de notícias.

A partir dos resultados apresentados nas pesquisas ao longo desta tese, em especial a pesquisa com usuários detalhada no item 3.2, ficou evidenciado que os aspectos contextuais propostos interferem no comportamento dos usuários no que diz respeito ao consumo de notícias. Na pesquisa, para cada uma das notícias apresentadas individualmente, seguiram-se duas perguntas: 1) “Você tem interesse por essa notícia? Avalie em uma escala de 1 a 5, sendo 1 = "nenhum interesse" e 5 = "estou louco para ler essa notícia"”; e 2) “Em quais dos seguintes momentos você teria grandes possibilidades de consumir essa notícia? (marque quantas opções quiser), cujas opções de resposta oferecidas foram: a) “em casa”; b) “no trabalho”; e c) “em uma viagem”.”

Na análise dos resultados, ficou demonstrado que mesmo notícias avaliadas como de “grande interesse” por parte de um usuário em específico poderiam não ser consumidas em um determinado contexto, ao passo que uma notícia de menor interesse poderia ter mais possibilidades de ser consumida nesse mesmo contexto. A partir do cruzamento dessas informações, é possível inferir que em diferentes contextos, diferentes conjuntos de notícias resultariam em maior ou menor engajamento dos usuários.

A seguir, ainda tomando como base os resultados dessa pesquisa e a consequente validação da hipótese formulada, é apresentada uma proposição de cenários teóricos para a recomendação de conteúdo aplicada a alguns dos participantes da pesquisa. Para tanto, são selecionados 3 participantes de diferentes perfis/áreas de atuação. Os cenários são propostos a partir do cruzamento das informações de nível de interesse e de contextos com maior possibilidade para o consumo de cada notícia.

**- Cenário 1: usuário “blogueira”**

O resultado da pesquisa aplicada ao 1º participante selecionado foi consolidado na tabela 7. Nela, é possível identificar o nível de interesse do usuário

em cada uma das 5 notícias propostas, bem como os contextos em que ele estaria mais propenso a consumir cada uma delas.

**Tabela 7 – Respostas do usuário 1 (blogueira) ao questionário.**

| Usuário 1 (blogueira) | Notícia 1 | Notícia 2 | Notícia 3 | Notícia 4 | Notícia 5 |
|---|---|---|---|---|---|
| Interesse pela Notícia | 5 | 3 | 4 | 5 | 4 |
| Momentos em que consumiria a notícia | Em casa | Em casa, No trabalho | Em casa, No trabalho, Em uma viagem | Em casa, No trabalho, Em uma viagem | Em casa No trabalho |

Fonte: Pesquisa realizada pelo autor (google forms).

A partir do cruzamento dos dados obtidos, é possível assumir que para cada contexto (ou mesmo desconsiderando o contexto e levando em conta somente o interesse do usuário), a oferta de diferentes conjuntos de notícias resultaria em diferenças na assertividade do sistema de recomendação.

Quando é considerado apenas o “interesse” assinalado, o melhor conjunto de notícias para a recomendação seria, pela ordem: Notícia 1 (5), Notícia 4 (5), Notícia 3 (4) e Notícia 5 (4) e Notícia 2 (3).

Quando a segmentação se dá apenas pelas notícias assinaladas como de grande possibilidade de consumo “Em casa”, o conjunto de notícias e a ordem da exibição permaneceria os mesmos, já que as 5 notícias foram assinaladas nesse contexto.

Entretanto, segmentando pelo contexto “No trabalho”, o melhor conjunto de notícias para a recomendação seria: Notícia 4 (5), Notícia 3 (4), Notícia 5 (4) e Notícia 2 (3). Neste caso, a notícia 1, apesar do interesse máximo assinalado (nível 5) seria descartada porque não seria consumida no contexto “No Trabalho”.

Da mesma forma, a segmentação pelo contexto “Em uma viagem” resultaria em um conjunto de notícias diferente de todos os demais: apenas a Notícia 4 (5) e Notícia 3 (4) seriam recomendadas. Todas as demais notícias seriam descartadas por não terem sido assinaladas para consumo neste contexto.

