MODELOS ESTOCÁSTICOS PARA LEITORES DE JORNAIS ONLINE

BRÁULIO MIRANDA VELOSO

MODELOS ESTOCÁSTICOS PARA LEITORES DE JORNAIS ONLINE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais – Departamento de Ciência da Computação. como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Renato Martins Assunção.
Coorientador: Nivio Ziviani.

Belo Horizonte

Agosto de 2016

Veloso, Bráulio Miranda

V443m Modelos estocásticos para leitores de jornais online. / Bráulio Miranda Veloso. — Belo Horizonte, 2016.
xxiv, 122 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Renato Martins Assunção.
Coorientador: Nivio Ziviani.

1. Computação — Teses. 2. Sistemas de Recomendação — Teses. 3. Processos Estocásticos — Teses. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Modelos estocásticos para leitores de jornais online

BRÁULIO MIRANDA VELOSO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. RENATO MARTINS ASSUNÇÃO - Orientador
Departamento de Ciência da Computação - UFMG

PROF. NIVIO ZIVIANI - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. BERTHIER RIBEIRO DE ARAÚJO NETO
Departamento de Ciência da Computação - UFMG

PROF. EDLENO SILVA DE MOURA
Departamento de Ciência da Computação - UFAM

PROF. RODRYGO LUIS TEODORO SANTOS
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 08 de agosto de 2016.

À memória de
Carina Parma, Gustavo Santos e Nancileia Carvalho,
amigos do maravilhoso tempo do ensino básico em Ouro Preto.

Agradecimentos

Em primeiro lugar gostaria de agradecer a Deus, pelo dom da vida e aos meus queridos pais José Geraldo Veloso e Maria do Carmo Miranda Veloso pelos sacrifícios realizados para esta conquista. Agradeço também a minha irmã Silvânia, ao cunhado Rafael e ao meu irmão Wanderson, pelo carinho, incentivo e ajuda nas horas difíceis.

Também gostaria de agradecer a minha noiva e futura esposa Gabriella Leone Fernandes pelo amor, cuidado e compreensão das horas de convivência que tivemos de abdicar para a produção desta dissertação. Obrigado Gabi, com a benção de Deus continuaremos firmes e felizes.

Agradeço aos amigos e colegas que me ajudaram de alguma maneira na elaboração deste material. Em especial, agradeço aos professores Nivio Ziviani e Renato Assunção, pelas orientações, transmissão de conhecimento, incentivo e grande apoio. Gostaria de agradecer também a todos os demais colegas e professores do Laboratório para Tratamento da Informação – LATIN.

Agradeço aos caríssimos professores que aceitaram compor a banca examinadora. Obrigado pela boa vontade e por cederem um pouco do tempo de vocês.

Agradeço também aos familiares e amigos que facilitaram a vinda e a permanência em BH. Obrigado Zé Geraldo e tia Maria, Camila e tia São, e a todos das famílias Miranda e Veloso que torceram por mim.

Não menos importante, gostaria de agradecer a todos os colegas da estatística, em especial ao colegas do Laboratório de Estatística Espacial – LESTE, pelos bolos, cafés, sofá, e principalmente pelas ajudas nas dúvidas de probabilidades e `R`.

Finalmente agradeço ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais pelo apoio e incentivo.

"Mesmo que eu tivesse o dom da profecia, e conhecesse todos os mistérios e toda a ciência; mesmo que tivesse toda a fé, a ponto de transportar montanhas, se eu não tiver caridade, não sou nada."

(Coríntios 13,2)

Resumo

O objetivo desta dissertação é estudar o comportamento de usuários durante a leitura de artigos online em jornais digitais. O estudo foi baseado em mais de 20 milhões de sessões compostas pelos clicks sucessivos por parte de usuários em notícias postadas em dois jornais online de grande porte no país. A motivação para este trabalho é que o entendimento acerca do comportamento da leitura sucessiva de artigos pode ajudar no desenvolvimento de sistemas de recomendação mais eficazes. A sessão de cada usuário foi reduzida à sequência dos tópicos das notícias lidas. Foram estudados 32 modelos estocásticos, cada um deles procurando capturar a essência do comportamento do usuário. Eles foram divididos em cinco categorias: modelos sem influência do passado, os que desprezam totalmente a informação do passado; modelos de memória curta, em que apenas as leituras recentes afetam o futuro; modelos de preferência revelada, nos quais o futuro é condicionado nas características de um tópico por vez; modelos de permanência geométrica, onde o comportamento de leitura é dividido entre as opções de permanecer no tópico atual de leitura seguindo uma distribuição geométrica e mudar de tópico segundo algumas regras; e finalmente modelos de vantagem cumulativa, nos quais as leituras prévias de um tópico aumentam as chances de sua leitura no futuro. Os modelos foram ajustados por máxima verossimilhança e comparados com base no critério de informação de Akaike e no escore de acurácia de predição de Brier. Os melhores modelos são aqueles em que o usuário se move pelos tópicos influenciado apenas pelas suas leituras mais recentes. Os modelos de vantagem cumulativa estiveram logo atrás, com predições ligeiramente piores, mas ainda assim bastante satisfatórias. Uma conclusão importante é que o rastreamento dos clicks de um usuário numa dada sessão de leitura pode ser explorado para recomendar dinamicamente notícias online com maior efetividade.

Palavras-chave: Comportamento de *Clicks*; Recomendação de Notícias; Modelagem do Comportamento de Usuários; Modelos Estocásticos.

Abstract

The aim of this thesis is the study of the behavior of online users of digital newspapers. We analyzed more than 20 million sessions composed by user's successive clicks in news posted in two large Brazilian online newspapers. The motivation for this work is that understanding the sequence of topics reading behavior can help to design better recommendation systems. Each user session was reduced to the sequence of topics read. We analyzed 32 stochastic models, each one trying to capture the essence of the user behavior. They are divided into five categories: models without past influence, those that totally disregard the information of the past; short memory models, where only the recent topics read affect the next one; preference revealed models which the future is conditioned on characteristics of a topic at a time; geometric permanence models where the reading behavior is divided among the options of remaining on the current topic of reading following a geometric distribution and changing of topic according to some rules; and finally models of cumulative advantage, in which previous readings of a topic increase its readings chances in the future. The models are fitted by maximum likelihood and compared according to goodness of fit and prediction power. The best models are those in which the user moves around the states influenced by his most recent readings. The cumulative advantage models were close behind, with slightly worse predictions but still quite satisfactory. We show how our findings can be explored for dynamically recommending online news to a user based on his clicks tracking in a given reading session.

Keywords: Click Behavior; News Recommendation; User Behavior Modeling; Stochastic Models.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Na sua interação com sistemas computacionais, usuários desejam ter resultados personalizados, refletindo experiências de navegação próprias (Shardanand & Maes [1995]). Contudo, eles não estão dispostos a gastar muito tempo especificando suas necessidades com informações pessoais. Em certos casos, o usuário nem sabe o que está procurando, simplesmente espera que o sistema mostre conteúdo do seu interesse (Das et al. [2007]). Sistemas de Recomendação (Jannach et al. [2011]) visam identificar automaticamente informações relevantes para um determinado usuário, aprendendo com os dados disponíveis e recomendando o conteúdo mais interessante para esse usuário.

Empresas do mundo todo que possuem lojas online utilizam sistemas de recomendação, tanto para melhorar o relacionamento com os usuários, quanto para aumentar seu faturamento (Wei et al. [2007]). Quanto melhor for a recomendação de itens em um site, maior é a probabilidade de compras por um usuário. Empresas de outros ramos além de *e-commerce* também utilizam sistemas de recomendação. Um caso particular é o de jornais online (Resnick et al. [1994]) que recomendam artigos potencialmente relevantes para seus usuários, buscando assim mantê-los em seus domínios por mais tempo.

1.1 Motivação

Os Sistemas de Recomendação ganharam muita atenção nas duas últimas décadas devido ao seu grande apelo comercial. A busca por melhorias nos algoritmos de recomendação é constante, mesmo que cada uma dessas melhorias traga um ganho pequeno. Segundo Wei et al. [2007], o acúmulo de pequenos ganhos pode significar grandes aumentos nos lucros das empresas.

No contexto de jornais online, os sites colocam à disposição do usuário informações acerca de assuntos muito variados. Os usuários possuem perfis muito diversos, alguns tendo interesses muito focados em poucos assuntos, enquanto outros leem notícias de assuntos bem variados. Numa sessão de leituras, o assunto do primeiro artigo depende do que o usuário procura inicialmente. Normalmente, o usuário lê outros artigos em seguida, dentre os disponíveis no site, por recomendações ou links disponíveis no layout do site, gerando assim um histórico de leituras. Mais raramente, ele pode fazer buscas no site do jornal procurando alguma notícia específica. O acompanhamento desse histórico de leituras permite estudar o processo de escolhas sucessivas dos usuários.

A principal motivação desta dissertação é aprender os padrões da leitura sequencial de artigos numa dada sessão por um usuário e procurar utilizar esse conhecimento para alavancar as recomendações a serem feitas ao longo da sessão. Nosso modelo classifica as notícias em tópicos ou assuntos. Com essa redução de informação, vamos estudar o passeio aleatório da população de usuários de jornais online através dos tópicos durante uma sessão de leituras de artigos sucessivos. Várias questões que podem ser úteis para melhorar um sistema de recomendação foram exploradas, tais como a distribuição de probabilidade do número de artigos lidos dentro de um tópico, ou se essa distribuição depende do tópico lido no momento ou do número de artigos lidos anteriormente. Se observamos simplesmente os tópicos dos artigos lidos, já teremos informações úteis à recomendação. A descoberta de padrões probabilísticos de leitura, imperceptíveis de início, pode sugerir estratégias de recomendação mais relevantes para o usuário.

1.2 Trabalhos Relacionados

Em um sistema de recomendação, existem usuários (U) e itens (I). Os itens podem ser produtos, artigos de jornal ou serviços (como os de hotéis ou restaurantes) que estão disponíveis para os usuários. Os usuários (clientes ou utilizadores finais) consomem ou utilizam os itens. Assume-se que, potencialmente, os usuários poderiam avaliar todos os itens disponíveis. Essa avaliação dos usuários atribuída aos itens é comumente chamada de *rating* (Kantor et al. [2011]). Dado o grande número de itens, na prática, os usuários avaliam apenas uma fração muito pequena dos itens disponíveis. Além disso, a variedade de itens avaliados costuma ser pequena (Candillier et al. [2009]). A escala de variação dos ratings depende do sistema. Alguns exemplos são: {ótimo, médio, ruim}, {0, 1} ou $\{1\star, 2\star, 3\star, 4\star, 5\star\}$.

Usuários de um determinado sistema avaliam itens com relação aos seus gostos e/ou experiências. Por exemplo, usuários de um site de produtos eletrônicos podem avaliar em quantas estrelas (1 a 5) cada eletrônico (item) pode ser rotulado. O cliente avalia com mais estrelas os produtos mais satisfatórios.

Um sistema de recomendação sugere novos itens para usuários a partir da experiência de todos os usuários, dos conteúdos dos itens e características dos usuários. O princípio fundamental é descobrir os padrões e correlações entre as avaliações de forma a explicitar as preferências dos usuários e identificar quais os itens ainda não avaliados são potencialmente de seus interesses. Quanto melhor for a recomendação do sistema, mais alta é a possibilidade de haver nova compra. Logo, os sistemas de recomendação visam recomendar itens com boa chance de agradar ao usuário (Jahrer et al. [2010]).

Os sistemas de recomendação mais famosos em empresas são os da *Amazon*[1] e da *Netflix*[2]. A Amazon começou sua empresa com venda de livros online e devido ao crescimento das vendas impulsionado pelo seu recomendador hoje ela vende diversos itens desde cd's a roupas. Na Netflix os itens recomendados são filmes, séries e documentários. Os sistemas dessas empresas mostraram que uma boa recomendação aumenta as vendas (Linden et al. [2003]; Weigend [2003]; Meuth et al. [2008]; Gomez-Uribe & Hunt [2016]).

Abordagens de Predição

Existem três principais abordagens para a implementação de sistemas de recomendação: *filtragem colaborativa*, *filtragem baseada em conteúdo* e *filtragem híbrida* (Adomavicius & Tuzhilin [2005]; Candillier et al. [2007]).

Em um sistema de *filtragem colaborativa* a entrada consiste de um conjunto de ratings definidos pelos usuários a itens aos quais eles tiveram acesso. A estratégia utilizada para fazer recomendações é baseada na predição de ratings que um usuário forneceria para itens desconhecidos, utilizando um conjunto de ratings que outros usuários forneceram a esses itens. Os usuários que rotularam um determinado conjunto de itens de forma similar, tendem a continuar rotulando os demais itens da mesma forma. Logo, se um determinado item desse conjunto não foi rotulado por um dos usuários, a comparação com usuários similares permite prever qual rating seria dado ao item (Schafer et al. [2007]).

Em um sistema de *filtragem baseada em conteúdo*, as recomendações são realizadas a partir do conteúdo descritivo dos itens, como nome, título, descrição, resumo,

[1] `http://www.amazon.com/`
[2] `http://www.netflix.com/`

etc. Nessa abordagem, a estratégia é recomendar itens cujo conteúdo descritivo seja semelhante ao de itens pelos quais o usuário já demonstrou interesse no passado.

Em um sistema de *filtragem híbrida* a recomendação é feita com a utilização de mais de uma abordagem de recomendação, ou seja, envolvendo filtragem colaborativa e filtragem baseada em conteúdo (Burke [2002]).

Há diversos exemplos de sistemas de recomendação que utilizam essas duas abordagens. Agrawal et al. [2009] demonstram como o uso conjunto de recomendação baseada em conteúdo e recomendação baseada em filtragem colaborativa melhora a eficiência e utilidade das notícias para seus usuários. Guimarães et al. [2013] utilizam uma abordagem híbrida adaptável para recomendar usuários nas redes sociais. Eles apresentam um algoritmo que aprende a combinar diferentes fontes de evidências, incluindo a saída de outros algoritmos, usando um modelo de regressão logística.

Zhao et al. [2015] criaram uma técnica baseada na fatoração de matriz (Koren et al. [2009]), chamada "fatoração de comportamento" para prever os melhores tópicos de notícias para recomendar aos usuários de uma rede a partir do comportamento prévio do usuário. Eles identificam possíveis comportamentos como comentar, curtir, compartilhar, gerar novos posts, etc. e recomendam tópicos aos seus usuários dependendo da ação que ele esteve fazendo.

Claypool et al. [1999] afirmam que a filtragem colaborativa combina as opiniões informadas pelos usuários para fazer previsões personalizadas e precisas, enquanto a filtragem baseada em conteúdo usa a velocidade dos computadores para fazer previsões completas e rápidas. Assim eles utilizam de uma abordagem híbrida para o problema de recomendação de notícias, tentando capturar as principais características de cada abordagem em separado.

Recomendação de Notícias

A recomendação clássica é a abordagem de filtragem colaborativa (Adomavicius & Tuzhilin [2005]). Dados n usuários e m itens, a matriz $Y \leftarrow U \times I$ é a matriz de ratings. Para cada usuário U_j, com $j \in \{1, \ldots, n\}$, e cada item I_k, com $k \in \{1, \ldots, m\}$, a entrada $Y_{j,k}$ é preenchida com um valor de rating, se o usuário U_j avaliou o item I_k, ou com um marcador de desconhecido (por exemplo: NA, ?, – ou $\varnothing$), caso contrário. O problema usual de recomendação é predizer automaticamente com qual valor de rating um usuário U_j preencheria os itens ainda não rotulados por ele (completar a coluna j da matriz Y com valores do conjunto de ratings) e recomendar ao usuário somente os itens com alto valor predito de rating.

Os dados utilizados neste trabalho são de jornais online e são constituídos pelas

leituras de artigos sucessivos feitas por usuários. Neste caso, os itens são os artigos disponíveis para leitura. O problema de recomendação de notícias é diferente da recomendação de itens de compra. Não há ratings disponíveis. Há somente a informação de que um usuário leu um determinado artigo. Para a maioria dos portais de notícias online, há somente a informação do click que o usuário gerou para ler um artigo. Não há uma classificação indicando se o artigo foi interessante para o usuário. Na prática, a informação de click em um artigo é considerada uma classificação positiva, indicando o interesse do usuário por aquele conteúdo. Entretanto, falta uma classificação negativa explícita já que não há informação de artigos pelos quais o usuário não se interessa (Das et al. [2007]; Esiyok et al. [2014]; Zhao et al. [2015]).

Comumente, a recomendação de novos artigos é feita baseada em itens similares aos vistos pelo usuário. Isso indica que a abordagem preferida no caso de notícias é aquela baseada em conteúdo. Como os jornais online possuem conteúdos muito dinâmicos, os métodos de filtragem colaborativa tradicionais são de difícil aplicação (Li et al. [2010]).

Os artigos de jornais online possuem classificação por tópicos tais como *esportes*, *entretenimento*, *política*, ou *saúde*. Normalmente, o tópico é levado em consideração na hora de recomendar novos artigos. Se um usuário está lendo artigos de um determinado tópico, outros artigos do mesmo tópico costumam ser recomendados a ele.

De acordo com Billsus & Pazzani [2007], a recomendação de notícias apresenta algumas particularidades em comparação com outros domínios. Essas particularidades incluem conteúdos extremamente dinâmicos, com novos itens sendo criados e itens mais antigos perdendo interesse rapidamente. Outro aspecto é a necessidade de gerar notícias frescas para manter o interesse do usuário. Notícias quentes, de última hora, ou simplesmente recentes. A maioria dos usuários interage com uma pequena fração de notícias disponíveis. Essa pequena fração avaliada rapidamente torna-se desatualizada e desinteressante. Logo, os sistemas de recomendação de notícias lidam com dados altamente esparsos num ambiente muito dinâmico.

Hsieh et al. [2016] propõem um modelo de recomendação de notícias e eventos centrada no usuário, denominado Recomendação Imersiva. O principal pressuposto é que indivíduos geram vestígios ou traços digitais quase permanentemente, tais como mensagens no *Twitter*[3], assuntos nos cabeçalhos de e-mails, o histórico do navegador web e registros de compras digitais. Esses traços refletem quem somos, o que fazemos, e no que estamos interessados. Os autores geram perfis de interesses dos usuários com base nos seus vestígios digitais encontrados em diferentes plataformas disponibilizadas

[3] `http://www.twitter.com/`

por cada usuário. Para recomendação, eles utilizam um algoritmo híbrido que junta as informações dos perfis pessoais identificados nos vestígios digitais com perfis de itens e ratings existentes. Porém ao analisar sua estratégia, eles esbarraram no problema da falta de disposição do usuário em fornecer informações. Pouquíssimos usuários permitiram acessos a suas contas de Twitter e e-mail para o sistema gerar os perfis de traços digitais.

Semelhante a esse último trabalho, De Francisci Morales et al. [2012] propõem recomendar notícias explorando informações da conta do Twitter dos usuários. O seu algoritmo de recomendação mescla informações de amizade, histórico de visualização e tópicos populares da rede social do usuário com a popularidade dos artigos. Os resultados mostram que a abordagem melhora a acurácia de predição, sendo um recurso útil para entender o comportamento do usuário. Entretanto, a melhora só foi calculada para o percentual de usuários para os quais havia informações disponíveis para coletar. A utilização de informação de outras plataformas, novamente, só é útil perante a disponibilidade dessas informações extras. Logo, nós acreditamos que a utilização de somente informações mais elaboradas da própria plataforma, como o comportamento normal dos usuários é uma fonte mais confiável para trabalhar pois sempre estará disponível.

Veja o exemplo de Li et al. [2010]. Esse trabalho modela o problema de recomendação de artigos de notícias personalizada como o problema do bandido contextual (*contextual bandit problem*). Nessa abordagem o algoritmo de aprendizagem seleciona sequencialmente os artigos para servir os usuários com base nas informações contextuais dos usuários e artigos. Simultaneamente, ele adapta a estratégia de seleção de artigos baseado no feedback dos cliques dos usuários, tentando assim maximizar o número de cliques dos usuários. Os autores mostram que sua abordagem é melhor que as abordagens de contexto tradicionais.