**- Cenário 2: usuário “designer”**

Na tabela 8 é apresentado o resultado consolidado das respostas do segundo participante selecionado. Da mesma forma que na tabela 7, é possível identificar o nível de interesse do usuário em cada uma das 5 notícias propostas, bem como os contextos em que ele estaria mais propenso a consumir cada uma delas.

**Tabela 8 – Respostas do usuário 2 (designer) ao questionário.**

| Usuário 1 (designer) | Notícia 1 | Notícia 2 | Notícia 3 | Notícia 4 | Notícia 5 |
|---|---|---|---|---|---|
| Interesse pela Notícia | 3 | 2 | 4 | 3 | 2 |
| Momentos em que consumiria a notícia | Em casa | No trabalho | Em casa, No trabalho | Em uma viagem | Em casa |

Fonte: Pesquisa realizada pelo autor (google forms).

O cruzamento das respostas novamente permite observar que para contextos diferentes a oferta de diferentes conjuntos de notícias poderia acarretar em melhores níveis de assertividade e engajamento a partir dos sitemas de recomendação.

Apenas pelo “interesse” demonstrado, o melhor conjunto de notícias para a recomendação seria, pela ordem: Notícia 3 (4), Notícia 1 (3), Notícia 4 (3), Notícia 2 (2) e Notícia 5 (2).

Quando segmentado o contexto “Em casa”, a recomendação seria: Notícia 3 (4), Notícia 1 (3) e Notícia 5 (2). A Notícia 4 (mesmo com maior interesse do que a Notícia 5) e a Notícia 2 seriam descartadas pois o usuário não consumiria estes conteúdos no contexto “em casa”.

Já no contexto “No trabalho”, o conjunto de notícias recomendadas seria: Notícia 3 (4) e Notícia 2 (2). Neste cenário, uma das notícias de interesse mais baixo (notícia 2) seria recomendada, ao passo que notícias de maior interesse seriam descartadas em função do contexto “no trabalho” não ter sido assinalado.

Por último, no contexto “em uma viagem”, apenas a Notícia 4 (3) deveria ser recomendada. Todas as demais notícias seriam descartadas do sistema de recomendação, independente do nível de interesse maior ou menor, uma vez que o usuário assinalou que não consumiria estes conteúdos durante uma viagem.

**- Cenário 3: usuário “pedagoga”**

Os resultados consolidados do terceiro e último participante selecionado são apresentados de forma consolidada na tabela 9, seguindo os mesmos padrões utilizados nas duas tabelas anteriores.

**Tabela 9 – Respostas do usuário 3 (pedagoga) ao questionário.**

| Usuário 3 (pedagoga) | Notícia 1 | Notícia 2 | Notícia 3 | Notícia 4 | Notícia 5 |
|---|---|---|---|---|---|
| Interesse pela Notícia | 4 | 5 | 1 | 4 | 3 |
| Momentos em que consumiria a notícia | Em casa, Em uma viagem | Em casa, No trabalho, Em uma viagem | No trabalho | Em casa, No trabalho, Em uma viagem | No trabalho |

Fonte: Pesquisa realizada pelo autor (google forms).

Ao considerar unicamente o “interesse” assinalado, desconsiderando os diferentes contextos, o melhor conjunto de notícias para o sistema de recomendação seria, pela ordem: Notícia 2 (5), Notícia 1 (4), Notícia 4 (4), Notícia 5 (3) e Notícia 3 (1).

Entretanto, para este mesmo usuário no contexto “em casa”, a melhor recomendação seria: Notícia 2 (5), Notícia 1 (4) e Notícia 4 (4). As notícias 3 e 5 seriam descartadas pois o usuário declarou que não consumiria estes conteúdos no contexto “em casa”.

Da mesma forma, no contexto “No trabalho”, um conjunto diferente de notícias seria recomendado, a saber: Notícia 2 (5), Notícia 4 (4), Notícia 5 (3) e Notícia 2 (1).

A notícia 1, apesar do grande interesse (4) seria descartada em função do contexto, em detrimento da Notícia 3, de baixíssimo interesse (1) mas que poderia ser consumida no contexto “no trabalho”.

Já “Em uma viagem”, o conjunto de notícias a serem recomendadas seria: Notícia 2 (5), Notícia 1 (4) e Notícia 4 (4). As demais notícias (3 e 5) seriam descartadas em função do contexto.

Ao analisar todos os cenários teóricos propostos e apresentados acima, primeiramente é possível identificar as diferenças entre os hábitos de consumo dos 3 usuários selecionados. Dessa maneira, fica novamente evidenciado que um mesmo elemento de recomendação de conteúdo, regido pelos mesmos critérios para a oferta, não apresentaria resultados satisfatórios nas métricas de engajamento para todos os 3 usuários ao mesmo tempo.

Além disso, é possível identificar que em apenas um dos cenários teóricos propostos (usuário 1 no contexto “em casa”) não houve diferenças entre o melhor conjunto de notícias a serem recomendadas em relação ao interesse demonstrado pelo usuário (no caso, “por interesse” e “em casa” apresentaram os mesmos resultados). Em todos os demais cenários, seja na comparação com o nível de interesse nas notícias, seja na comparação entre diferentes contextos, observa-se que o conjunto de notícias selecionados para a recomendação mais assertiva apresentou variações entre si.

Dessa maneira, a proposição de cenários reafirma que aspectos contextuais interferem no consumo de notícias. Mesmo notícias de maior interesse poderiam não ser consumidas em um determinado contexto, em detrimento de notícias até de menor interesse, mas que ainda assim teriam maior probabilidade de serem acessadas naquele mesmo contexto.

Nesse caso, a proposição mostra que o simples conhecimento de uma variável contextual (se o usuário está acessando de casa, do trabalho ou durante uma viagem) tornaria possível segmentar a recomendação de conteúdo de diferentes maneiras, aumentando a possibilidade de ser assertivo e engajar os usuários. Evidentemente, é possível deduzir por associação que o conhecimento a respeito de outras variáveis contextuais permitiria aos sites de notícias implementar sistemas de recomendação com resultados ainda mais expressivos em relação às métricas de engajamento.

Do ponto de vista prático, apenas para indicar algumas possibilidades baseadas em exemplos de sistemas de recomendação de outras áreas apresentados neste capítulo, as seguintes soluções tornariam viável a aferição de informações contextuais por parte dos sites de notícias:

1) Perguntar ativamente para o usuário algo sobre seu contexto durante o acesso à primeira notícia (ao invés de exigir um cadastro com atividade logada, por exemplo, como é feito em muitos sites atualmente);
2) Integrar o sistema de cadastro do site de notícias com o de redes sociais (como facebook, google, linkedin, etc) onde haveriam informações suficientes sobre o usuário para identificar aspectos contextuais (e de fato utilizar-se dessas informações).
3) Considerar a média da população geral para recomendar conteúdos específicos em cada contexto (ex: em média, o trabalho acontece no horário comercial; fora disso, a pessoa estaria em casa; Viagens seriam em períodos de férias ou em datas próximas a feriados). Assim, a recomendação para o usuário seria diferente dependendo do dia e horário em que ele acessasse.

Tais soluções, apesar de simples, são exemplos de recursos que ainda não são utilizados pelos sites de notícias brasileiros avaliados nesse estudo. A aplicação dessas soluções, bem como a maior inclusão de aspectos contextuais nos modelos de sistemas de recomendação automatizados em sites de notícia permanecem como um vasto campo a ser pesquisado não apenas pela área do Design, como também por áreas ligadas às Ciências da Computação, *Machine Learning*, Inteligência Artificial, Estatística e Comunicação, dentre outras.

## CONSIDERAÇÕES FINAIS

A medida que o crescente aumento na produção e publicação de conteúdo online apresenta-se de maneira inquestionável nos mais diversos domínios e áreas de conhecimento, os sistemas de recomendação automatizados vêm ganhando espaço como solução para a filtragem e consequente sugestão de conteúdos de

forma cada vez mais personalizada.

Entretanto, se tais sistemas apresentam resultados impressionantes no que diz respeito aos níveis de assertividade e estímulo ao engajamento de usuários em áreas como recomendação de filmes, livros ou artigos para compra, no âmbito das notícias, os resultados até o momento não têm sido satisfatórios. O baixo engajamento decorrente em parte do insucesso na utilização destes sistemas de recomendação vem provocando quedas nas métricas e mudanças nos modelos de negócio dos sites de notícias a nível mundial.