Comportamento de Usuários

Enquanto trabalhos sobre recomendação de notícias são comuns, estudos sobre o comportamento dos usuários para dar substrato ao sistema de recomendação são mais raros. A seguir, vamos revisar alguns dos principais trabalhos cujo foco é a caracterização do comportamento dos usuários na web, não necessariamente no contexto de recomendação de notícias.

Agichtein et al. [2006] afirmam que avaliar as preferências de usuários de máquinas de buscas na web é crucial para o desenvolvimento, entendimento e manutenção desses sistemas. A premissa é que eles podem transformar as interações dos usuários com o

sistema de busca em julgamento de relevância. Eles estudam o comportamento dos usuários através da sua preferência em relação aos resultados da pesquisa fornecidos pela máquina de busca. Os autores criam um modelo para esse comportamento e mostram como ele pode prever as preferências nas buscas futuras.

Kwak et al. [2010] estudaram o comportamento de usuários do Twitter e identificaram algumas peculiaridades nos hábitos dos usuários dessa rede social. Eles identificaram que as principais mensagens classificadas como o assunto do momento (*trending topics*) são, na sua maioria, mensagens contendo manchetes ou conteúdo abreviado de notícias de jornais online. A partir dessa constatação eles perguntam se o Twitter é uma rede social ou uma mídia de notícias. Eles também identificaram que, quando uma mensagem é retuitada por outro usuário, ela tende a receber mais retuítes e que os usuários não seguem uns aos outros segundo a lei de potência.

Kumar & Tomkins [2010] pediram permissão a um conjunto de usuários para coletar informação de visualização de páginas para tentar caracterizar o seu comportamento online. Eles obtiveram mais de 50 milhões de *pageviews* em uma semana. Os autores identificaram que os usuários ficam online normalmente uma hora seguida e visualizam uma média de 59 páginas diferentes. A grande maioria dos usuários volta a usar a web 12 horas após finalizar o seu acesso. Poucos são os usuários que só acessam a web uma vez por dia. Eles identificaram os sites mais populares e criaram uma taxonomia para classificar as páginas: de conteúdo (sites de notícias e sites de multimídia tais como vídeos, músicas ou imagens, portais de conteúdo, sites de compras e sites de conteúdo adulto); de comunicação (e-mail, redes sociais, blogs e fóruns); e de busca (páginas de máquinas de buscas, páginas de busca em sites de multimídia e em sites de compras). Metade das páginas visualizadas na web são de conteúdo, um terço são de páginas de comunicação e o restante é formado por páginas de busca. Em geral, um usuário que começa em uma página de um desses três tipos tende a não mudar o tipo de página ao longo de sua sessão de acesso.

O trabalho de Chen et al. [2015] estuda o comportamento dos usuários de um jornal online. Eles caracterizam as leituras dos usuários pelos domínios como um fluxo de dados e propõem dois modelos estatísticos. As notícias do jornal estudado foram classificadas em 22 tópicos considerados como nós de um grafo. Foi criado um grafo dinâmico, que é atualizado a cada 5 minutos de observação da massa de usuários. Ao invés de acompanhar o usuário individualmente, esses autores monitoraram apenas as contagens determinadas pelos fluxos entre os tópicos criando uma matriz de transição dinâmica entre os tópicos. Um das principais conclusões foi que a maioria dos visitantes fica em apenas um tópico, em vez de passear pelos tópicos disponíveis.

O trabalho de Esiyok et al. [2014] é o mais próximo do trabalho desenvolvido nesta

dissertação. Eles estudaram os hábitos de leitura dos usuários de um portal de notícias online. A premissa motivadora deles é a mesma que a nossa: o estudo dos hábitos de leitura dos usuários pode fornecer uma visão útil para projetar melhores sistemas de recomendação de notícias. Eles modelaram o processo de leitura sequencial como um processo de Markov estacionário de primeira ordem e estimaram as probabilidades de transição entre as categorias de notícias presentes. Não foram considerados outros modelos alternativos. Embora o trabalho seja apresentado como um estudo preliminar, não encontramos outros trabalhos dos mesmos autores posteriores a esse. Nosso trabalho também estudou o modelo Markoviano de primeira ordem, porém fomos mais além. Comparamos esse modelo com diversos modelos, dentre eles, modelos Markovianos de ordem superior.

1.3 Objetivo

A hipótese principal com que iniciamos este trabalho foi que os usuários de jornais online leem as notícias de modo similar aos usuários de jornais impressos. Nossa suposição era que os usuários leriam artigos de um determinado tópico até esgotarem seu interesse por aquele domínio. Eles então encerrariam a sessão ou passariam a ler artigos de outro tópico até esgotar aquele assunto. Nossa concepção inicial imaginava cada sessão como um passeio pelos tópicos sem muita chance de retornar a um tópico após sair dele. Uma sessão típica seria, por exemplo, composta de 3 artigos de esportes lidos em sequência, seguidos por 2 artigos de política, quando então o usuário leria 3 artigos seguidos sobre entretenimento. Se esse fosse o comportamento do usuário, poderíamos construir sistemas de recomendação para explorá-lo seguindo uma modelagem simples: Quando um usuário abre uma sessão por um artigo de esportes, deve-se preferencialmente recomendar artigos desse mesmo tópico. Ao se aproximar do número médio de artigos lidos de um assunto, deve-se passar a recomendar artigos de outro tópico. Ao entrar num artigo de um terceiro tópico, as recomendações de artigos dos dois primeiros tópicos deveriam ser desestimuladas.

Como os sistemas de recomendação de notícias normalmente recomendam mais de um artigo para o usuário ler em seguida, várias políticas poderiam ser desenvolvidas. Por exemplo, para um usuário que começa a ler artigos de esportes deveríamos aumentar aos poucos o número de artigos de outros tópicos recomendados. Identificar a velocidade em que isso deveria ser feito e quais outros tópicos deveriam ser recomendados era um dos objetivos principais a ser explorado na dissertação.

Entretanto, quando começamos a analisar os dados dos jornais online verificamos

rapidamente que nossa hipótese sobre o comportamento do usuário não tinha sustentação nos dados. Os usuários não liam da forma que imaginamos. Eles retornam com frequência a notícias de tópicos que já haviam sido abandonados. Assim, é comum termos uma sessão formada por um artigo de esportes, seguido por um de política, novamente um de esporte e terminando com mais um de política. Isso aconteceu com frequência suficiente para que nosso modelo inicial fosse completamente descartado como bom descritor dos dados.

Fizemos então uma mudança drástica de objetivo. Nosso interesse passou a ser a caracterização do comportamento dos usuários de notícias online. Devido à grande variabilidade presente nos dados, nossa intenção é propor modelos probabilísticos que capturem a essência dos hábitos de leitura dos usuários. A sequência de tópicos lidos em uma sessão de um usuário é vista como uma instanciação de uma trajetória de um processo estocástico. O objetivo principal deste trabalho é **propor um modelo probabilístico que descreva de forma sucinta e aproximada os hábitos de leitura dos usuários de jornais online.** A intenção é formular uma estrutura matemática simples, mas não trivial, que represente os aspectos essenciais e mais relevantes do fenômeno. Semelhante a uma caricatura, um bom modelo probabilístico não é um retrato fiel e perfeito de um indivíduo, mas um esboço que reproduz e até amplifica ou exagera os seu traços mais marcantes de forma a torná-lo facilmente reconhecível. Esses princípios guiam a modelagem do processo de leitura sequencial de artigos e tópicos de jornais online desenvolvidos nesta dissertação.

1.4 Contribuições

Após um estudo exploratório da permanência e transição entre tópicos, desenvolvemos modelos estocásticos para prever o próximo tópico a ser lido levando em conta diferentes resumos da história de leitura anterior na sessão. Ao todo, foram 32 modelos que podem ser divididos em cinco categorias. A primeira categoria possui os **modelos sem influência do passado**, onde a informação de tópicos prévios é totalmente desconsiderada. Esses modelos constituem uma espécie de *straw man alternatives*, modelos muito simples e pouco realistas, que são considerados apenas para medirmos quão afastados eles estão dos dados empíricos. A segunda categoria é composta pelos **modelos de memória curta**, em que apenas as leituras recentes afetam o futuro. O modelo clássico nessa categoria é o modelo de cadeia de Markov, que condiciona o futuro pelos tópicos das leituras do passado na ordem em que ocorreram. A terceira categoria é composta pelos **modelos de preferência revelada**, onde o futuro é condicionado somente a

características de um tópico por vez. A quarta categoria é a de **modelos de permanência geométrica**. Nessa classe, decompomos o processo de leitura como entrada num tópico, um tempo aleatório de permanência nele e transição para um novo tópico em função do passado de leituras. A modelagem seguiu essa decomposição conceitual, dividindo o problema em dois módulos distintos: estudar as transições entre tópicos e estudar a permanência num tópico. Finalmente, a quinta categoria é composta pelos **modelos de vantagem cumulativa**, nos quais as leituras prévias dos tópicos aumentam as chances de sua leitura no futuro. O princípio é o de que uma pequena vantagem inicial, ocasionada pelo interesse exibido na primeira leitura, vai acumular vantagens adicionais com o tempo. Isso implica numa memória de longo prazo, em que pequenas perturbações iniciais podem se propagar no tempo, impactando bastante o futuro.

Todos os modelos foram ajustados por máxima verossimilhança e comparados de acordo com a qualidade do ajuste (*goodness of fit*), complexidade do modelo e capacidade de previsão. Usamos o critério de Informação de Akaike para avaliar o ajuste e a complexidade e o escore de Brier para avaliar a predição. Os melhores modelos são aqueles nos quais o usuário se move pelos tópicos influenciado pelos tópicos e a ordem de suas leituras mais recentes, os modelos de memória curta. Os modelos de vantagem cumulativa vieram logo atrás, com previsões ligeiramente piores, mas ainda bastante satisfatórias e competitivas com os modelos de memória curta.

Em resumo, as principais contribuições desta dissertação são as seguintes:

- Estudo exploratório e estatístico de duas grandes grandes bases de dados de jornais online descrevendo suas principais características.
- Proposta de cinco categorias de modelos estocásticos, bem como 32 instanciações desses modelos, para descrever o comportamento do usuário de jornais online.
- Ajuste por máxima verossimilhança e comparação dos modelos quanto ao ajuste, complexidade e capacidade preditiva do próximo tópico lido a partir da sequência de tópicos anteriores.

A seguir, no Capítulo 2, os modelos e as métricas utilizadas nessa dissertação serão melhores apresentados. No Capítulo 3, os dados de jornais online são apresentados pelas suas características básicas e no Capítulo 4, as análises exploratórias nos mostram características mais complexas desses dados. No Capítulo 5, os resultados dos experimentos são apresentados e a dissertação é concluída com as considerações finais do Capítulo 6. Alguns cálculos, tabelas e gráficos adicionais que explicitam certas partes do texto podem ser vistos nos Apêndices.

Capítulo 2

Modelos para Tópicos

> *"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."*
>
> — John von Neumann

Neste capítulo, apresentamos toda a modelagem conceitual utilizada nesse trabalho. As dez primeiras seções apresentam os modelos estocásticos. Na seção 2.11 os modelos são divididos por grupos e na seção 2.12 as métricas utilizadas para comparar os modelos são apresentadas.

Seja $u \in U = \{1, 2, \ldots, N\}$ o índice de um usuário e n_u o número de itens que ele lê em uma sessão S_u. Uma sessão é formada pela informação da sequência sucessiva dos tópicos dos artigos lidos e é denotada por $S_u = (T_{u,1}, T_{u,2}, \ldots, T_{u,n_u})$, onde $T_{u,i}$ é o tópico do i-ésimo item lido. Temos $T_{u,i} \in \mathcal{L} = \{1, 2, \ldots, L\}$, o conjunto dos *rótulos* identificadores dos tópicos.

Vamos adotar um modelo probabilístico para representar a coleção de sessões. Para isso, veremos os caminhos como trajetórias de processos estocásticos a tempo discreto com espaço de estados $\mathcal{L}$. Cada sessão gera um caminho aleatório $(1, T_{u,1}),\ (2, T_{u,2}),\ \ldots,\ (n_u, T_{u,n_u})$ no reticulado $\mathbb{N} \times \mathcal{L}$.

A Figura 2.1 ilustra a sessão S_u vista como a trajetória de um passeio aleatório. No eixo vertical, temos os estados (ou tópicos). No eixo horizontal, o índice da ordem de leitura dos artigos na sessão. No exemplo da figura, o usuário u fez uma sessão de leituras que gerou a trajetória $S_u = (T_{u,1},\ T_{u,2},\ T_{u,3},\ T_{u,4},\ T_{u,5},\ T_{u,6}) = (1, 8, 1, 6, 6, 6)$ no reticulado. O usuário começou lendo um artigo do tópico 1 e em seguida leu um artigo do tópico 8. Depois voltou a ler um artigo do tópico 1, e por último leu três artigos do tópico 6.

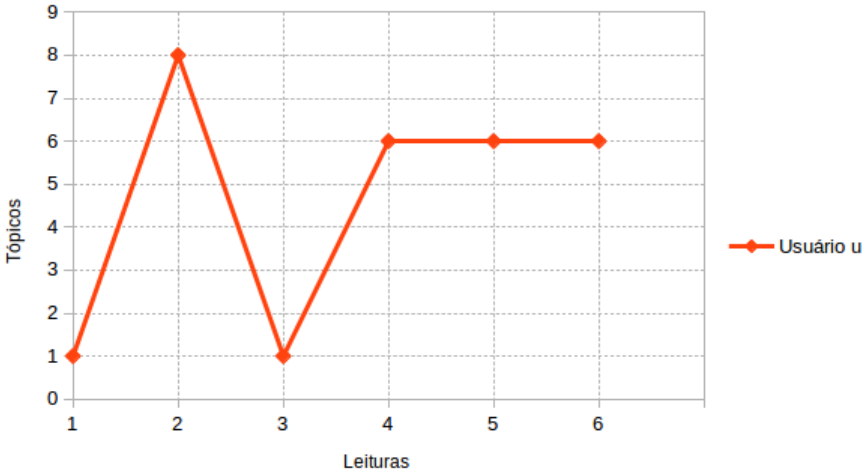


Figura 2.1: Exemplo de sessão vista como trajetória. O usuário u gerou a sessão $S_u = (T_{u,1},\ T_{u,2},\ T_{u,3},\ T_{u,4},\ T_{u,5},\ T_{u,6}) = (1, 8, 1, 6, 6, 6)$.

A probabilidade de qualquer sequência de tópicos numa sessão é dada pela multiplicação das probabilidades das leituras sucessivas condicionadas nas leituras anteriores:

$$\mathbb{P}(T_1, T_2, \ldots, T_n) = \mathbb{P}(T_1)\ \mathbb{P}(T_2 \mid T_1)\ \mathbb{P}(T_3 \mid T_1, T_2) \ldots \mathbb{P}(T_n \mid T_1, T_2, \ldots, T_{n-1}) \quad (2.1)$$

Assim, a caracterização completa da distribuição de probabilidade que governa um processo estocástico requer a especificação de dois componentes:

1. A distribuição de probabilidade inicial ou do primeiro tópico lido numa sessão:

$$\mathbb{P}(T_{u,1} = l), \quad l \in \mathcal{L} \quad (2.2)$$

   Esse primeiro componente requer a especificação de $L = |\mathcal{L}|$ probabilidades.

2. A classe das distribuições de probabilidade do i-ésimo tópico lido condicionada na sequência de todos os tópicos lidos anteriormente:

$$\mathbb{P}(T_{u,i} = l_i \mid T_{u,1} = l_1, \ldots,\ T_{u,i-1} = l_{i-1}), \quad l_i \in \mathcal{L} \text{ e } 1 < i \leq n_u \quad (2.3)$$

   Esse segundo componente requer a especificação de um número muito maior de probabilidades que o primeiro. Para um número natural $n_u = n > 1$ arbitrário e para cada possível trajetória prévia $(l_1, \ldots, l_{n-1})$, precisamos especificar probabilidades para os L possíveis estados no passo n condicionados nos estados passados. Logo, são ao todo L^n probabilidades associadas com o

passo n. Portanto, para processos observados até um tempo N, isto requer $\sum_{n=2}^{N} L^n = (L^{N+1} - L^2)/(L-1) = O(L^N)$ elementos.

Dada uma trajetória parcial de um usuário numa sessão, desejamos estimar o tópico do próximo artigo a ser lido. Se tivermos as probabilidades (2.2) e (2.3) para todo n e toda trajetória, basta utilizar a probabilidade conjunta expressa como produto probabilidade condicional em (2.1) e identificar qual é o tópico mais verossímil para próxima leitura.

Todos os modelos probabilísticos impõem restrições na coleção de probabilidades condicionais em (2.3) que reduzem drasticamente o número de probabilidades necessárias para especificar completamente a distribuição do processo. Essas restrições apresentam-se sob a forma de suposições que procuram capturar toda a essência probabilística do processo. Como dissemos na introdução, o objetivo é encontrar uma formulação probabilística simples que capture os principais aspectos da leitura online.

Vamos começar assumindo que as probabilidades (2.2) e (2.3) aplicam-se a todos usuários e sessões. Como consequência, elas não dependem de u. Logo, a distribuição $\mathbb{P}(T_{u,1})$ do tópico inicial pode ser escrita simplesmente como $\mathbb{P}(T_1)$, e o número total de tópicos em uma sessão pode ser simplificado apenas como n. Vamos também nos referir à i-ésima leitura como sendo o instante de tempo i de uma sessão.

2.1 Abordagens Irrealistas e Elementares

Os primeiros modelos não são realistas e são utilizados apenas como ponto de referência para comparação com outros modelos. Um primeiro modelo básico, chamado **M-Uniforme**, adota probabilidades iguais para todos os tópicos em qualquer instante n. Esse modelo reduz a probabilidade condicional (2.3) e do primeiro tópico (2.2) a:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l) = 1/L \tag{2.4}$$

para $n \geq 1$ e onde $L = |\mathcal{L}|$ é o tamanho do conjunto dos tópicos.

Outro modelo irrealista, chamado de **M-Alta Permanência**, utiliza somente a informação do último tópico. Ele assume uma grande probabilidade para a permanência no mesmo tópico da leitura corrente e probabilidades iguais e pequenas para as transições aos demais tópicos. A probabilidade do primeiro tópico (2.2) é assumida uniforme, como no modelo anterior: $\mathbb{P}(T_1 = l\ ) = 1/L$. Para as probabilidades

condicionais em (2.3), esse modelo estabelece que:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \begin{cases} p, & \text{se } l = l_{n-1} \\ (1-p)/(L-1), & \text{se } l \neq l_{n-1} \end{cases} \tag{2.5}$$

onde p é a probabilidade de permanência no último tópico lido.

2.2 Probabilidade Inicial mais Realista

Para flexibilizar os modelos irrealistas vistos na seção anterior, vamos começar considerando um modelo mais adequado para as probabilidades iniciais em (2.2). Tipicamente, alguns tópicos atraem mais que outros fazendo com que as probabilidades $\mathbb{P}(T_1 = l)$ sejam não uniformes. Como não temos nenhuma especificação prévia para essas probabilidades, em todos os próximos modelos elas serão consideradas completamente não estruturadas, podendo assumir qualquer valor no simplex $(\alpha_1, \ldots, \alpha_L)$ onde $\alpha_l \geq 0$ e $\alpha_1 + \cdots + \alpha_L = 1$.

Essas probabilidades iniciais possuem uma estimativa óbvia, ótima e simples, a estimativa de máxima verossimilhança. Neste problema, essa estimativa é simplesmente a frequência empírica da seleção dos diversos tópicos na primeira leitura dos usuários. Essa estimativa será utilizada em todos os modelos considerados nesta dissertação com exceção daqueles da seção anterior.

Essa probabilidade inicial não é de muito interesse em nosso problema, em que o objetivo principal é predizer os tópicos das leituras sucessivas de artigos, e não o primeiro deles. As probabilidades $\mathbb{P}(T_1 = l)$ são necessárias apenas para a completude dos modelos.