Partindo deste problema de pesquisa, o estudo apresentado nesta tese apresentou um levantamento teórico e prático acerca dos sistemas de recomendação de conteúdo e de sua utilização nos sites de notícias, em especial os sites brasileiros, que foram o foco de toda a pesquisa.

Em um primeiro momento, foram propostas três hipóteses na tentativa de identificar alguns dos fatores responsáveis para o problema de pesquisa apresentado. Para tanto, foram utilizadas como abordagens metodológicas a análise de métricas de utilização do site do jornal O Globo e também pesquisas com usuários de sites de notícias.

Nesse ponto, por conta das dificuldades técnicas e da limitação de recursos, não foi possivel fazer um estudo mais abrangente, contemplando outras ferramentas de análise de sites de notícias. Por essa razão, as hipóteses foram validadas de forma experimental, o que não esgota a questão acerca do tema, mas conduz a proposição de caminhos e de ferramentas para a validação completa, bem como para o avanço no campo da recomendação de conteúdo em notícias.

A seguir, a partir do levantamento prático acerca do estado atual da recomendação de conteúdo nos sites de notícias do Brasil, foi possível chegar a uma proposta de categorização para a recomendação nestes sites.

Entendendo que são vários os fatores que poderiam impactar nos resultados insatisfatórios apresentados pelos sistemas automatizados de recomendação de conteúdo noticioso, o enfoque desta tese se deu sobre os critérios adotados para a escolha do conteúdo a ser recomendado - ou seja, para o melhor entendimento acerca do funcionamento dos algoritmos por trás dos elementos responsáveis pela recomendação de conteúdo. Apesar de entender que outros fatores como o design visual destes elementos, bem como sua usabilidade, podem igualmente ter influência nos baixos resultados apresentados, este estudo não aprofunda a

compreensão a respeito de tais fatores.

Além disso, utilizou-se uma abordagem metodológica do Design acerca dos elementos e sistemas de recomendação na tentativa de identificar soluções dos problemas típicos e muitas vezes exclusivos ou acentuados no âmbito das notícias.

A partir dos resultados obtidos na análise de métricas e pesquisas com usuários, foi possível identificar e validar experimentalmente que a inclusão de questões contextuais de cada usuário pode ser vista como um caminho para a melhoria na recomendação de conteúdo no âmbito da notícia. Utilizando os dados obtidos na pesquisa com usuários, foi possível alcançar o objetivo de propôr cenários teóricos que levam em conta o contexto de cada usuário para recomendar conteúdos específicos, com maior probabilidade de consumo. A aplicação dos fatores contextuais na recomendação de conteúdo de notícias resultaria em diferentes critérios para a seleção e priorização dos conteúdos a serem recomendados em cada acesso de cada usuário diferente, variando em função dos diferentes contextos.

Apesar da relevância do tema, os sistemas de recomendação dos sites de notícias brasileiros em sua maioria não se utilizam de fatores contextuais para a recomendação de conteúdo, entendendo o usuário como um ser de interesses muitas vezes estáticos e pouco influenciados pelo contexto. Entretanto, o que se verifica é que os hábitos de consumo de conteúdo noticioso são amplamente impactados por variações contextuais, e paratanto, a inclusão desses fatores nos algoritmos de recomendação de conteúdo se apresentam como um fator chave para a melhoria dos resultados apresentados por esses sistemas. Especificamente no âmbito das notícias, o comportamento e os interesses dos usuários tendem a ser mais afetados pelo contexto do que em outras áreas onde os sistemas de recomendação têm obtido melhores resultados em sua utilização.

Em relação a dificuldades e complexidades deste estudo em específico, um primeiro e importante fator a ser relatado é a baixa disponibilidade de métricas e dados aprofundados dos sites de notícias. Por serem sites pertencentes a empresas privadas e concorrentes entre si, a disponibilidade de informações técnicas sobre os sistemas de recomendação, suas regras e critérios para a oferta de conteúdo não são disponibilizados de forma aberta, sob o risco de serem utilizados pelas demais empresas.

Da mesma forma, outra grande dificuldade no entendimento dos fatores

responsáveis pelos resultados insatisfatórios no uso de sistemas de recomendação em sites de notícia é complexidade em isolar tais fatores para análise. As métricas podem ser mensuradas, mas não explicam o por quê de tais resultados. Para uma avaliação mais precisa a respeito do quanto cada fator (design visual, usabilidade, critérios para a recomendação, etc) tem influência nos resultados apresentados, seria necessário ter total liberdade para fazer alterações pontuais em um site de notícias e analisar o resultado de cada uma dessas alterações individualmente.