2.3 Independência

O modelo **M-Independência** assume que os tópicos das leituras sucessivas são variáveis aleatórias independentes. Assim, os tópicos já vistos não influenciam no tópico do próximo artigo. A probabilidade condicional em (2.3) fica reduzida a:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l) \tag{2.6}$$

As probabilidades de um mesmo tópico l em diferentes instantes de tempo podem ser diferentes. Assim, podemos ter $\mathbb{P}(T_n = l) \neq \mathbb{P}(T_m = l)$ se $n \neq m$. Nesse modelo, nenhuma estrutura é imposta no vetor de probabilidades associado com o tempo n.

A estimativa de máxima verossimilhança de $\mathbb{P}(T_n = l)$ é trivial, sendo igual à frequência de leitura do tópico l no instante n. Assim, para todo instante arbitrário n, é necessário especificar L probabilidades.

Outro modelo mais restritivo assume que $\mathbb{P}(T_n = l)$ não varia com n. Isto é, a popularidade de um tópico independe do passado e do instante de leitura n. Esse modelo, chamado de **M-Independência Homogênea**, especifica que:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l) = \mathbb{P}(T_1 = l) \tag{2.7}$$

para todo n. Os valores das probabilidades são estimados como a proporção de leitura do tópico independente do instante de leitura. Dessa forma, é necessário especificar somente L probabilidades.

2.4 Markov

Na fatoração da Equação (2.3) observe que a probabilidade associada com a i-ésima probabilidade condicional é uma i-upla que requer a especificação de L^i valores reais. Em toda sua generalidade, ela assume que a probabilidade de escolher o próximo tópico depende de toda a história pregressa na sessão. Uma simplificação muito popular para reduzir ao máximo essa complexidade é assumir um modelo de *cadeia de Markov*. Nesse modelo, a probabilidade condicional depende apenas do último tópico lido. Isto é, a probabilidade do próximo tópico depende apenas do tópico que está sendo lido no momento, e não do modo como se chegou a esse tópico.

O modelo **M-Markov-I** assume que:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, \ T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l \mid T_{n-1} = l_{n-1}) \tag{2.8}$$

para todo $n \geq 2$ e para toda sequência de tópicos. Para cada instante $n \geq 2$, é necessário especificar as L^2 probabilidades em (2.8).

Uma simplificação adicional assume que essas probabilidades são invariantes com n, ficando constantes a medida que o tempo passa:

$$\begin{aligned} \mathbb{P}(T_n = l \mid T_{n-1} = l_{n-1}) = \mathbb{P}(T_2 = l \mid T_1 = l_{n-1}), \\ \forall l \in \mathcal{L} \text{ e } \forall\ n > 1. \end{aligned} \tag{2.9}$$

Esse processo é chamado de *cadeia de Markov homogênea* e será denotado nesta dissertação como **M-Markov-I Homogêneo**. Ele requer a especificação de apenas

L^2 probabilidades para definir todas as probabilidades condicionais em (2.3).

Os modelos de Markov apresentados anteriormente levam em consideração somente o último instante, sendo por isso chamados de modelos de primeira ordem. Quando o modelo retém a memória de mais instantes do passado nas probabilidades condicionais, esse é chamado de modelo de ordem superior.

O modelo de Markov de 2ª ordem leva em consideração os dois instantes anteriores para o cálculo da probabilidade do próximo tópico. O modelo **M-Markov-II** é definido como:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l \mid T_{n-2} = l_{n-2}, T_{n-1} = l_{n-1}) \quad (2.10)$$

Para $n > 2$ arbitrário, é necessário especificar L^3 probabilidades. Quando $n = 2$, só há um instante no passado. Nesse caso, o modelo acima se reduz ao modelo de Markov de primeira ordem.

Como no modelo de primeira ordem, podemos simplificar o modelo Markoviano de segunda ordem forçando as probabilidades de transição em (2.10) a serem constantes no tempo, não variando com n. Isso dá origem ao modelo **M-Markov-II Homogêneo**:

$$\mathbb{P}(T_n = l \mid T_{n-2} = l_{n-2},\ T_{n-1} = l_{n-1}) = \mathbb{P}(T_3 = l \mid T_1 = l_{n-2},\ T_2 = l_{n-1}), \quad \forall l \in \mathcal{L} \text{ e } \forall\ n > 2. \quad (2.11)$$

Podemos considerar modelos de Markov de ordem mais elevada, necessitando somente haver histórico suficiente para estimar o modelo. Por exemplo, a partir do 3º artigo lido, temos histórico para calcular um modelo de Markov de 3ª ordem. E com 4 instantes ou mais, podemos calcular um modelo de 4ª ordem.

O modelo de Markov de ordem k precisa de no mínimo k instantes de passado e especifica L^{k+1} probabilidades em cada instante. Caso utilizemos os modelos invariantes com n, é necessário somente L^{k+1} probabilidades para todos os instantes $n > k$. Para os demais instantes $n \leq k$, utilizamos os modelos de ordem menor.

Os modelos de ordem superior requerem a estimação de muitas probabilidades. Para n relativamente grande teremos poucas sessões disponíveis para realizar a estimação já que apenas as sessões de tamanho maior que n podem ser utilizadas. Por causa disso, para ordem igual ou maior a 3, nós utilizaremos os modelos de ordem simplificados (invariantes com n). Não vamos considerar modelos de ordem acima de 4 nesta dissertação.

Os modelos **M-Markov-III Homogêneo** e **M-Markov-IV Homogêneo** são definidos respectivamente em 2.12 e 2.13.

$$\begin{aligned}\mathbb{P}(T_n = l \mid T_{n-3} = l_{n-3}, T_{n-2} = l_{n-2}, T_{n-1} = l_{n-1}) = \\ \mathbb{P}(T_4 = l \mid T_1 = l_{n-3}, T_2 = l_{n-2}, T_3 = l_{n-1}), \quad \forall l \in \mathcal{L} \text{ e } \forall\ n > 3.\end{aligned} \tag{2.12}$$

$$\begin{aligned}\mathbb{P}(T_n = l \mid T_{n-4} = l_{n-4}, T_{n-3} = l_{n-3}, T_{n-2} = l_{n-2}, T_{n-1} = l_{n-1}) = \\ \mathbb{P}(T_5 = l \mid T_1 = l_{n-4}, T_2 = l_{n-3}, T_3 = l_{n-2}, T_4 = l_{n-1}), \quad \forall l \in \mathcal{L} \text{ e } \forall\ n > 4.\end{aligned} \tag{2.13}$$

2.5 Histórico de Visitas

Neste modelo, vamos reduzir a história completa da sessão permitindo que o passado mais longínquo afete as probabilidades do próximo passo. Não vamos usar apenas a memória mais recente do processo. Precisamos considerar um resumo do passado que não escale com n. Assim, vamos assumir que a probabilidade de leitura do tópico l no artigo n depende apenas do número de vezes que o usuário visitou o tópico l previamente. Essa é a única característica do passado que vai influenciar a próxima escolha. Não importa quais outros tópicos foram visitados ou a ordem em que isso foi feito.

Seja $S^l_{n-1} = \sum_{i=1}^{n-1} I[T_i = l]$. Isto é, S^l_{n-1} é o número de vezes que o usuário leu artigos do tópico l numa sessão de tamanho $n-1$. O modelo **M-Histórico de Visitas** determina que:

$$\begin{aligned}\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l \mid S^l_{n-1} = s) = f_n(l, s) \\ s = 0, \ldots, n-1.\end{aligned} \tag{2.14}$$

Não estamos assumindo nenhuma forma particular de variação. Podemos ter essa probabilidade aumentando com s. Ou podemos ter a probabilidade aumentando com s até certo ponto quando então ela passa a diminuir. Esse modelo teria uma memória de cada tópico visitado e as probabilidades do próximo tópico variam com o número de visitas anteriores ao tópico em questão. Nesse caso, especificar as probabilidades requer a estimação da coleção de valores $f_n(l, s)$. Para um instante n arbitrário, teremos $L \times n$ probabilidades.

Esse modelo conta quantas leituras prévias foram do tópico alvo l no histórico da sessão. Essa contagem assume valores de 0 a $n-1$, aumentando linearmente com n. Uma alternativa a esse modelo restringe a contagem aos últimos m artigos da sessão.

Seja $S^l_{n-1,m} = \sum_{i=n-m}^{n-1} I[T_i = l];\ 2 \leq m < n-1$. Isto é, $S^l_{n-1,m}$ é o número de vezes que o usuário leu artigos do tópico l nos últimos m artigos numa sessão de tamanho $n-1$. O modelo **M-Últimas m Visitas** determina que:

$$\mathbb{P}(T_n = l \mid T_1 = l_1,\ ..,\ T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l \mid S^l_{n-1,m} = s) = f_{n,m}(l,s) \quad s = 0, \ldots, m. \tag{2.15}$$

A probabilidade de leitura do tópico l no passo n depende apenas do número de vezes que o usuário visitou o tópico l nos últimos m artigos. Veja que, nesse caso, especificar as probabilidades requer a estimação da coleção de valores $f_{n,m}(l,s)$. Para $2 \leq m < n-1$, isto significa $L \times m$ termos.

Se $m = n-1$ teríamos exatamente o modelo anterior. Com $m = 1$, o modelo utilizaria somente o último estado, assemelhando-se com o modelo Markoviano. Ele não seria idêntico ao modelo de Markov de primeira ordem pois a probabilidade condicional ao último tópico lido tem somente duas possibilidades: o último estado é igual ao tópico alvo ou o último estado foi um estado diferente do tópico alvo. Com $m = 0$, o modelo não utilizaria nenhuma informação de passado se assemelhando ao modelo de independência, o modelo (2.6).

2.6 Duração no Estado

Este modelo propõe outro mecanismo para a transição entre leituras sucessivas. Nele, a probabilidade de visitar o tópico l no passo n depende apenas de quantos artigos desse mesmo tópico foram lidos contando retroativamente a partir do último artigo até a mudança mais recente de tópico. Isto é, de maneira retrospectiva a partir do instante atual, conta-se quantas vezes o tópico l foi lido de forma ininterrupta. A característica de interesse passa a ser o tempo de duração no estado atual.

Sendo mais específico, com $k \in \{0,\ 1,\ \ldots,\ n-1\}$, define-se $D^l_{n-1} = k$ se $S = (\ldots, T_{n-(k+1)} \neq l, T_{n-k} = l, \ldots, T_{n-1} = l)$. Isto é, as últimas k leituras consecutivas foram do tópico l. Observe que podemos ter $D^l_{n-1} = 0$, caso o último tópico seja diferente de l. Outro caso particular é quando não existe o tópico $T_{n-(k+1)} \neq l$. Isto é, a sessão já começou no tópico-alvo l e, até o instante atual, todas as leituras foram desse mesmo tópico. Logo, $k = n-1$ quando não existe $T_{n-(k+1)} \neq l$.

Sendo mais rigoroso, podemos definir:

$$D^l_{n-1} = \begin{cases} \min\{k < n-1; \quad T_{n-(k+1)} \neq l\}, & \text{caso exista } T_{n-(k+1)} \neq l; \\ n-1, & \text{caso contrário.} \end{cases}$$

O modelo **M-Duração no Estado** reduz a probabilidade de leitura da seguinte forma:

$$\mathbb{P}(T_n = l \mid T_1 = l_1,\ \ldots,\ T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l \mid D^l_{n-1} = k) = g_n(l,k) \qquad k = 0, \ldots, n-1. \tag{2.16}$$

Esse modelo requer a especificação das probabilidades $\mathbb{P}(T_n = l \mid D^l_{n-1} = k)$. Para cada $n > 1$ arbitrário, isto significa especificar $L \times n$ probabilidades.

Como no modelo (2.15), esse modelo é limitado aos últimos m artigos para evitar o seu crescimento com n. O modelo reduzido (**M-Duração no Estado Últimos m Artigos**) assume que:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l \mid D^l_{n-1,m} = k) = g_{n,m}(l,k) \qquad k = 0, \ldots, m \tag{2.17}$$

onde:

$$D^l_{n-1,m} = \begin{cases} \min\{k < m; \quad T_{n-(k+1)} \neq l\}, & \text{caso exista } T_{n-(k+1)} \neq l; \\ m, & \text{caso contrário.} \end{cases}$$

e $m \in \{2, \ldots, n-2\}$. Esse modelo requer a especificação das L probabilidades $g_{n,m}(l,k) = \mathbb{P}(T_n = l \mid D^l_{n-1,m} = k)$ para cada valor de k possível. Logo, a cada instante $n > 1$ é necessário especificar $L \times m$ probabilidades.

2.7 Duração da Última Visita

O modelo de duração no estado tem um inconveniente. Se o estado atual é l, o modelo volta retrospectivamente para conhecer D^l_{n-1} e saber quanto tempo ele está nesse estado l. A chance de permanecer no estado l depende desse tempo de visita D^l_{n-1}. Entretanto, para todos os estados j diferentes do estado atual l, teremos $D^j_{n-1} = 0$ e a probabilidade $g_n(j,0)$ terá pouca flexibilidade, podendo assumir apenas um único valor para cada estado j, independentemente do restante da história da sessão. Parece uma restrição muito severa. O modelo a ser apresentado a seguir procura melhorar esse aspecto.

Dado que $T_{n-1} \neq l$, o modelo **M-Duração da Última Visita** considera que a probabilidade de ler artigo do tópico l no instante n depende da quantidade de artigos que foram lidos desse tópico em sequência pela última vez. Isto é, volta-se na história verificando quanto tempo durou a última visita feita ao tópico l, não importa quão

distante do passado, nem por qual tópico a visita foi interrompida.

Seja:

$$L_{n-1}^{l} = \max_{j}\{T_j = l\} - \min_{j}\{T_{j+1} = l \wedge T_j \neq l\}$$

o tempo de duração da última visita ao estado l. Podemos ter $L_{n-1}^{l} = 0$ se o estado l ainda não foi visitado na sessão. No outro extremo, se $L_{n-1}^{l} = n-1$, a sessão está no tópico l desde a primeira leitura. Reduzimos, então, a probabilidade de transição da seguinte forma:

$$\begin{aligned}\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l \mid L_{n-1}^{l} = k) = d_n(l,k) \\ k = 0, \ldots, n-1.\end{aligned} \tag{2.18}$$

Esse modelo requer a especificação das probabilidades $\mathbb{P}(T_n = l \mid L_{n-1}^{l} = k)$. Para $n > 1$ arbitrário, isso significa especificar $L \times n$ probabilidades.

2.8 Leituras Pós Saída

O próximo modelo chamado de **M-Leituras Pós Saída** considera que a probabilidade de ler um artigo do tópico l no próximo passo n depende da quantidade de artigos que foram lidos após sair pela última vez do tópico l. Defina $E_{n-1}^{l} = k$, com $k = n-1-j$ se a sessão for do seguinte tipo $S = (\ldots, T_j = l, T_{j+1} \neq l, \ldots, T_{n-1} \neq l)$. Ou seja,

$$E_{n-1}^{l} = n - 1 - \max_{j}\{[T_j = l]\}\,.$$

O modelo também define que $E_{n-1}^{l} = \infty$ se ainda não tiver havido leitura do tópico l na sessão. Assim, reduzimos a probabilidade de transição da seguinte forma

$$\begin{aligned}\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_n = l_n) = \mathbb{P}(T_n = l \mid E_{n-1}^{l} = k) = h_n(l,k) \\ k = 0, \ldots, n-2 \text{ ou } \infty.\end{aligned} \tag{2.19}$$

Observe que, quando $j = n-1$, a sessão termina em $T_j = l$ e nesse caso não houve a saída do tópico l ainda. Logo, $E_{n-1}^{l} = 0$. O modelo requer a especificação das probabilidades $\mathbb{P}(T_n = l \mid E_{n}^{l} = k)$, para $n > 1$ arbitrário, isso significa especificar $L \times n$ probabilidades.

2.9 Vantagem Cumulativa

A próxima classe de modelos assume que os tópicos escolhidos inicialmente possuem impacto permanente no restante da história da sessão de leituras. Os tópicos lidos no início ganham uma vantagem sobre os demais tópicos na forma de um bônus adicionado à sua probabilidade inicial. Essa vantagem permanece com o tópico ao longo da sessão e é incrementada de forma cumulativa a medida que ocorrem mais visitas ao tópico. Os dois modelos aumentam a probabilidade dos tópicos já vistos na sessão acumulando um bônus para o tópico caso ele seja lido.

Seja $S_{n-1}^l = \sum_{i=1}^{n-1} I[T_i = l]$ o número de vezes que o usuário leu artigos do tópico l numa sessão de tamanho $n-1$. Seja $\pi(l)$ a probabilidade inicial de leitura e β_l o parâmetro de bônus para o tópico l. O modelo **M-Vantagem Cumulativa A** assume que o bônus é acrescido de forma cumulativa e *aditiva* à uma probabilidade de base $\pi_n(l)$:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) \propto \pi_n(l) + \beta_l S_{n-1}^l \tag{2.20}$$

As probabilidades de base $\pi_n(l)$ são valores que podem variar com a ordem de leitura n. Observe que não é só a trajetória de leituras que é aleatória. Nesse modelo, as *probabilidades* condicionais são também aleatórias, variando estocasticamente ao sabor da história anterior de leituras. A cada passo, as probabilidades condicionais resultantes são normalizadas para somarem 1.

O modelo **M-Vantagem Cumulativa B** é uma versão multiplicativa do modelo anterior:

$$\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) \propto \pi_n(l)(1 + \beta_l S_{n-1}^l) \tag{2.21}$$

Como antes, as probabilidades devem ser normalizadas a cada passo.

Os valores de bônus β_l são parâmetros dos modelos a serem estimados por máxima verossimilhança. Fixados os valores dos L bônus, precisamos também especificar L probabilidades para o valores de base $\pi_n(l)$ para cada $n = 1, 2, \ldots$ Usamos duas abordagens. Na primeira, o valor é invariante nos instantes, deixando que a variação só aconteça pelos cálculos acumulativos: $\pi_n(l) = \pi(l)$. Na segunda abordagem, deixamos os valores de $\pi_n(l)$ variando livremente a cada instante, com a única restrição de que somem 1 sobre os tópicos. Assim, um tópico pode ser pouco provável num certo instante mas em outro instante ele pode ter uma chance maior de ser escolhido, independentemente da história prévia de leitura.

2.10 Permanência Geométrica

Uma última classe de modelos assume que há um padrão de permanência nos tópicos seguindo a distribuição geométrica e as transições de tópicos seguem funções específicas mais elaboradas. São essas funções de mudança que vão diferenciar os modelos. Para esta classe de modelos, vamos representar de outra forma a sessão de n artigos. A representação $S = (T_1, T_2, \ldots, T_n)$ será substituída por outra em que vamos mostrar cada tópico lido junto com o tempo de permanência neste tópico. Por exemplo, a sessão $S = (2, 5, 5, 5, 5, 2, 2, 2)$ será representada por $S' = (E_1 = 2,\ N_1 = 1,\ E_2 = 5,\ N_2 = 4,\ E_3 = 2,\ N_3 = 3)$ onde toda variável E_i denota o i-ésimo tópico que o usuário passou a ler e N_i é o número de artigos lidos naquele tópico. Um mesmo tópico pode aparecer mais de uma vez se sua leitura for interrompida por outro tópico. No exemplo acima, isso ocorre com o tópico 2, que foi o primeiro e terceiro tópico lido ($E_1 = 2$ e $E_3 = 2$).

Dada a sequência de tópicos de uma sessão de n artigos $S = (T_1, T_2, \ldots, T_n)$, definimos um vetor associado com indicadores binários dos pontos de mudança na sequência de tópicos. Mais formalmente, seja $\delta = [\delta_1, \delta_2, \ldots, \delta_n, \delta_{n+1}]$ onde $\delta_1 = 1$, $\delta_{n+1} = 1$ e, para $2 \leq i \leq n$:

$$\delta_i = \begin{cases} 0, & \text{se } T_i = T_{i-1} \\ 1, & \text{se } T_i \neq T_{i-1} \end{cases} .$$

Por exemplo, a sessão $S = (2, 5, 5, 5, 5, 2, 2, 2)$ possui o seguinte vetor de indicadores binários $\delta = (1, 1, 0, 0, 0, 1, 0, 0, 1)$.

Para cada par (i, j) onde $\delta_i = 1,\ i \leq n$ e $j = \min\{j > i;\ \delta_j = 1\}$, definimos $E_k = T_i$ e $N_k = j - i$, com $k \in \mathbb{N}$. E_k é o k-ésimo **estado** e N_k é a quantidade de artigos lidos em sequência do tópico do k-ésimo estado, ou simplesmente a **duração no estado**. Escrevemos uma sessão S na forma $S' = (E_1,\ N_1,\ E_2,\ N_2,\ \ldots,\ E_m,\ N_m)$.

Para estimar os parâmetros do modelo, precisamos obter a probabilidade da sessão S nessa nova representação S'. Isto é, precisamos saber calcular

$$\mathbb{P}(S) = \mathbb{P}(S') = \mathbb{P}(E_1,\ N_1,\ E_2,\ N_2,\ \ldots,\ E_m,\ N_m)$$

quando S for representada como S'. Regras usuais de probabilidade permitem escrever

$$\begin{aligned} \mathbb{P}(S') = \mathbb{P}(E_1,\ N_1,\ E_2,\ N_2,\ \ldots,\ E_m,\ N_m) = \\ \mathbb{P}(E_1) \times \mathbb{P}(N_1 \mid E_1) \times \mathbb{P}(E_2 \mid E_1, N_1) \times \mathbb{P}(N_2 \mid E_1, N_1, E_2) \times \ldots \\ \times\ \mathbb{P}(E_m \mid E_1,\ N_1, \ldots,\ E_{m-1},\ N_{m-1}) \times\ \mathbb{P}(N_m \mid E_1, N_1, \ldots,\ N_{m-1},\ E_m) \end{aligned}$$

Assumindo que a distribuição de N_i só depende do E_i e que E_i é independente de N_j para $i \neq j$, temos:

$$\begin{aligned}\mathbb{P}(S') = \mathbb{P}(E_1) \times \mathbb{P}(N_1 \mid E_1) \times \mathbb{P}(E_2 \mid E_1) \times \mathbb{P}(N_2 \mid E_2) \times \ldots \\ \times \mathbb{P}(E_m \mid E_1, , \ldots, E_{m-1}) \times \mathbb{P}(N_m \mid E_m)\end{aligned} \tag{2.22}$$

Esta classe de modelos será denotada em geral por **M-PG**, um acrônimo para **M-Permanência Geométrica**. A probabilidade da sessão $\mathbb{P}(S')$ requer o cálculo de $\mathbb{P}(E_1)$. Essa é simplesmente a probabilidade da escolha do tópico da primeira leitura e é idêntico a $\mathbb{P}(T_1)$. Vamos estimar essa probabilidade como antes, usando simplesmente a frequência empírica do primeiro tópico lido.

A seguir consideramos os fatores $\mathbb{P}(N_i \mid E_i)$ em (2.22). O modelo assume que todos esses fatores representam probabilidades de uma variável aleatória com distribuição geométrica. Seja $x_i \in \{1, 2, \ldots\}$ um dos possíveis valores para o número de leituras no i-ésimo tópico da sequência. Então,

$$\mathbb{P}(N_i = x_i \mid E_i = l_i) = \begin{cases} \mathbb{P}(\text{Geom}(\theta_{l_i}) = x_i) = (1-\theta_{l_i})^{x_i - 1}\theta_{l_i}, & \text{se } i < m \\ \mathbb{P}(\text{Geom}(\theta_{l_i}) \geq x_i) = 1 - \sum_{j=1}^{x_1 - 1}(1-\theta_{l_i})^{j-1}\theta_{l_i}, & \text{caso } i = m \end{cases} \tag{2.23}$$

onde θ_{l_i} é o parâmetro da distribuição geométrica. Este parâmetro pode ser específico por tópico ou ter um valor único para todos os tópicos.

Resta especificar um modelo para os fatores $\mathbb{P}(E_i \mid E_1, \ldots, E_{i-1})$ em (2.22). Esses fatores são calculados de três formas diferentes, gerando três variações do modelo **M-PG**. A primeira variação calcula o fator das probabilidades condicionais dos estados como uma simples renormalização das probabilidade removendo o estado anterior. Essa variação é chamada de **M-PG-A**. Como $E_i \neq E_{i-1}$ por construção da sessão S', definimos

$$\mathbb{P}(E_i \mid E_1, \ldots, E_{i-1}) = \begin{cases} \frac{\pi(E_i)}{1-\pi(E_{i-1})}, & \text{se } E_i \neq E_{i-1} \\ 0, & \text{caso contrário} \end{cases} \tag{2.24}$$

onde $\pi(E)$ é a probabilidade de leitura do tópico E calculado como a frequência relativa a todos os tópicos.

A segunda variação do modelo, chamada de **M-PG-B**, aumenta a probabilidade de volta a um estado já visto. Seja $I[E_i \neq E_j]$ a função indicadora valendo 1 se o tópico E_i é diferente do tópico E_j, e valendo 0, caso contrário. No momento da leitura do i-ésimo tópico E_i, a variável $s_i = \sum_{j=1}^{i-1} I[E_j \neq E_i]$ conta o número de tópicos já visitados que são diferentes do tópico E_i. Isto é, s_i mede quantas transições não

levaram à entrada no tópico E_i no histórico prévio da sessão. Como antes, $\pi(E)$ é a probabilidade de leitura do tópico E calculado como a frequência relativa a todos os tópicos. Dessa forma, especificamos o fator das probabilidades condicionais dos estados da seguinte forma:

$$\mathbb{P}(E_i|\ E_1, \ldots, E_{i-1}) \propto \begin{cases} \pi(E_i)^{s_i}, & \text{se } E_i \neq E_{i-1} \\ 0, & \text{caso contrário} \end{cases} \tag{2.25}$$

No momento em que vai ocorrer uma transição para um novo tópico, as probabilidades são modificadas. A probabilidade condicional de entrar no tópico E_i é modificada por s_i. Se s_i é uma contagem grande, o tópico foi pouco visitado e assim sua probabilidade é drasticamente reduzida ao tomarmos $\mathbb{P}(E_i|\ E_1, \ldots, E_{i-1})$ proporcional a $\pi(E_i)^{s_i}$, pois $0 < \pi(E_i) < 1$.

Todas as probabilidades de leituras dos tópicos (π) são elevadas a uma potência, a soma s_i das indicadoras, diminuindo as demais probabilidades. Quanto mais estados anteriores não forem do tópico avaliado, maior é a diminuição da probabilidade. Para que no final elas somem 1, há uma normalização sobre a soma de todas as probabilidades com exceção da probabilidade do tópico que está no estado anterior, E_{i-1}.

Quando $i = 2$, tendo somente o E_1 como passado, o fator (2.25) fica igual ao fator calculado pela primeira variação em (2.24).

A última variação do modelo (**M-PG-C**) também considera como alta a probabilidade de volta a um estado visto *recentemente*. Para isso, ela adiciona um bônus para o caso de retorno ao penúltimo estado E_{i-2}:

$$\mathbb{P}(E_i|\ E_1, \ldots, E_{i-1}) = \begin{cases} \frac{\pi(E_i)+\beta_{E_{i-2}}}{1-\pi(E_{i-1})+\beta_{E_{i-2}}}, & \text{se } E_i = E_{i-2} \\ \frac{\pi(E_i)}{1-\pi(E_{i-1})+\beta_{E_{i-2}}}, & \text{caso contrário} \end{cases} \tag{2.26}$$

onde $\pi(E)$ é a probabilidade de leitura do tópico E, e $\beta_{E_{i-2}}$ é um valor de bônus dado ao tópico do estado E_{i-2}. Como E_{i-2} pode assumir qualquer valor em $\mathcal{L}$, L valores para o parâmetro beta são necessários para esse modelo. Quando $i = 2$, não existe E_{i-2}. Logo o modelo (2.26) considera $\beta_{E_0} = 0$, resultando em cálculo igual ao da primeira variação (2.24).

2.11 Grupos de modelos

Todos esses modelos apresentados foram agrupados pelas suas caraterísticas. Os modelos foram divididos em cinco categorias: modelos sem influência do passado; modelos de memória curta; modelos de preferência revelada; modelos de permanência geométrica; e modelos de vantagem cumulativa.

Os modelos sem influência do passado são os modelos que associam probabilidades para o próximo tópico seguindo alguma regra categórica que despreza a informação dos tópicos do passado. Por exemplo, o modelo **M-Uniforme** assume que todos os tópicos são equiprováveis. Já os modelos **M-Independência** e **M-Independência Homogênea** associam probabilidades independentemente do tópico atual. Eles contabilizam as frequências de leituras dos tópicos na base de dados de todas as sessões como as probabilidades de próximo tópico.

Os modelos de memória curta são os modelos que condicionam as probabilidades do próximo tópico em um número fixo de passado, observando a sequência das leituras. O primeiro modelo desse grupo é o modelo **M-Alta Permanência** que assume como alta a probabilidade de permanência no tópico atual e baixa a probabilidade de mudança de tópico. Os demais modelos desse grupo são todos Markovianos, variando a ordem e a invariância ou não das probabilidades pelo instante avaliado. O valor da ordem denota a quantidade de 'memória' de passado que é avaliada para a predição do próximo tópico. **M-Markov-I** e **M-Markov-I Homogêneo** são modelos de primeira ordem, condicionam o futuro somente no ultimo tópico, o atual. Enquanto o modelo de alta permanência assume valores fixos para as probabilidades, os modelos Markovianos de primeira ordem calculam as probabilidades a partir dos dados. Além dos modelos de primeira ordem temos os modelos de segunda ordem: **M-Markov-II** e **M-Markov-II Homogêneo** que condicionam o futuro nos dois últimos tópicos; de terceira ordem: **M-Markov-III Homogêneo** que condiciona o futuro nos três últimos tópicos; e de quarta ordem: **M-Markov-IV Homogêneo** que condiciona o futuro nos quatro últimos tópicos. Os modelos ditos homogêneos adotam probabilidades invariáveis com o número da leitura. Os modelos Markovianos de ordem superior observam a ordem em que os tópico ocorrem, gerando probabilidades diferentes dependente da ordem dos tópicos no passado. Por exemplo, a probabilidade de um determinado tópico dado o passado em dois instantes $\mathbb{P}(T_n = l \mid T_{n-2} = x, T_{n-1} = y)$ tende a ser diferente da probabilidade desse mesmo tópico se o passado tiver os tópicos em ordem diferente: $\mathbb{P}(T_n = l \mid T_{n-2} = y, T_{n-1} = x)$.

Os modelos de preferência revelada são os modelos que utilizam de uma função focada em um tópico-alvo para predizer a probabilidade desse ser o tópico da próxima

leitura. A função coleta informação de todo passado para esse tópico-alvo, retornando um valor. Esse valor retornado nos fornece a probabilidade do tópico, pela tabela de verossimilhança ajustada ao modelo. O primeiro modelo desse grupo, o **M-Histórico de Visitas** utiliza uma função que contabiliza quantas vezes o tópico-alvo já foi lido no passado da sessão. Em uma sessão de n leituras, o valor retornado s varia de $0 \ldots n$, quanto maior o n mais leituras desse tópico podem ter sido feitas. Assim, a probabilidade do próximo tópico depende do instante atual n e do tópico-alvo. Observe: dados $n = 3$ e o histórico de visitas de um tópico-alvo l ser $s = 2$, temos uma certa probabilidade $\mathbb{P}(T_4 = l \mid s = 2)$. Se o instante avaliado for $n = 9$ e o histórico de vistas desse tópico não mudar, teremos a probabilidade $\mathbb{P}(T_{10} = l \mid s = 2)$. Essa última probabilidade deve ser menor que a primeira, pois se passaram 6 leituras e o tópico-alvo l não foi lido mais vezes. Essa é uma suposição simples. Na verdade, são os dados que nos falarão se a probabilidade é maior, menor ou a mesma ao decorrer da sessão. Os valores dessas probabilidades estão presentes nas matrizes de transição ajustadas previamente com os dados das sessões de treino. O modelo **M-Últimas M Visitas** é variação do modelo função de histórico de visitas. A única diferença é que o histórico analisado é condicionado às últimas m leituras. Outro modelo desse grupo é o modelo **M-Duração no Estado**. Ele utiliza uma função de duração no tópico atual, que contabiliza há quanto tempo o tópico alvo vem sendo lido. Esse modelo também possui variação: **M-Duração no Estado Últimos M Artigos**. A diferença entre esses dois modelos é que o último avalia somente os últimos m artigos da sessão. No modelo de duração, quando o tópico alvo não é igual ao tópico atual, a duração resulta em valor zero. Já o modelo **M-Duração da Última Visita** observa a duração da última visita ao tópico alvo, possibilitando valores diferentes de zero quando o tópico alvo não é igual ao último tópico. O último modelo desse grupo é o **M-Leituras Pós Saída**, esse modelo utiliza a função que contabiliza quantas leituras ocorreram depois da saída do tópico alvo. Caso o tópico nunca tenha sido visitado anteriormente, a função assume valor infinito, e o modelo contabiliza a probabilidade de aparição desse novo tópico.

O próximo grupo é formado pelos modelos que classificam a sessão em intervalos de permanência e mudança. Todos os modelos adotam que o tempo de permanência segue uma distribuição geométrica. Isto justifica a escolha do nome do grupo como sendo modelos de permanência geométrica. Os modelos desse grupo se diferenciam basicamente pelo função de transição de tópico. Por exemplo, **M-PG A** adota uma função de renormalização das probabilidades de entrada nos tópicos simplesmente removendo a probabilidade do tópico do qual se está saindo. Já o modelo **M-PG B** altera as probabilidades a partir do histórico de quantos estados foram vistos e não

estavam no tópico alvo. Todos os tópicos recebem ajustes e é feita uma normalização removendo o tópico do último estado. Já no modelo **M-PG C**, essa normalização é feita como um bônus dado para o próximo tópico. O bônus é um parâmetro específico para cada tópico.

O último grupo de modelos é o dos modelos de vantagem cumulativa. Nesse grupo, os modelos alteram as probabilidades iniciais a partir de cada escolha do usuário de forma permanente. O tópico que o usuário lê ganha um acréscimo na sua probabilidade, um aumento aditivo (**M-Vantagem Cumulativa A**) ou multiplicativo (**M-Vantagem Cumulativa B**). O incremento desses modelos são parâmetros de entrada e assumem valores para cada tópico em separado. Desse modo, pode-se dar mais incremento para um tópico do que para outro tópico quando visitado.

2.12 Avaliação dos Modelos

Duas métricas foram usadas para a comparação entre os modelos: o critério de informação de Akaike e o escore de predição de Brier. A seguir apresentamos brevemente as métricas. Os resultados obtidos em cada métrica estão presente no capítulo 5.

Critério de Informação de Akaike

A famosa frase de John von Neumann presente na epígrafe deste capítulo nos lembra que não é surpreendente que um modelo complexo, com muitos parâmetros, ajuste-se bem a um conjunto de dados. Com um número grande de parâmetros, um modelo pode se adaptar a qualquer conjunto de dados. Entretanto, um ajuste excessivamente grande a um conjunto de dados tipicamente implica em menos capacidade preditiva para novos dados. Diz-se que o modelo não generaliza muito bem. O **critério de informação de Akaike (AIC)** avalia dois aspectos de um dado modelo: seu ajuste aos dados e sua complexidade. O AIC valoriza modelos com bom ajuste mas penaliza aqueles com muitos parâmetros.

Seja $L(M)$ o valor da máxima verossimilhança de um modelo arbitrário M e $df(M)$ o número de parâmetros independentes estimados no modelo, também chamado de graus de liberdade. O valor do critério de informação de Akaike (**AIC**) é dado por:

$$AIC(M) = 2 \ln L(M) - 2df(M) \tag{2.27}$$

Num conjunto de modelos, $M_1, \ldots, M_k$, deve-se preferir aquele com maior valor do AIC. O AIC de um modelo leva em conta dois aspectos. De um lado, queremos

que a log-verossimilhança $\ln L(M)$ seja grande. Entretanto, para evitar *over-fitting*, a medida penaliza os modelos muito complexos, com muitos parâmetros. A maneira exata como estes dois aspectos devem ser combinados é obtida através da teoria de informação por Akaike (Akaike [1974]) e levou à sua fórmula (2.27).

Escore de Brier

Dado um modelo M qualquer em um instante n, e o conjunto de rótulos dos tópicos $\mathcal{L} = \{1, 2, .., L\}$, temos L probabilidades, uma para cada possível tópico:

$$\begin{aligned} \mathbb{P}^M(T_n = 1|T_1, ..., T_{n-1}) &= p_1^{n,M} \\ \mathbb{P}^M(T_n = 2|T_1, ..., T_{n-1}) &= p_2^{n,M} \\ \vdots \qquad & \qquad \vdots \\ \mathbb{P}^M(T_n = L|T_1, ..., T_{n-1}) &= p_L^{n,M} \end{aligned} \tag{2.28}$$

Como escolher qual o tópico para recomendar? O que tiver a maior probabilidade? O que atingir certo nível mínimo de probabilidade tal como 51%, 80% ou 100%? Ou, os melhores k tópicos ranqueados? Uma solução encontrada para avaliar a qualidade de previsões foi a definição de regras de pontuação (*scoring rules*). Essas regras procuram mensurar o quanto as probabilidades associadas com as previsões estão nos direcionando para o tópico certo. Dentre estas regras decidimos usar uma das mais populares, o escore de Brier (Ferri et al. [2011]; Hernández-Orallo et al. [2012]), que descrevemos a seguir.

Suponha que o conjunto de tópicos é constituído por 4 assuntos: $\mathcal{L} = \{1, 2, 3, 4\}$. Considere uma sessão de um usuário que já leu dois artigos, sendo o primeiro artigo do tópico $T_1 = 1$ e o seguinte, do tópico $T_2 = 4$. No instante $n = 3$, três modelos hipotéticos (X, Y e Z) nos forneceram as seguintes probabilidades:

Modelo M	$p_1^{3,M}$	$p_2^{3,M}$	$p_3^{3,M}$	$p_4^{3,M}$
X	0,09	0,005	0,005	0,90
Y	0,45	0,02	0,03	0,50
Z	0,30	0,20	0,15	0,35

Tabela 2.1: Exemplo de previsão de 3 modelos fictícios para um conjunto de 4 tópicos possíveis, dada a sessão prévia $S = (1, 4)$.

Os modelos forneceram probabilidades diferentes para a escolha do tópico T_3 mas todos concordam que o tópico mais provável é 4. No caso do usuário realmente ler um artigo do tópico 4, todos os modelos terão acertado em sua previsão. Contudo, podemos

falar que o modelo X teve um acerto melhor que os demais por ter a probabilidade do tópico 4 bem alta.

Suponha agora que $T_3 = 1$. Isto é, no instante 3, o usuário lê um artigo do tópico 1. Nenhum modelo colocou esse tópico como o mais provável. Todos erraram. Mas qual dos três modelos errou menos? Esse tópico é o segundo mais provável para todos os modelos, mas há modelos que alocam mais probabilidade a ele que outros. E se o usuário ler um artigo de tópico diferente dos dois mais prováveis? Por exemplo, se $T_3 = 3$, como avaliar a qualidade da predição exibida na Tabela 2.1? Qual dos modelos errou menos nesse caso?

Para responder à pergunta de qual modelo erra menos nas probabilidades de predições foi utilizado o **escore de Brier**. O escore de Brier calcula a diferença média (ao quadrado) entre as probabilidades de previsão e os resultados reais. Segundo a sua fórmula (2.29), podemos calcular o valor do erro de cada modelo somando sobre todas as opções de tópicos a diferença ao quadrado entre a probabilidade predita para o tópico menos o acontecimento real (1, se o tópico foi o escolhido e 0, caso contrário). Para um melhor entendimento do escore de Brier, vamos ilustrar os cálculos.

Nosso conjunto de dados de teste é formado pelas três sessões que descrevemos acima, $S_1 = (1, 4, 4), S_2 = (1, 4, 1), S_3 = (1, 4, 3)$. As probabilidades dos modelos são aquelas apresentadas na tabela (2.1) e o instante avaliado é $n = 3$. O escore de Brier para um modelo M no instante $n = 3$ é dado por:

$$BS(i = 3, M) = \frac{1}{N_i} \sum_{S \in \mathbb{S}; n \geq i} \sum_{l \in \mathcal{L}} (f_M(S, i, l) - o(S, i, l))^2$$

Para o **modelo X**, o escore de Brier é calculado como a seguir. Para S_1, o somatório mais interno fica igual a: $(0,09-0)^2+(0,005-0)^2+(0,005-0)^2+(0,9-1)^2 =$ **0,01815**. Para S_2, o somatório mais interno fica igual a: $(0,09-1)^2+(0,005-0)^2+(0,005-0)^2+(0,9-0)^2 =$ **1,63815**. Para S_3, o somatório mais interno fica igual a: $(0,09-0)^2+(0,005-0)^2+(0,005-1)^2+(0,9-0)^2 =$ **1,80815**. Somando as três parcelas e dividindo por $N_3 = 3$ temos: $BS(i = 3, M = 1) \cong$ **1,16**.