Como resultado direto, um terceiro fator de grande complexidade neste estudo é a implementação prática das soluções em um cenário real e sua consequente validação prática. Para isso, em determinado momento da pesquisa, pensou-se na possibilidade de prototipar uma solução que simulasse a recomendação de conteúdo em sites de notícias.

Entretanto, tal solução esbarra no quarto e último fator de limitação para as pesquisas sobre o tema, que é a complexidade (do ponto de vista técnico) de se desenvolver e implementar sistemas de recomendação automatizados, principalmente com o nível de profundidade proposto – com a inclusão de fatores contextuais. Desta maneira, a proposta de cenários teóricos para a recomendação, a partir de dados reais obtidos em pesquisas com usuários, se mostrou o caminho mais acertado para demonstrar que contextos afetam o comportamento dos usuários no âmbito da notícia. Tais contextos demandariam, portanto, diferentes critérios para a recomendação de notícias de maneira mais assertiva.

É importante ressaltar que por ser uma área de pesquisa muito nova, inúmeros estudos acerca do tema vêm sendo feitos e não é possível ainda apontar um caminho único e definitivo para a melhoria dos sitemas de recomendação. A medida em que os recursos tecnológicos evoluem com grande velocidade, é possível que sejam encontradas formas cada vez mais inovadoras de automatizar a recomendação de conteúdos a partir de critérios mais individualizados e específicos, oferecendo experiências de uso cada vez mais otimizadas e aprofundadas.

A respeito da abordagem contextual como critério para a recomendação automatizada de conteúdo, o estudo deixa uma série de possibilidades e caminhos para novos trabalhos e pesquisas acerca do tema: quais aspectos contextuais promovem maior interferência no padrão de consumo de notícias? Qual a melhor solução do ponto de vista tecnológico para obter o máximo de informações contextuais sem ultrapassar os limites da invasão de privacidade? Além da inclusão

dos aspectos contextuais na lógica algorítmica dos sistemas de recomendação, que outras alterações poderiam ser feitas nos elementos de recomendação de conteúdo para elevar sua assertividade (como mudanças de design ou na usabilidade desses elementos)? Todas essas questões se apresentam como pertinentes para futuros estudos, trazendo maior compreensão acerca do tema e possibilitando a obtenção de melhores resultados na implementação de sistemas de recomendação de conteúdo em sites de notícias.

**REFERÊNCIAS**

ADOMAVICIUS, G. TUZHILIN, A. *Toward the Next Generation of Recommender Systems*: A Survey of the State-of-the-Art and Possible Extensions. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, junho 2005.

ASSOCIATED PRESS. *A New Model for News - Studying the Deep Structure of Young-Adult News Consumption:* A Research Report from The Associated Press and the Context-Based Research Group. Jul. 2008 Disponível em: <http://rumble.me/wp-content/uploads/2013/02/A-New-model-for-news.pdf> Acesso em: 04 jan. 2015

BORTHWICK, J. *The rise of social distribution networks*. Mai. 2009. Disponível em: <http://www.businessinsider.com/henry-blodget-the-rise-of-social-distribution-networks-2009-5>.
Acesso em: 07 jan. 2015.

CARNEIRO, R. *Qual a maturidade de sua empresa para a área de métricas*? Web analytics – uma visão brasileira. E-book. Abr. 2008.

DAVID, A. B. *Context design:* How to anticipate users' needs before they're needed .The Next Web, 28 abr. 2014. Disponível em: https://thenextweb.com/dd/2014/04/28/context-design-anticipate-users-needs-theyre-needed/
Acesso em: 30 dez 2017.

DIAKOPOULOS, N. *Algorithmic accountability reporting:* on the investigation of black boxes. Columbia Journalism School, Tow Center for Digital Journalism, Columbia. 2014.

GARCIN, F; ZHOU, K.; FALTINGS, B. *Personalized News Recommendation Based on Collaborative Filtering*. Artificial Intelligence Lab, Ecole Polytechnique Federale de Lausanne – Switzerland. 2016.