Agora, o cálculo do escore de Brier para o **modelo Y**. Para S_1, o somatório mais interno fica igual a: $(0,45-0)^2+(0,02-0)^2+(0,03-0)^2+(0,5-1)^2 =$ **0,4538**. Para S_2, o somatório mais interno fica igual a: $(0,45-1)^2+(0,02-0)^2+(0,03-0)^2+(0,5-0)^2 =$ **0,5531**. Para S_3, o somatório mais interno fica igual a: $(0,45-0)^2+(0,02-0)^2+(0,03-1)^2+(0,5-0)^2 =$ **1,3938**. Somando as três parcelas e dividindo por $N_3 = 3$ temos: $BS(i = 3, M = 2) \cong$ **0,8**.

Finalmente, o resultado do escore de Brier para o **modelo Z**. Para S_1, o somatório

mais interno fica igual a: $(0,3-0)^2+(0,2-0)^2+(0,15-0)^2+(0,35-1)^2=$ **0,575**. Para S_2, o somatório mais interno fica igual a: $(0,3-0)^2+(0,2-0)^2+(0,15-0)^2+(0,35-1)^2=$ **0,675**. Para S_3, o somatório mais interno fica igual a: $(0,3-0)^2+(0,2-0)^2+(0,15-0)^2+(0,35-1)^2=$ **0,975**. Somando as três parcelas e dividindo por $N_3=3$ temos: $BS(i=3,M=3)\cong$ **0,74**.

Quando as probabilidades estão bem ajustadas aos dados e os modelos não são triviais, o escore de Brier tende a variar entre 0 e 1. Os limites teóricos para o escore de Brier são 0 e 2.

Observando o resultado final do escore de Brier, podemos falar que o modelo Z é o que errou menos. Agora, observemos sessão por sessão. Para o caso de S_1, o modelo que errou menos foi realmente o modelo X. Em S_2, o usuário lê um artigo do tópico 1 no instante 3 ao invés de ler um artigo do tópico 4 (o mais provável). Neste caso, o modelo que erra menos é o modelo Y. Porém, se o terceiro artigo lido for de um tópico menos provável como o tópico 3 em S_3, temos que o modelo que erra menos é o modelo Z. Observe que cada modelo errou menos que os demais dependendo da sessão de teste. Dependendo da frequência das sessões na base de dados em que um modelo supera os demais, o escore de Brier final nos mostrará em termos gerais qual modelo erra menos.

Como dito anteriormente, o escore de Brier (**BS**) é uma função de pontuação que mede a acurácia das estimativas probabilísticas (Brier [1950]; Murphy [1973]). A sua versão multi categoria foi a utilizada para medir as predições de próximo tópico.

Sejam $\mathcal{S}$ a coleção de sessões de teste, n o tamanho de uma sessão $S\in\mathcal{S}$, N_i o total de sessões da coleção de teste que contêm no mínimo i instantes, e $\mathcal{L}$ o conjunto de tópicos. O escore de Brier para o modelo M no instante i é dado por:

$$BS(i,M)=\frac{1}{N_i}\sum_{S\in\mathcal{S};n\geq i}\sum_{l\in\mathcal{L}}(f_M(S,i,l)-o(S,i,l))^2 \tag{2.29}$$

onde, $o(S,i,l)$ é a informação binária de que o i-ésimo artigo da sessão S é ou não do tópico l, e $f_M(S,i,l)$ é a probabilidade do próximo tópico i ser igual a l pelo modelo M. Quanto menor o escore de Brier calculado melhores são as predições. O escore de Brier é uma das medidas que atendem a requisitos formais para avaliação de predições, sendo um dos exemplos mais famosos de uma *proper scoring rule* (Gneiting & Raftery [2007]).

Capítulo 3

Dados de Jornais Online

Neste trabalho foram utilizados dados reais de 2 jornais online que serão identificados como **Jornal Online A** e **Jornal Online B**. Os dados foram gentilmente cedidos por uma empresa de recomendação com a autorização dos jornais online. Nas figuras e nos gráficos a seguir há o seguinte padrão de cor para facilitar o reconhecimento do jornal/base a que se referem: **verde para Jornal Online A** e **azul para Jornal Online B**.

3.1 Coleta de Dados

Foram coletados dados referentes a leitura de usuários dos jornais no período de 01/02/2015 a 31/03/2015. O resultado foi a obtenção de mais de 80 Gb de dados em formato Json, com as seguintes informações: item_id (identificador único do artigo lido), timestamp (tempo em milissegundo do acesso ao artigo), user_id (hash que anonimiza o usuário), recs (lista dos item_ids recomendados presentes na página do artigo lido), click (informação se a visualização veio de um click em link recomendado), dentre outras. Havia linhas nulas, mal formatadas e duplicadas na base de dados. Elas foram removidas.

Os jornais possuem públicos-alvo diferentes. Enquanto um jornal foca em entretenimento (**Jornal Online A**), o outro foca em notícias mais sérias (**Jornal Online B**). Mesmo assim, os dois jornais possuem altos índices de leituras, feitas por usuários fiéis e visitantes ocasionais. Na Figura 3.1 é possível comparar os acessos aos dois jornais ocorridos no período da coleta de dados. O **Jornal Online A** recebe normalmente mais acessos que o **Jornal Online B**. Não há um padrão regular de acesso semanal exceto aos domingos, quando os dois jornais recebem menos acessos do que nos demais dias.

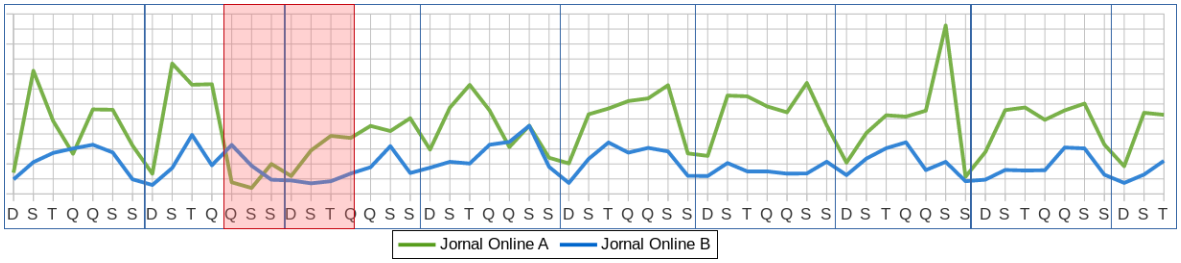


Figura 3.1: Volume de visitação dos jornais nos meses de fevereiro e março de 2015. As letras D, S, T, Q, Q, S, e S são relacionados aos dias da semana começando no Domingo. O primeiro D corresponde ao dia 01/02/15, e o último T ao dia 31/03/15.

No período analisado houve o Carnaval de 2015. Nesses dias ambos os jornais possuem os menores índices de acessos. O carnaval aconteceu entre o fim da segunda semana e o início da terceira. Na Figura 3.1 a caixa em vermelho ressalta os dias da festividade.

Uma segunda parte dos dados contendo as informações dos artigos também foi coletada. Temos mais de 300 mil artigos do **Jornal Online A** e quase 450 mil artigos do **Jornal Online B**. Para cada artigo, temos as seguintes informações: item_id (id do artigo), groups (grupo em que o artigo se enquadra no jornal), html (texto contendo o caminho dentro do site para o artigo), title (titulo do artigo), author (nome do autor), body (o corpo do artigo).

3.2 Sessões de Leituras

Com a junção das duas partes dos dados geramos as sessões de leitura. Uma sessão de leitura é formada pelas leituras consecutivas de artigos diferentes por um mesmo usuário. Dois artigos sucessivos devem estar espaçados por 30 minutos no máximo. Caso o tempo entre duas leituras ultrapasse esse valor, consideramos que a sessão de leituras estava encerrada e outra sessão do mesmo usuário teve início. Quando há leituras repetidas de um mesmo artigo em sequência, somente a primeira leitura é considerada. Para cada leitura, somente as informações de item_id e timestamp são mantidas. Uma sessão apresenta além das informações de leituras em ordem cronológica, o identificador do usuário que gerou a sessão.

Dentre todas as sessões resultantes, somente as *sessões relevantes* foram selecionadas. Sessões relevantes são aquelas que possuem no mínimo 2 e no máximo 90 artigos lidos e que possuem uma duração total inferior a 90 minutos. Duração total é a diferença entre o tempo de acesso do primeiro e do último artigo da sessão. As sessões que não se enquadram nesses critérios foram desconsideradas pois sessões compostas

da leitura de um único artigo, chamadas de sessões unas, não contêm transição entre artigos, não contendo informação útil para o objetivo deste trabalho. Além disso, as sessões com mais de 90 artigos ou mais de 90 minutos de duração total são muito longas para leituras humanas usuais. Um humano que lê mais de 90 artigos em 90 minutos, leu mais de um artigo por minuto, em média. Se a duração total for menor, isso implica em leituras ainda mais rápidas, sendo potencialmente uma sessão de leitura não-humana, tal como um acesso por computador *bot* ao jornal online. O percentual de filtragem dessa etapa está na Tabela 3.1. Apesar do percentual de sessões relevantes ser menor que o percentual de sessões unas, ainda é alto o volume de dados restante após eliminar as sessões de um único artigo.

	Sessões	*Sessões Grandes*	Sessões Relevantes	
Base	*Unas*	*e/ou Longas*	*Percentual*	*No. Absoluto*
Jornal Online A	75,87%	0,01%	24,12%	17.558.933
Jornal Online B	85,83%	0,01%	14,16%	5.632.371

Tabela 3.1: Resumo da filtragem de sessões relevantes. Sessões unas são aquelas com a leitura de um único artigo. As sessões grandes/longas são as de mais de 90 artigos e ou com duração total acima de 90 minutos.

3.3 Tamanho das Sessões

Ao longo da leitura de um artigo ou ao seu término, os usuários podem interessar-se por outros artigos disponíveis, e assim podem clicar nos links disponíveis, continuando no mesmo assunto ou transitando para artigo de outro assunto. Essas leituras sucessivas de artigos geram as sessões que esta dissertação utiliza como principal base de dados.

Nos dados que coletamos, há muitas sessões compostas por mais de um artigo. O tamanho de uma sessão é definida como a quantidade de artigos lidos, e esses tamanhos são bem variados. Os gráficos da Figura 3.2 ilustram a distribuição dos tamanhos das sessões relevantes. O eixo horizontal mostra o tamanho das sessões e o eixo vertical mostra a quantidade de sessões. Nota-se que a distribuição dos dois jornais está concentrada nas sessões curtas, com poucas leituras.

Embora pequenas, as contagens na cauda superior da distribuição não são zeros. Para visualizar melhor o decaimento dessa cauda, produzimos os gráficos da Figura 3.3 onde o eixo vertical apresenta as contagens em escala logarítmica de base 10. É possível observar mais claramente agora que os usuários do **Jornal Online A** tendem a ler mais artigos do que os usuários do **Jornal Online B**, gerando um número maior de sessões longas.

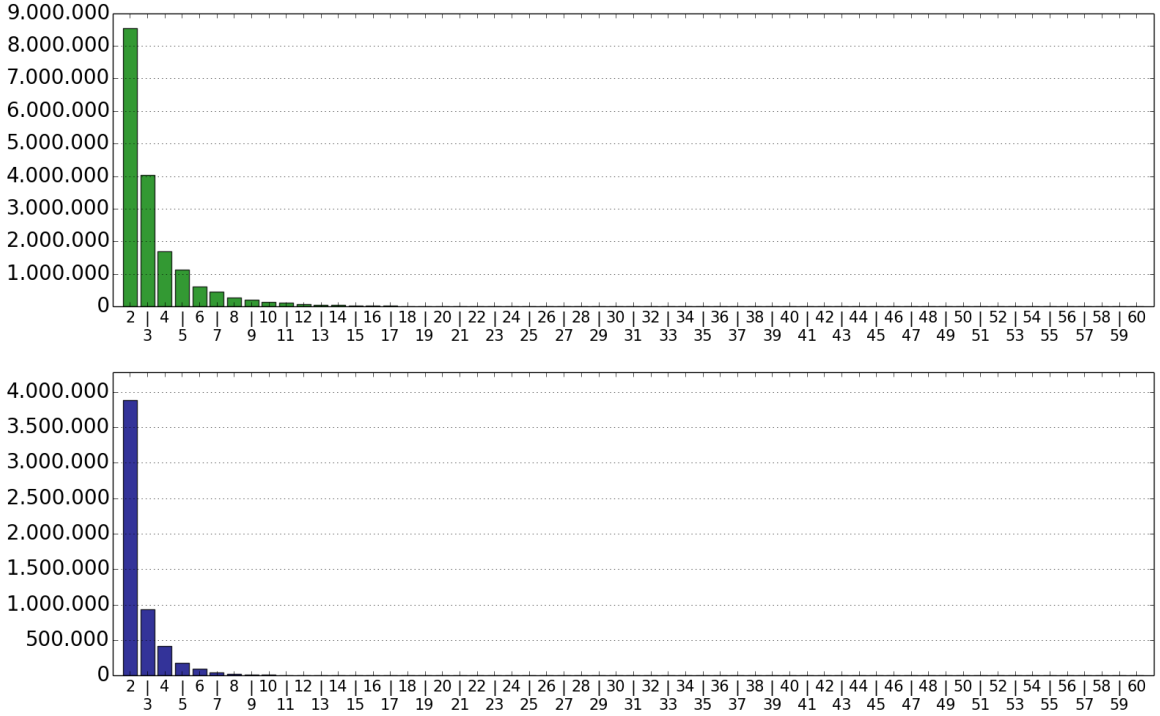


Figura 3.2: Distribuição dos tamanho das sessões relevantes do **Jornal Online A** e do **Jornal Online B**. O eixo X mostra o tamanho das sessões e o eixo Y mostra a quantidade de sessões.

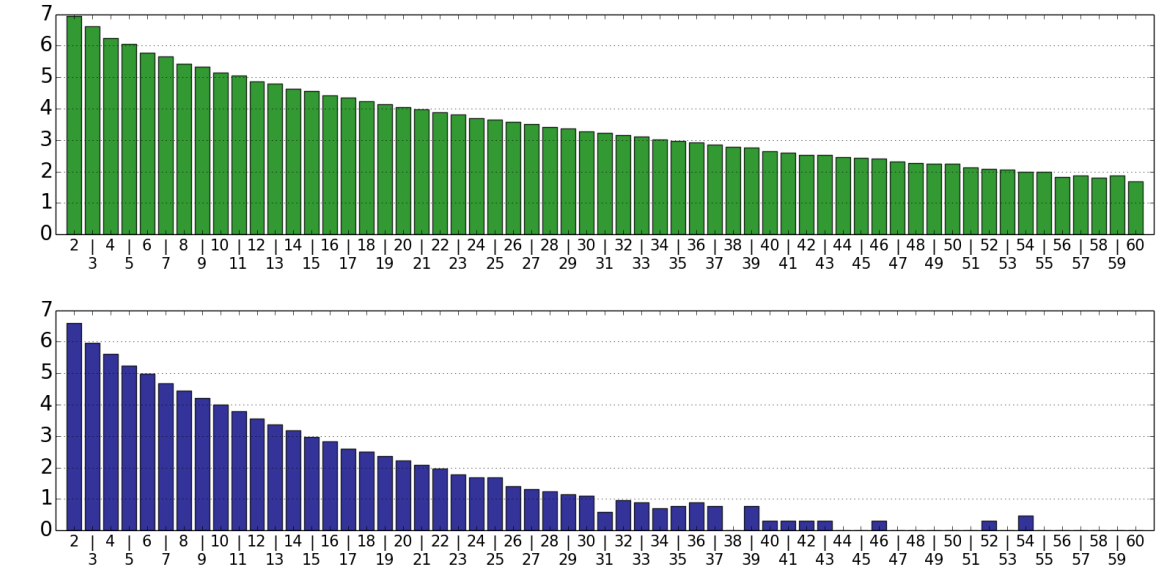


Figura 3.3: Distribuição dos tamanho das sessões relevantes em escala logarítmica do **Jornal Online A** e do **Jornal Online B**. O eixo X mostra o tamanho das sessões e o eixo Y mostra a quantidade de sessões, em escala logarítmica de base 10.

Além disso, esse gráfico possui um decaimento aproximadamente linear. Isso leva à conjectura de que as duas distribuições possuem cauda pesada, ou seja, elas são do tipo *power laws* (Newman [2005]). Nesse tipo de distribuição, se o tamanho de uma sessão é representado pela variável aleatória N, temos $\mathbb{P}(N = n) \propto n^{\alpha}$ onde $\alpha > 1$. Isso implica que $\log(\mathbb{P}(N = n))$ é uma função linear de $\log(n)$, como podemos ver na Figura 3.4. Nesta figura há dois gráficos, um para cada base de dados, que nos mostram os valores de tamanho de sessão versus a contagem de casos, ambos em escala logarítmica de base 10. Podemos ver que o decrescimento é praticamente linear, como suspeitamos anteriormente. Newman [2005] demonstra que podemos estimar o parâmetro α por máxima verossimilhança:

$$\alpha = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1}$$

onde as quantidades $x_i,\ i = 1, \ldots, n$ são os valores medidos da variável resposta x e x_{min} é o menor valor de x. Estimamos por essa máxima verossimilhança os valores do parâmetro α para cada jornal, encontrando $\alpha_A = 1.206$ e $\alpha_B = 1.264$.

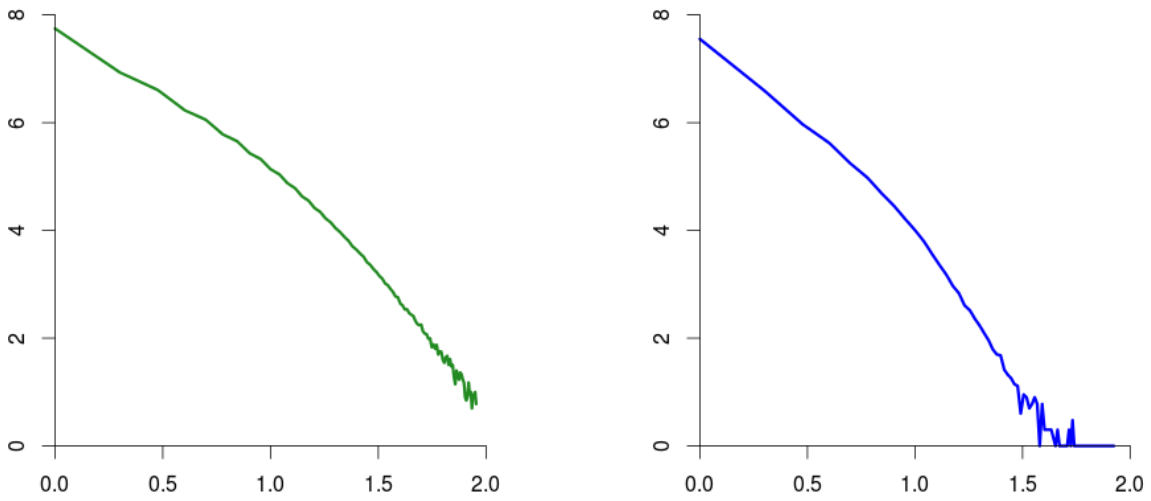


Figura 3.4: Distribuição dos tamanho das sessões relevantes. Ambos eixos em escala logarítmica de base 10.

3.4 Acessos dos Usuários

Na base de dados do **Jornal Online A**, depois de todo o tratamento, foram identificados 8.886.039 usuários distintos. A grande maioria (73%) gerou apenas uma única sessão de leitura, mas houve usuários que chegaram a gerar até 281 sessões de leitura

nesses 2 meses avaliados. Na base do **Jornal Online B**, também depois do tratamento inicial, encontramos 3.846.728 id's distintos de usuários que geraram de uma a 191 sessões. Novamente, a maioria deles (80%) criou apenas uma única sessão durante o período de análise. Como é pequena a proporção de usuários que aparecem com sessões múltiplas na base, nós decidimos não estudar a variação do comportamento de um mesmo usuário na base. Sentimos que não teríamos uma base grande para generalizar para a população de usuários. Assim, nossa análises e modelos ignoram que um mesmo usuário pode aparecer mais de uma vez na base.

Na Tabela 3.2 temos um resumo sobre a frequência de sessões geradas pelos usuários. Os leitores do **Jornal Online A** são um pouco mais fiéis que os leitores do **Jornal Online B**. 26,6% dos usuários do primeiro jornal geraram mais de uma sessão de leitura, contra 20,0% dos leitores do segundo jornal online. Em números absolutos são 2.363.686 contra 769.346 usuários, respectivamente.

Frequências Acumulada dos usuários		*Máximo de sessões*
Jornal Online A	**Jornal Online B**	*distintas geradas*
73,4%	80,0%	1
84,7%	91,4%	2
89,6%	95,4%	3
92,4%	97,2%	4
94,2%	98,2%	5
95,4%	98,7%	6
96,3%	99,0%	7
97,0%	99,2%	8
97,5%	99,4%	9
97,9%	99,5%	10
-	100%	191
100%	-	281

Tabela 3.2: Frequência acumulada dos usuários pelo máximo de sessões geradas.

Dos 8.886.039 usuários do **Jornal Online A**, 26,6% deles geraram 11.036.580 sessões (62,9% do total de sessões). No caso do **Jornal Online B**, 769.346 usuários do total de 3.846.728 (ou 20% dos usuários) geraram 2.554.989 sessões (45,4% do total de sessões). Resumindo, a porcentagem de usuários que gerou mais de uma sessão é maior no **Jornal Online A** e esse grupo de usuários gerou mais sessões que o grupo correspondente do outro jornal.

3.5 Tópicos dos Artigos

Cada artigo possui a informação do grupo de notícias ao qual ele pertence. Essa denominação não é única pois um artigo pode pertencer a mais de um grupo. Além disso, a classificação pode conter erros (há erros de digitação e de redação). A informação do grupo de notícias de cada artigo foi comparada com outras duas informações: o html e o título da notícia. A informação mais significativa dentre as disponíveis foi escolhida manualmente para designar o pré-tópico do artigo. Em seguida, os pré-tópicos gerados foram comparados e agrupados por semelhança de conteúdo. Essa comparação e agrupamento foi feita nas duas bases de dados em separado.

No final, obtivemos 10 tópicos para cada base de dados. Os rótulos dos tópicos foram anonimizados e serão apresentados pela letra do jornal seguida por um número de 0 a 9. Assim, os tópicos do **Jornal Online A** são `A0`, `A1`, `A2`, `A3`, `A4`, `A5`, `A6`, `A7`, `A8` e `A9`; e os tópicos do **Jornal Online B** são `B0`, `B1`, `B2`, `B3`, `B4`, `B5`, `B6`, `B7`, `B8` e `B9`. Um tópico `Ax` do primeiro jornal não tem necessariamente relação com o tópico `Bx` do outro jornal. Alguns tópicos existem em ambos jornais, e outros são específicos de um jornal.

3.6 Frequência dos Tópicos, Postagem e Acesso

A Figura 3.5 compara os tópicos dos artigos publicados e dos artigos lidos pelos usuários nas sessões do **Jornal Online A**. Os tópicos mais publicados são `A4` e `A7` compondo mais do 60% do volume total dos artigos postados. Entretanto, esse grande volume de artigos postados é pouco lido. Menos de 25% dos artigos lidos vem desses dois tópicos mais comuns no **Jornal Online A**. O tópico mais lido desse jornal é o `A5`, que retém praticamente metade de todos os acessos. Esses artigos buscados avidamente representam apenas 3.4% do número total de artigos exibidos pelos jornal.

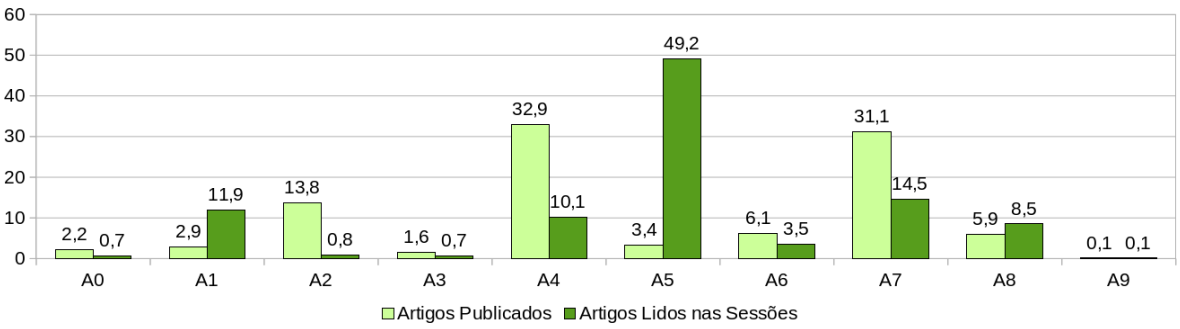


Figura 3.5: Comparativo das frequências de leitura e publicação do **Jornal Online A**.

O contexto no **Jornal Online B** é diferente, como mostra a Figura 3.6. Predomina a publicação de artigos de um tópico, enquanto as leituras são mais homogêneas. O conteúdo postado do Tópico `B7` representa mais da metade de todos os artigos postado no **Jornal Online B**. Os tópicos `B1`, `B3`, `B8`, `B4`, `B6` e `B7` são os mais acessados nessa ordem e todos possuem valores de leituras entre 25% e 10%. Novamente, os tópicos mais lidos, `B1` e `B3` não são os mais publicados, mesmo comportamento encontrado no **Jornal Online A**.

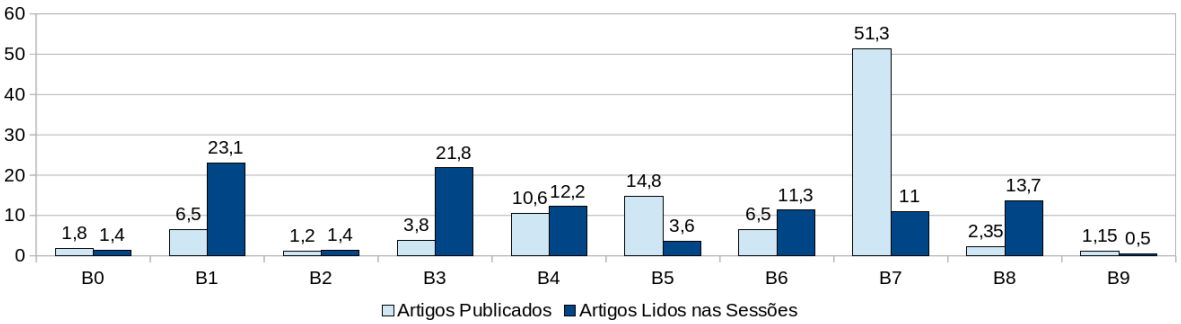


Figura 3.6: Comparativo das frequências de leitura e publicação do **Jornal Online B**.

3.7 Quantidade de Tópicos por Sessão

O total de sessões relevantes é da ordem de vários milhões, como vimos na Tabela 3.1. Há sessões bem longas, logo os usuários podem transitar por mais de um tópico dentro das sessões. A Tabela 3.3 mostra quantos tópicos distintos uma sessão possui. Pelos percentuais podemos ver que os leitores do **Jornal Online B** leem mais diversificadamente que os leitores do **Jornal Online A**. No **Jornal Online A** as sessões são, em sua maioria, de um único tópico, enquanto que, no **Jornal Online B** a maioria das sessões são compostas por dois tópicos.

Sessões	**Jornal Online A**	**Jornal Online B**
de um único tópico	48,3%	37,2%
de dois tópicos	41,8%	52,4%
de três tópicos	7,7%	8,2%
de quatro tópicos	1,7%	1,7%
de cinco ou mais tópicos	0,5%	0,5%

Tabela 3.3: Resumos dos dados pela quantidade de tópicos diferentes em cada sessão.

Em geral, há poucos tópicos distintos nas sessões para ambos os jornais. Praticamente 90% dos casos são de dois tópicos, no máximo. Essa concentração em poucos

tópicos é causada pelo grande número de sessões formadas por 3 artigos ou menos. Numa sessão composta por 2 artigos, só é possível ter um ou dois tópicos distintos. Com 3 artigos lidos, o número de tópicos é limitado por 3 também, e assim por diante. No caso do **Jornal Online A** a quantidade de sessões de tamanhos 2 e 3 representam respectivamente 48,6% e 22,9% de todas as sessões. E no caso do **Jornal Online B** esses valores são 69,0% e 16,6%. As sessões maiores são mais diversas mas não representam a maioria das sessões como veremos a seguir.

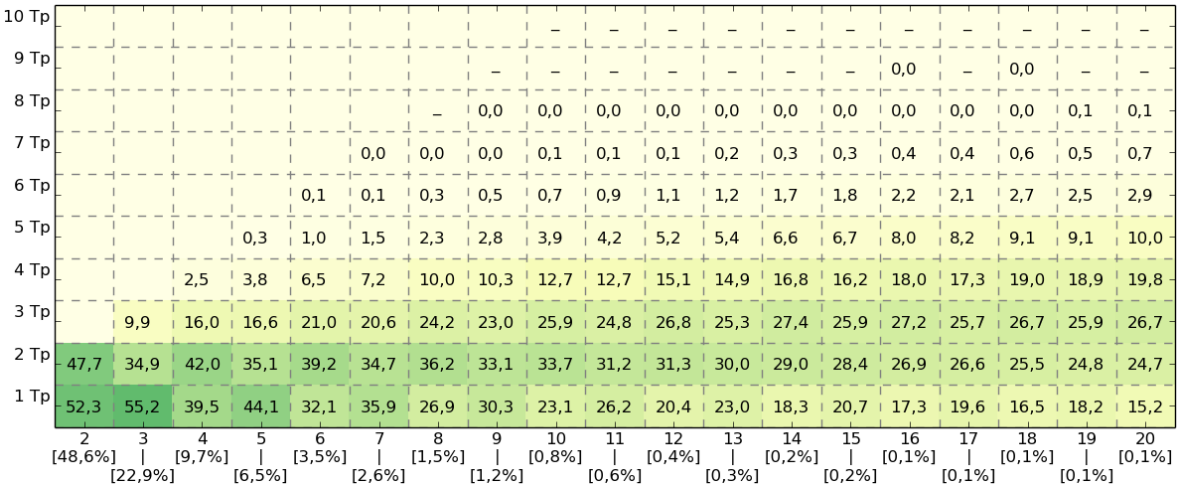


Figura 3.7: Distribuição da quantidade de tópicos pelo tamanho das sessões no **Jornal Online A**.

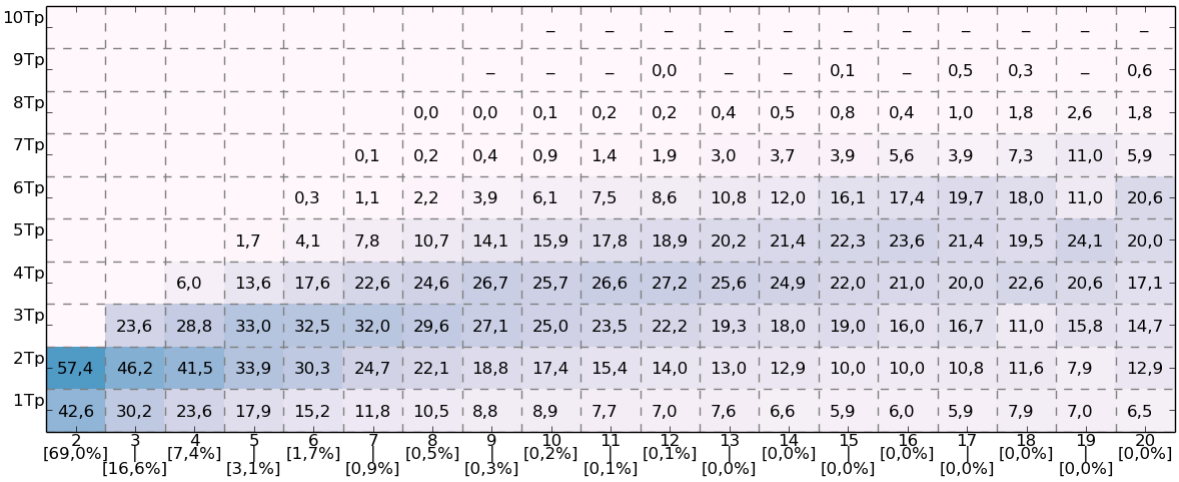


Figura 3.8: Distribuição da quantidade de tópicos pelo tamanho das sessões no **Jornal Online B**.

Nas Figuras 3.7 e 3.8 temos a distribuição da quantidade de tópicos distintos classificada pelo tamanho das sessões. No eixo das ordenadas está o tamanho das sessões, limitado em até 20 leituras. Entre colchetes, mostramos o percentual de sessões

que são daquele tamanho. No eixo das abcissas temos os números de tópicos distintos. No corpo do gráfico, mostramos a porcentagem somando 100% para cada tamanho de sessão. Quanto mais forte a cor mais alta a porcentagem.

No gráfico do **Jornal Online A** a mancha de cor mais forte que representa as maiores porcentagem se mantém entre 2, 3 e 4 tópicos. Já no gráfico do **Jornal Online B** a mancha de cor mais forte cresce mais rápido, atingindo 4, 5 e 6 tópicos distintos. Além disso, a concentração de cor se dilui à medida que cresce o tamanho da sessão indicando um aumento também do desvio padrão. Esse comportamento condiz com aquele visto anteriormente: os usuários do **Jornal Online A** tendem a ler de forma menos diversificada do que os usuários do **Jornal Online B**.

No próximo capítulo, apresentaremos os resultados de diversas análises exploratórias que fizemos nos dados. Essas análises nos mostraram mais do que estatísticas das bases de dados de jornais onlines, como os resultados deste capítulo, elas nos forneceram informações sobre o comportamento de leitura dos usuários.

Capítulo 4

Análise Exploratória dos Dados

Neste capítulo, vamos explorar características mais complexas das duas bases de dados de notícias online. Inicialmente, vamos analisar o tempo gasto lendo cada artigo numa sessão e como esse tempo está associado com o tamanho da sessão de leituras. Em seguida, analisamos como varia a distribuição da popularidade dos tópicos com a ordem de leitura. Queremos investigar se tópicos muito frequentes entre os primeiros artigos permanecem populares depois que vários artigos foram lidos numa sessão. Analisamos a seguir os padrões de transição entre os tópicos. Dado que um artigo de certo tópico foi lido, obtemos quais os tópicos mais prováveis da próxima leitura. Após analisarmos a transição, estudamos o tempo de permanência num dado tópico. A duração num dado tópico pode variar por tópico e pode depender do momento de entrada no tópico. Investigamos se a permanência num tópico pode ser maior no início da sessão do que no meio da sessão, quando uma eventual entrada no tópico ocorre após vários artigos lidos. Em seguida, analisamos a história dos tópicos identificando os padrões mais comuns das sequências de tópicos ignorando quanto tempo o usuário permanece no tópico. Analisamos esses padrões levando em conta também a duração em cada tópico.

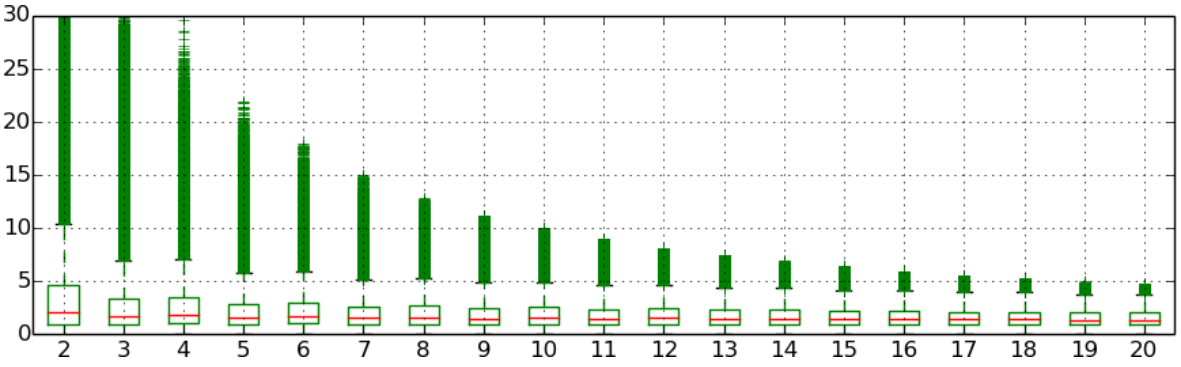
4.1 Intervalo de Leitura

Vamos analisar o tempo gasto na leitura dos artigos. Em particular, estamos interessados em saber como o tempo gasto em cada artigo está associado com o tamanho da sessão de leitura. Queremos saber se existe algum mecanismo de compensação tal que o maior tempo gasto numa sessão longa seja parcialmente descontado por um tempo mais curto gasto na leitura de cada artigo individual.

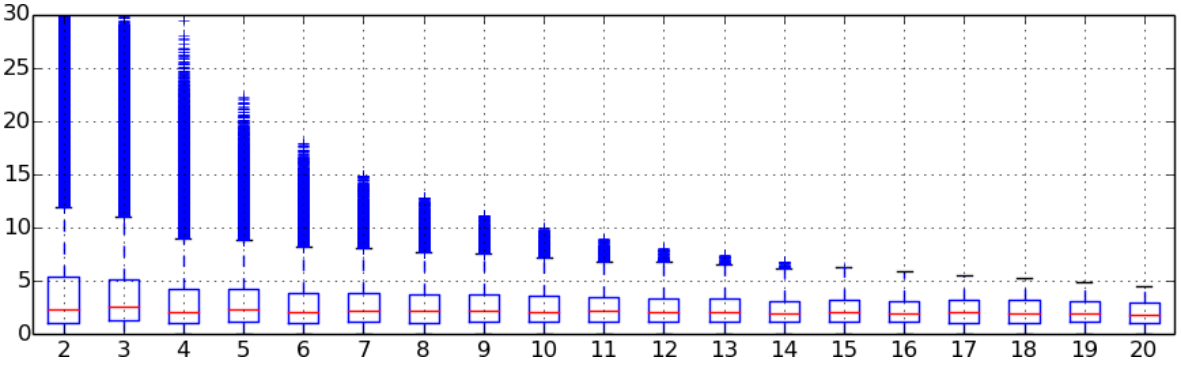
Cada leitura gerada pelos usuários possui a informação do tempo em que foi iniciada. Seja τ_i o tempo de início da i-ésima leitura em uma sessão. Defina o i-ésimo

intervalo de leitura $I_i = \tau_{i+1} - \tau_i$. Assumimos que o tempo de início de leitura do próximo artigo coincide com o tempo de término da leitura do artigo anterior. Numa sessão com n artigos lidos, temos $n-1$ valores $I_1, \ldots, I_{n-1}$. Observe que o tempo de leitura do último artigo não está disponível.

Para cada sessão com $n \geq 2$ artigos, calculamos o valor médio $\bar{I} = (I_1 + \ldots + I_{n-1})/(n-1)$. A Figura 4.1 mostra a distribuição desse tempo médio por artigo versus o tamanho da sessão para cada uma das duas bases. Para cada tamanho de sessão, representamos a distribuição usando um *boxplot*. A caixa de um boxplot delimita 50% dos dados centrais, os dados compreendidos do primeiro ao terceiro quartil. A linha central em vermelho mostra a mediana dos dados. Os fios (ou bigodes) que saem da caixa para cima e para baixo demonstram onde normalmente os dados estão compreendidos. O comprimento de cada bigode é calculado como 1,5 vezes a distância interquartil (altura da caixa). Contudo, quase sempre há dados atípicos, ou *outliers*. Eles são mostrados como pequenos traços horizontais. Esses são os elementos que compõem os boxplots, que estão nas Figuras a seguir.