GOLDBERG, D.; NICHOLS, D.; OKI, B. M.; TERRY, D. *Using collaborative filtering to weave an information tapestry*. Communications of the ACM, 35(12):61–70, 1992.

GONÇALVES, S. Para não naufragar no oceano da informação. Universidade de São Paulo, Escola de Comunicação e Artes. 2009.

GUO, G. *Resolving Data Sparsity and Cold Start in Recommender Systems.* In: Masthoff J., Mobasher B., Desmarais M.C., Nkambou R. (eds) User Modeling, Adaptation, and Personalization. UMAP 2012. Lecture Notes in Computer Science, vol 7379. Springer, Berlin, Heidelberg, 2012.

HAILE, T. *What You Think You Know About the Web Is Wrong*. 2015. Disponível em: <http://time.com/12933/what-you-think-you-know-about-the-web-is-wrong>
Acesso em: 04 abr. 2015.

HENDRIX, P. *5 ways startups are narrowing the content discovery gap*. 2014. Disponível em:< https://gigaom.com/2014/03/15/5-ways-startups-are-narrowing-the-content-discovery-gap >. Acesso em: 04 abr. 2015.

HIGGINS, E. SCHOLER, A. *Engaging the consumer*: The science and art of the value creation process. Journal of Consumer Psychology, v. 19 (n. 2), pp. 100-114. 2009.

INGRAM, M. *Drinking from the Twitter firehose:* I love the stream, but I need more filters and bridges. Jan. 2014. Disponível em:
<https://gigaom.com/2014/01/09/drinking-from-the-twitter-firehose-i-love-the-stream-but-i-need-more-filters-and-bridges/>.
Acesso em: 30 dez. 2014.

INGRAM, M. *If the news is important, it will find me.* Mar. 2008. Disponível em:
<http://www.mathewingram.com/work/2008/03/27/if-the-news-is-important-it-will-find-me>.
Acesso em: 30 dez. 2014.

INGRAM, M. *The Future of Media: Storify and the Curatorial Instinct.* Abr 2011. Disponível em: <https://gigaom.com/2011/04/25/the-future-of-media-storify-and-the-curatorial-instinct/>.
Acesso em: 30 dez. 2014.

JOLLY, J. *How algorithms decide the news you see.* Mai. 2014. Disponível em: <http://www.cjr.org/news_literacy/algorithms_filter_bubble.php?page=all>.
Acesso em: 30 dez. 2014.

JONES, M. *Introduction to approaches and algorithms.* Dez. 2013. Disponível em:https://www.ibm.com/developerworks/library/os-recommender1/os-recommender1-pdf.pdf
Acesso em: 23 nov. 2017.

KENNEDY, J. *No Time Left for Reluctant Transformers.* Dez. 2008. Disponível em: <http://niemanreports.org/articles/no-time-left-for-reluctant-transformers/>.
Acesso em: 04 jan. 2015.

LAFRANCE, A. *Coming in the side door*: The value of homepages is shifting from traffic-driver to brand. Ago. 2012.
Disponível em: <http://www.niemanlab.org/2012/08/coming-in-the-side-door-the-value-of-homepages-is-shifting-from-traffic-driver-to-brand/>
Acesso em: 11 set. 2016.

LAKATOS, E. M. *Fundamentos de metodologia científica 1*. 5ª ed. Atlas. São Paulo. 2003.

LEÃO, L. *O Labirinto da Hipermídia*: Arquitetura e navegação no ciberespaço. Editora Iluminuras. São Paulo, 1999.

LOHMANN, A. *Fluxos da navegação contextual na interação com o conteúdo noticioso.* Anais do Interaction South America. 2013.

MARRA, G. *Engajamento do Consumidor*: revisão teórica do conceito e seus antecedentes. Revista REGE , São Paulo – SP, Brasil, v. 20, n. 2, p. 233-249, abr./jun. 2013

MELVILLE, P; SINDHWANI, V. *Encyclopedia of Machine Learning*. 2010.