(a) **Jornal Online A**

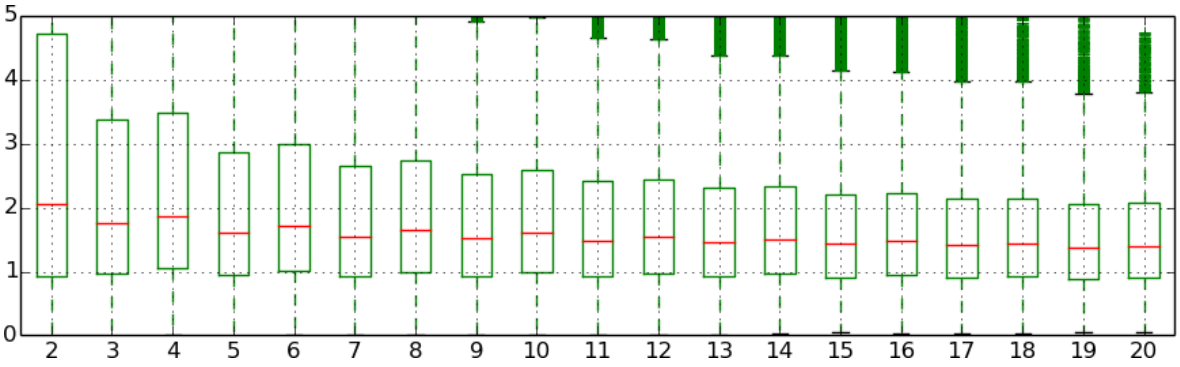


(b) **Jornal Online B**

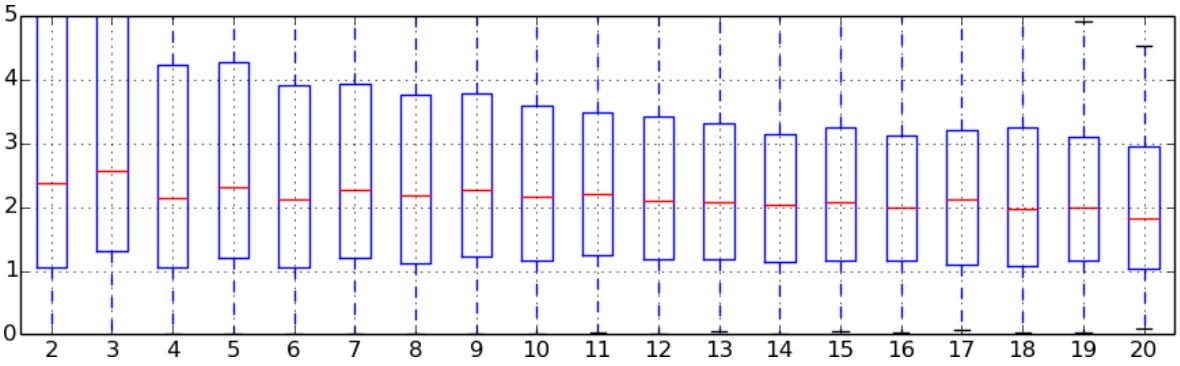
Figura 4.1: Distribuição do tempo médio de leitura por tamanho de sessão. No eixo vertical temos o tempo em minutos. No eixo horizontal, o tamanho das sessões.

Pela Figura 4.1, vemos que a distribuição do tempo médio de leitura não varia muito com o tamanho da sessão. A Figura 4.2 reproduz os gráficos da Figura 4.1 truncando o eixo vertical de forma a mostrar mais claramente o núcleo central da distribuição. O tempo máximo de um intervalo de leitura é 30 minutos, por definição da regra que define uma sessão. No gráficos da Figura 4.1, as partes mais significativas ficam bem abaixo desse valor. Por exemplo, os valores mais significativos das médias estão abaixo de 10 minutos. Logo para uma melhor visualização o valor máximo dos próximos gráficos estará normalmente abaixo de 30 minutos, como na Figura 4.2, que é um zoom dos gráficos da Figura 4.1.

Fica mais claro agora que o tempo médio por artigo diminui ligeiramente com o tamanho da sessão. No gráfico do **Jornal Online A**, vemos que a mediana do tempo médio por artigo varia entre aproximadamente 1,5 e 2 minutos. Já o tempo de leitura médio do **Jornal Online B** é um pouco maior que do **Jornal Online A**. Na maioria dos casos, a mediana do tempo médio de leitura está entre 2 e 2,5 minutos, e a parte superior da caixa é mais alta, ou seja, os percentis 75 são mais altos. Em geral, podemos dizer que os leitores do **Jornal Online B** demoram mais na leitura dos artigos do que os leitores do **Jornal Online A**.



(a) **Jornal Online A**



(b) **Jornal Online B**

Figura 4.2: Distribuição do tempo médio de leitura por tamanho da sessão, ampliada nas parte mais significativa. Eixo X tamanho das sessões e eixo Y tempo em minutos.

4.1.1 Filtragem pelo Intervalo de Leitura

A distribuição do intervalo médio de leitura mostrou que as leituras são, em média, mais rápidas com mais artigos em uma sessão. Procuramos entender melhor se esse efeito é geral em todos os artigos ao longo da sessão ou se é variável, ficando mais curto a medida em que os artigos são lidos.

Nas Figuras 4.3 e 4.4 temos amostras dos intervalos de leitura filtrados pela ordem da leitura e agregados pelo tamanho final da sessão. Por exemplo, na Figura 4.3, podemos ver a distribuição do tempo gasto no segundo artigo lido em sessões de tamanho 2, 3, etc. Nesse gráfico, a maioria das caixas estão compreendidas entre 1 e 3,5 minutos, exceto para as sessões de 2 artigos, onde a caixa do boxlplot está entre 1 e 5 minutos aproximadamente. A partir do segundo gráfico da mesma Figura, não há muita variação nem nas caixas nem nas medianas. Esse comportamento demonstra que a quantidade de artigos que o usuário lê em uma sessão não influencia o tempo mediano dos intervalos de leitura artigo a artigo.

Para os gráficos do **Jornal Online B**, presentes na Figura 4.4, os intervalos de leitura variam quanto ao tamanho da sessão. Observe no gráfico do 1º intervalo de leitura como as caixas crescem e decrescem ligeiramente ao longo do eixo horizontal. Esse comportamento também aparece nos dados do **Jornal Online A**, só que de forma mais sutil. Essa flutuação continua nos outros intervalos, demonstrando uma tendência constante. Como os dados do outro jornal, se comparamos os gráficos intervalo a intervalo vemos que as caixa do boxplot vão diminuindo. A diminuição do **Jornal Online B** é menor que a do **Jornal Online A**.

4.1.2 Filtragem pelo Tamanho da Sessão

Nesta análise, filtramos os tempos de leitura pelo tamanho da sessão e mostramos os valores particionados pelos intervalos existentes nas sessões. As Figuras 4.5, do **Jornal Online A**, e 4.6, do **Jornal Online B**, mostram alguns dos resultados obtidos nesta análise. Esses gráficos nos mostram mais claramente que a primeira leitura é sempre a mais demorada. Os usuários de ambos jornais tendem a gastar mais tempo na leitura do primeiro artigo. O tempo de leitura no **Jornal Online A** decresce constantemente e devagar em todos os gráficos, para todos os tamanhos de sessão. Já no caso do **Jornal Online B**, há decrescimento no tempo ao longo dos intervalos de leitura também, porém alternado com alguns leves incrementos. Esse comportamento fica mais claro no gráfico das sessões de tamanho 17 (4.6e).

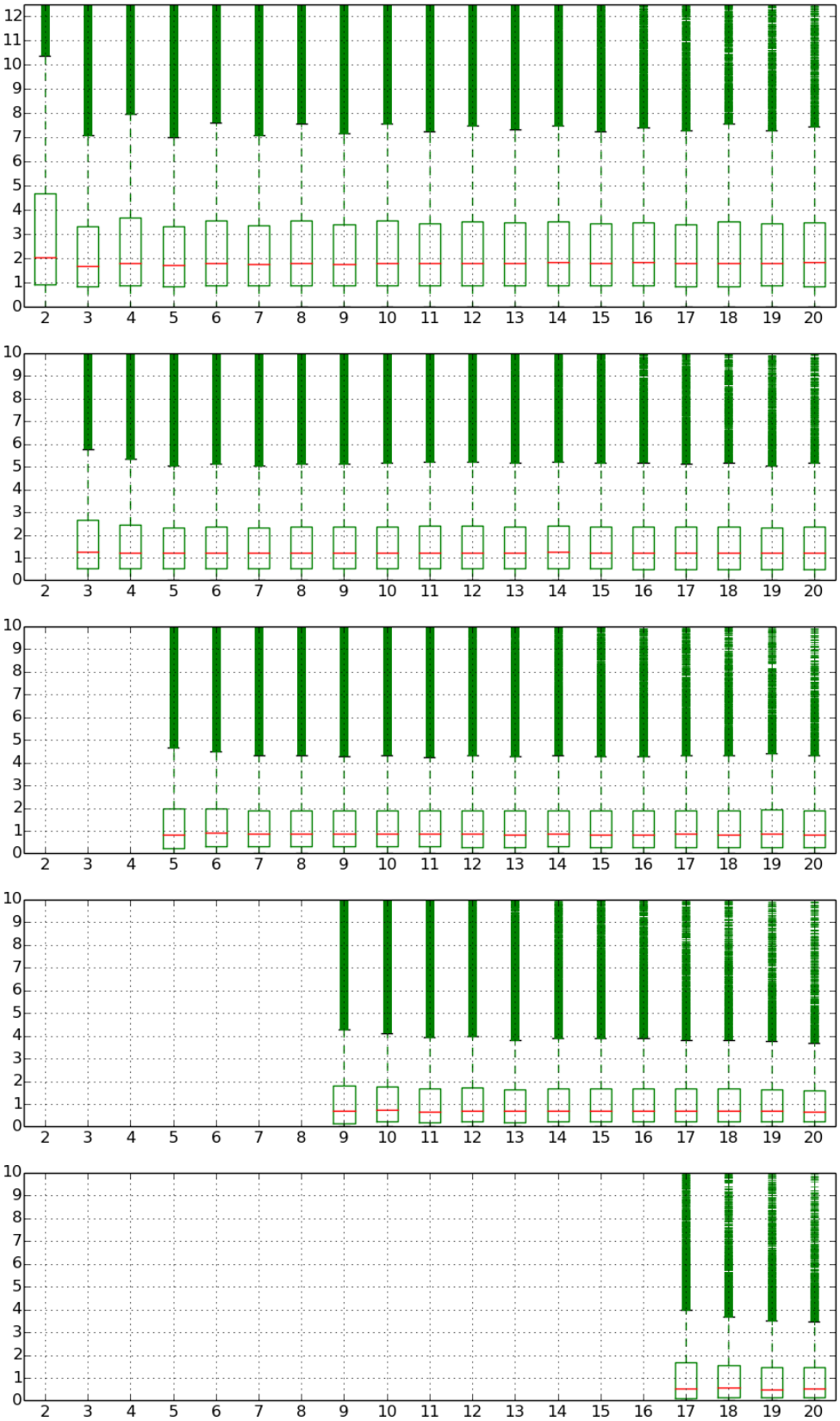
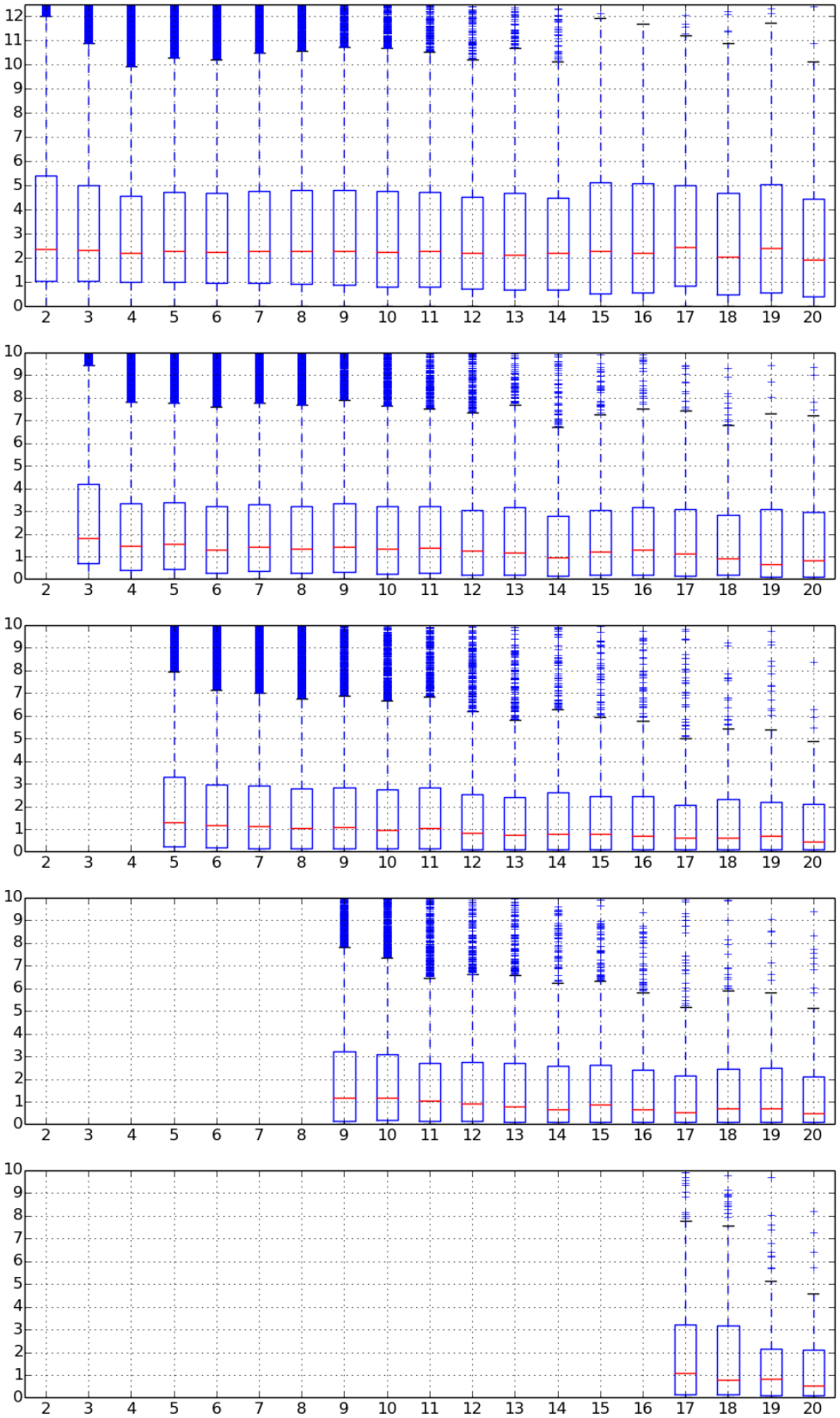


Figura 4.3: Intervalos de leitura do **Jornal Online A** filtrados pela ordem do intervalo e plotados pelo tamanho da sessão (eixo X). Eixo Y é o tempo em minutos.



1º intervalo de leitura.

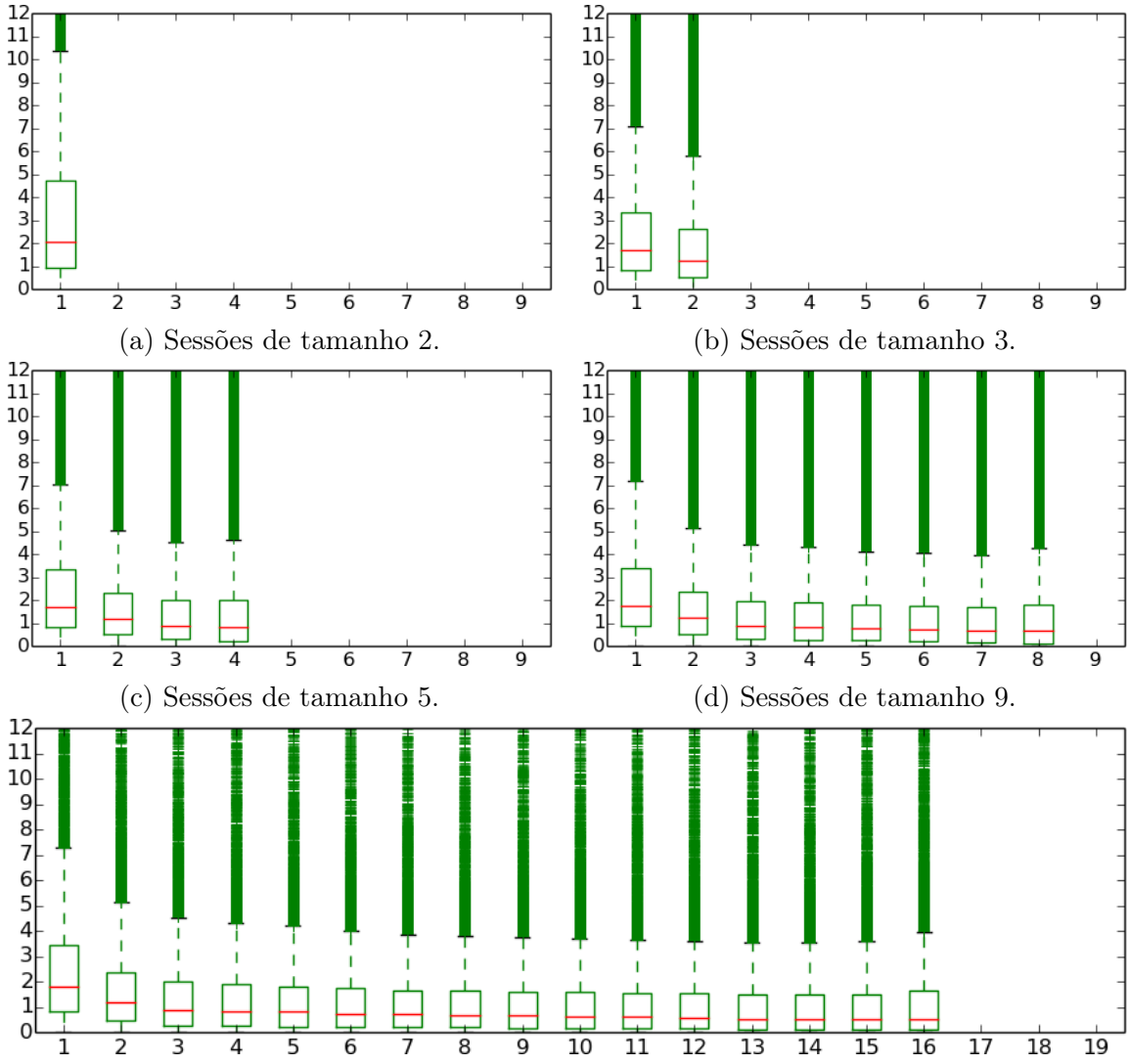
2º intervalo de leitura.

4º intervalo de leitura.

8º intervalo de leitura.

16º intervalo de leitura.

Figura 4.4: Intervalos de leitura do **Jornal Online B** filtrados pela ordem do intervalo e plotados pelo tamanho da sessão (eixo X). Eixo Y é o tempo em minutos.



(a) Sessões de tamanho 2.