MITCHEL, A.; PAGE, D. *Social, search & direct:* pathways to digital news*.* 2014. Disponível em:<http://www.journalism.org/files/2014/03/SocialSearchandDirect_PathwaystoDigitalNews.pdf>. Acesso em 14 abr. 2014

MITCHEL, A.; ROSENTIEL, T. *What Facebook and Twitter Mean for News.* 2012. Disponível em: <http://stateofthemedia.org/2012/mobile-devices-and-news-consumption-some-good-signs-for-journalism/what-facebook-and-twitter-mean-for-news> Acesso em 10 abr. 2014

MITCHELL, T. *Machine Learning*. Publisher: McGraw-Hill Science/Engineering/Math. ISBN: 0070428077. 1997.

NEWMAN, N. *Overview and Key Findings of the 2016 Report*. University of Oxford, 2016. Disponível em: http://www.digitalnewsreport.org/survey/2016/overview-key-findings-2016/#sharetab-0. Acesso em: 21 Mar. 2016.

OKADA, S. *Web Analytics*: Modelos de Métricas de Engajamento em Mídias Emergentes. Revista Brasileira de Marketing, 10, feb. 2012.

PARISER, E. *O filtro invisível*: o que a internet está escondendo de você. Zahar, 2012.

RAMPTON, J. *The Past, Present, and Future of Content Discovery*. Dez. 2014. Disponível em: <http://www.forbes.com/sites/johnrampton/2014/09/12/the-past-present-and-future-

of-content-discovery/>.
Acesso em: 02 jan. 2015.

RESNICK, P.; VARIAN, H. R. *Recommender Systems*, Communications of the ACM 40. 1997.

REZ, R. *Marketing de conteúdo*: a moeda do século XXI. DVS Editora, 2017.

RICCI, F. et al. (eds.). *Recommender Systems Handbook*, Springer Science Business Media, LCC 2011. DOI 10.1007/978-0-387-85820-3_1.

RICCI, F. et al. (eds.). *Recommender Systems*: The Textbook, Springer International Publishing Switzerland. 2016. DOI 10.1007/978-3-319-29659-3 1

SEWARD, Z. *The homepage is dead, and the social web has won—even at the New York Times.* Mai. 2014. Disponível em: <http://qz.com/209950/the-homepage-is-dead-and-the-social-web-has-won-even-at-the-new-york-times/>.
Acesso em: 30 dez. 2014.

SHALEV-SHWARTZ, S. BEN-DAVID, S. *Understanding Machine Learning*: From Theory to Algorithms. Cambridge University Press. 2014

SHARDANAND, U. MAES, P. *Social information filtering*: algorithms for automating “word of mouth”. In: Proceedings of the SIGCHI conference on Human factors in computing systems, p. 210–217. ACM Press/Addison-Wesley Publishing Co., 1995.

SMOLA, A. VISHWANATHAN, S. *Introduction to Machine Learning*. Yahoo! Labs Santa Clara –and– Departments of Statistics and Computer Science Purdue University –and– College of Engineering and Computer Science Australian National University. 2008.

STRAY, J. *Are we stuck in filter bubbles? Here are five potential paths out.* Jul. 2012. Disponível em:

<http://www.niemanlab.org/2012/07/are-we-stuck-in-filter-bubbles-here-are-five-potential-paths-out/>.
Acesso em: 04 jan. 2015.

STRAY, J. *Designing journalism to be used*. Set. 2010. Disponível em: <http://jonathanstray.com/designing-journalism-to-be-used>.
Acesso em: 03 jan. 2015.

STRAY, J. *There's no such thing as an objective filter:* Why designing algorithms that tell us the news is hard. Jun. 2012. Disponível em: <http://www.niemanlab.org/2012/06/theres-no-such-thing-as-an-objective-filter/>.
Acesso em: 04 jan. 2015.

STRAY, J. *Who should see what when? Three principles for personalized news.* Jul. 2012. Disponível em: <http://www.niemanlab.org/2012/07/who-should-see-what-when-three-principles-for-personalized-news/>.
Acesso em: 03 jan. 2015.

THOMPSON, M. *An Antidote for Web Overload:* With a hunger for explanatory guidance amid the raging storm of Web news flashes, a journalist stresses context to attract digital users**.** Set. 2009. Disponível em: <http://niemanreports.org/articles/an-antidote-for-web-overload/>.
Acesso em: 02 jan. 2015.

TURCHI, S. *Estratégias e métricas digitais*. Abr. 2011.