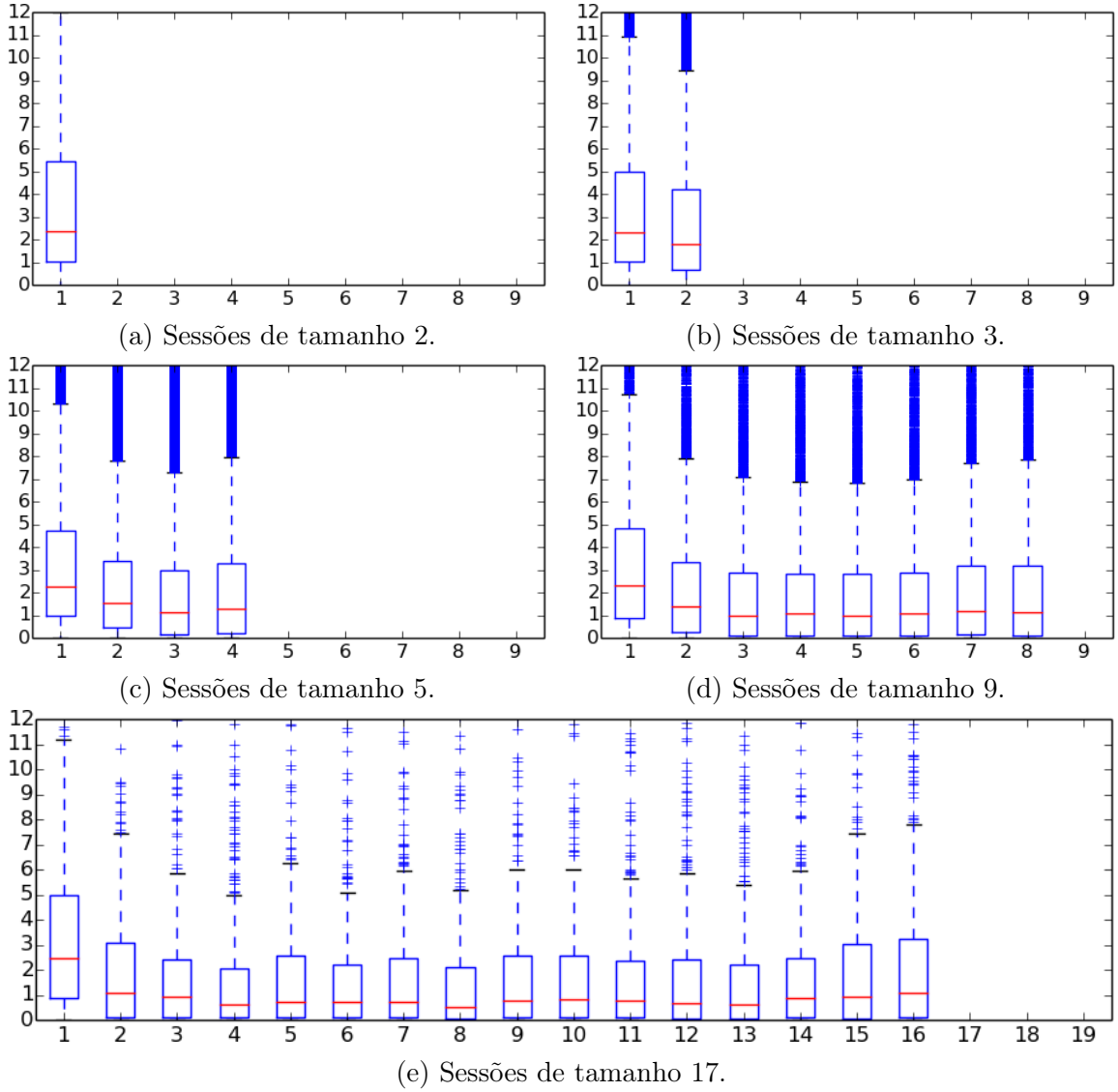
(b) Sessões de tamanho 3.

(c) Sessões de tamanho 5.

(d) Sessões de tamanho 9.

(e) Sessões de tamanho 17.

Figura 4.5: Distribuição do intervalo de leitura do **Jornal Online A** filtrados pelo tamanho da sessão e plotados por intervalo (eixo X). Eixo Y é o tempo em minutos.



(a) Sessões de tamanho 2.

(b) Sessões de tamanho 3.

(c) Sessões de tamanho 5.

(d) Sessões de tamanho 9.

(e) Sessões de tamanho 17.

Figura 4.6: Distribuição do intervalo de leitura do **Jornal Online B** filtrados pelo tamanho da sessão e plotados por intervalo (eixo X). Eixo Y é o tempo em minutos.

4.2 Intervalos de Leituras Consecutivas

Como vimos, os valores dos intervalos de leitura são bem variados nos inícios das sessões e tendem a diminuir a variância quanto maior for o tamanho da sessão. A próxima análise a ser apresentada tenta identificar a dependência entre o tempo de leituras consecutivas. Dado que o primeiro intervalo de leitura foi pequeno, o próximo intervalo tende a ser pequeno também? Ou, pelo contrário, tende a ser mais demorado?

As Figuras 4.7 e 4.8 mostram como é a distribuição do intervalo de leituras em dois instantes consecutivos condicionado ao primeiro instante. Os gráficos da primeira

linha são das sessões de tamanho 3. Os gráficos da segunda linha pertencem às sessões de tamanho 4, e os gráficos da terceira linha são das sessões de 5 artigos. Cada coluna mostra instantes de tempo diferentes, que estão listados abaixo das figuras. Os números somam 100 ao longo de cada coluna e representam porcentagens dentro da coluna. Para todos os gráficos, o primeiro instante de tempo I_i está no eixo X e o instante de tempo conseguinte I_{i+1} no eixo Y.

Os valores de intervalo de tempo variam de 0 a 30 minutos. Nessa análise utilizamos nos gráficos uma granularidade não linear nos eixos para melhor representatividade. Os intervalos apresentados são: 0-3 segundos, 3-10 segundos, 10-30 segundos, 30 segundos a 1 minuto, 1-3 minutos, 3-10 minutos e 10-30 minutos. Esses valores foram escolhidos após a análise dos resultados da sessão anterior. Observamos que a duração normal do intervalo de leitura é entre 1 e 3 minutos. Logo, destacamos esse intervalo nos gráficos com círculos vermelhos nas colunas e círculos cinzas nas linhas.

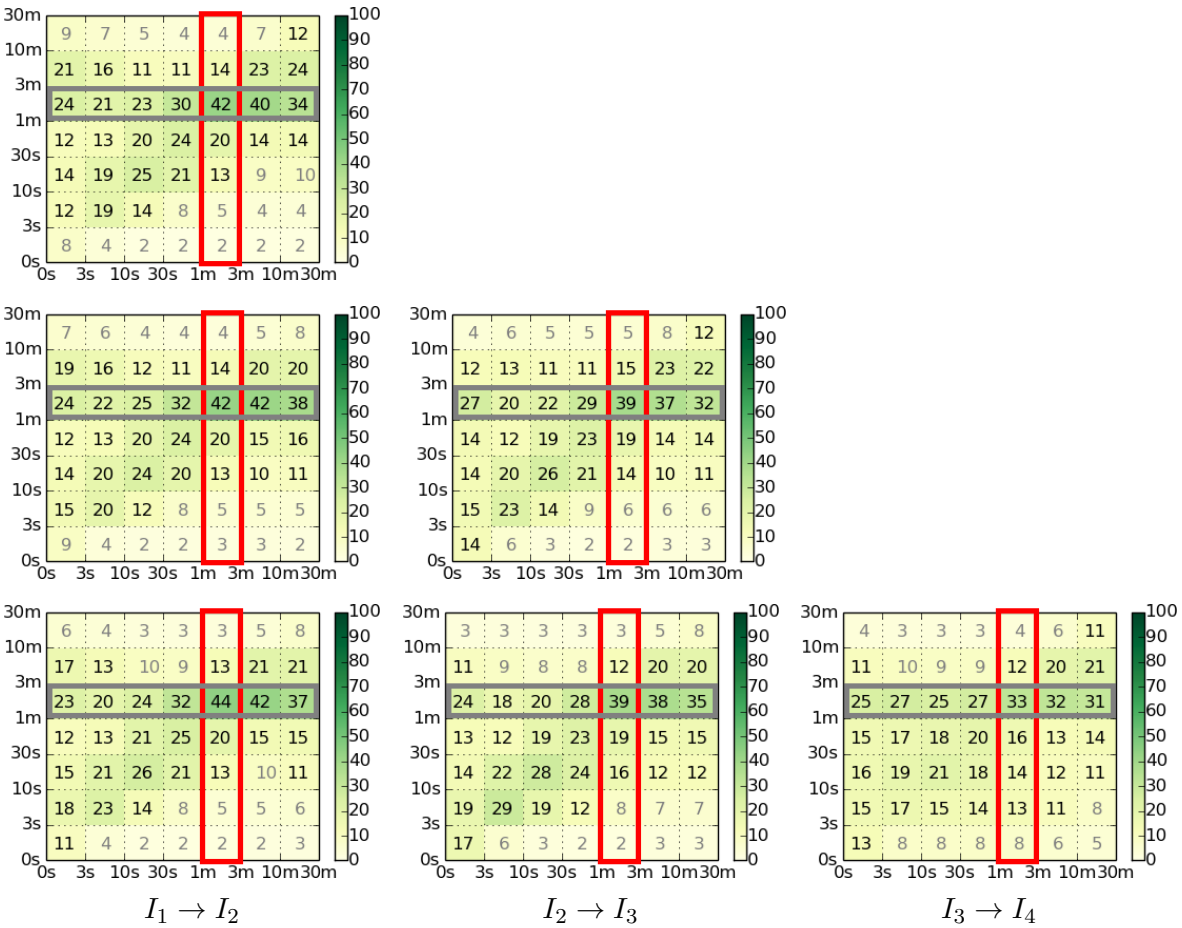


Figura 4.7: Distribuição do intervalo de leituras em dois instantes consecutivos condicionado no primeiro, dados do **Jornal Online A**.

No caso do **Jornal Online A** (Figura 4.7), podemos ver que há uma fraca re-

lação de dependência entre os intervalos de leitura. Todos os gráficos mostram que é alta a porcentagem de vezes que o intervalo de tempo I_{i+1} está entre 1 e 3 minutos, independentemente do tempo gasto no intervalo de leitura I_i. Contudo, os valores não são sempre os mesmos. Quando o usuário gasta pouco tempo no primeiro instante, por exemplo menos de 3 segundos, ele tem alta chance de gastar menos de 3 minutos no segundo intervalo. Quando o tempo do primeiro instante vai aumentando, a probabilidade dos tempos menores no segundo instante vão diminuindo, seguindo a diagonal. Temos assim uma leve dependência positiva entre os tempos gastos em uma leitura e a seguinte.

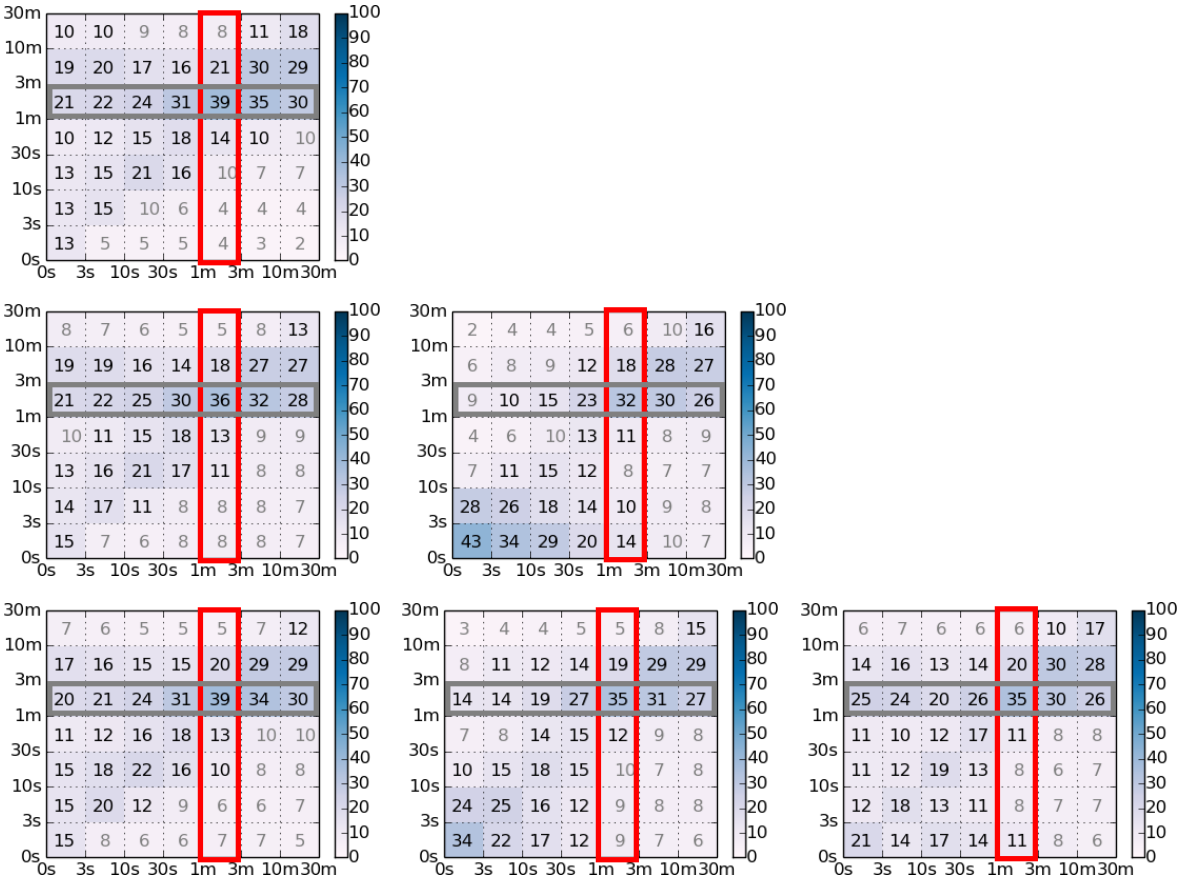


Figura 4.8: Distribuição do intervalo de leituras em dois instantes consecutivos condicionado no primeiro, dados do **Jornal Online B**.

No caso do **Jornal Online B** (Figura 4.8) há uma dependência maior do que aquela encontrada no **Jornal Online A**. O tempo de leitura de um artigo está influenciando o tempo de leitura do artigo seguinte. Como antes, há uma concentração de leitura durando de 1 a 3 minutos, mas não em todos os casos, e o crescimento das probabilidades acompanha a diagonal com mais força.

4.3 Os Tópicos ao Longo das Leituras

No capítulo anterior, listamos os tópicos e a informação daqueles mais acessados. Nesta próxima análise tentamos identificar se os acessos aos tópicos são distribuídos homogeneamente pela ordem das leituras. Os resultados obtidos são mostrados a seguir nos gráficos das Figuras 4.9 e 4.10. Plotamos os percentuais de leitura dos tópicos do primeiro ao vigésimo artigo lido separadamente. Os percentuais somam 100% ao longo de cada coluna. Quanto mais forte a cor, mais alta a porcentagem.

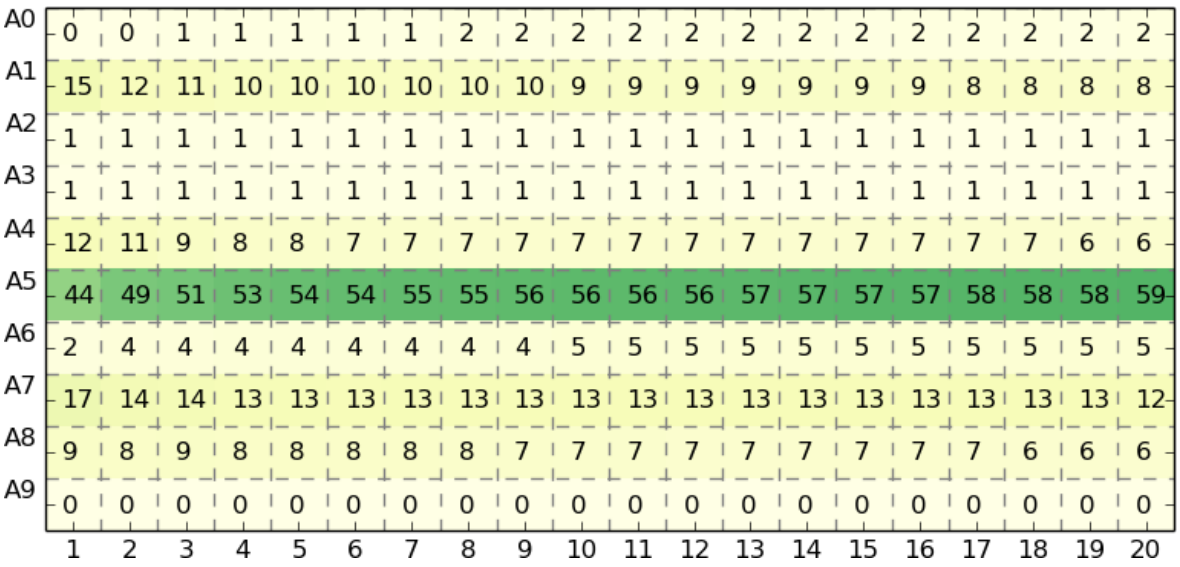


Figura 4.9: Distribuição da quantidade de tópicos lidos ao longo das leituras do **Jornal Online A**. No eixo X está a ordem dos artigos lidos e no eixo Y os tópicos.

Nos dados do **Jornal Online A** (Figura 4.9) as probabilidades se mantêm quase constantes ao longo das leituras. Os tópicos mais lidos no primeiro artigo permanecem como os mais lidos em todas os demais artigos lidos numa sessão. A mudança mais significativa é que o tópico mais lido, `A5`, aumenta sistematicamente seu peso com os outros tópicos tendo suas proporções diminuídas. O decrescimento da leituras dos demais tópicos é leve e não há mudança na ordem de preferência de leituras dos tópicos.

No caso do **Jornal Online B** (Figura 4.10), temos maior variabilidade ao longo das leituras. O tópico mais lido (`B1`) se mantém quase constante. O segundo mais lido inicialmente (`B3`) cresce sua participação rapidamente até a décima leitura quando então diminui o seu peso. O terceiro mais lido inicialmente (`B8`) cresce sua porcentagem de forma sistemática, chegando a ser o segundo tópico mais lido nos instantes acima de 10. Artigos de `B3` deveriam ser fortemente recomendados a partir da segunda leitura até a décima. Artigos do tópico `B8` deveriam ser mais recomendados se o leitor já leu mais de 10 artigos.

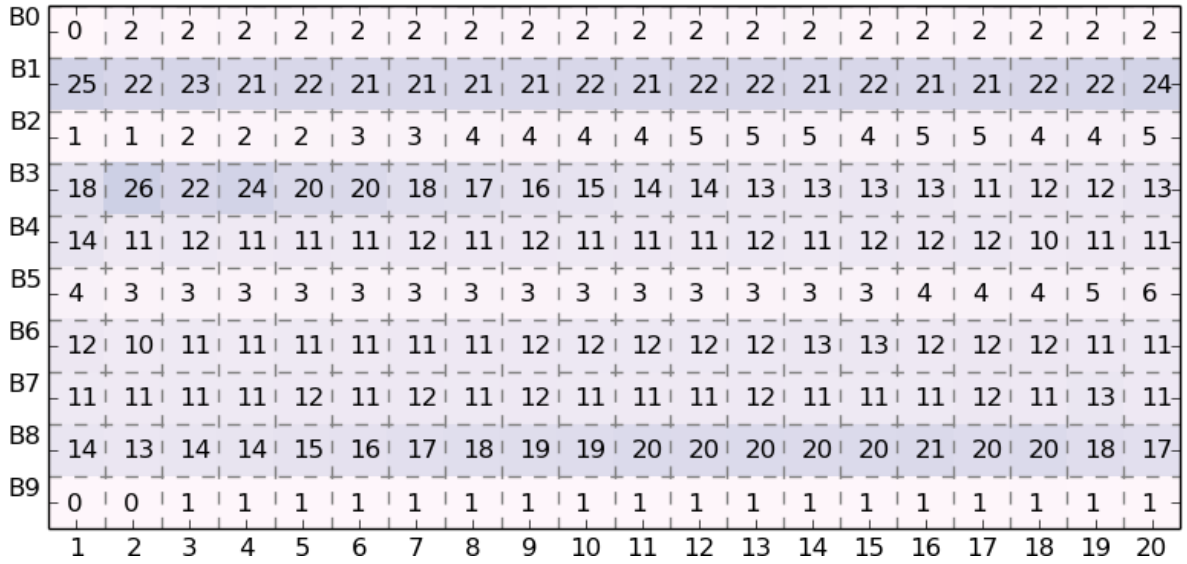


Figura 4.10: Distribuição da quantidade de tópicos lidos ao longo das leituras do **Jornal Online B**. No eixo X está a ordem dos artigos lidos e no eixo Y os tópicos.

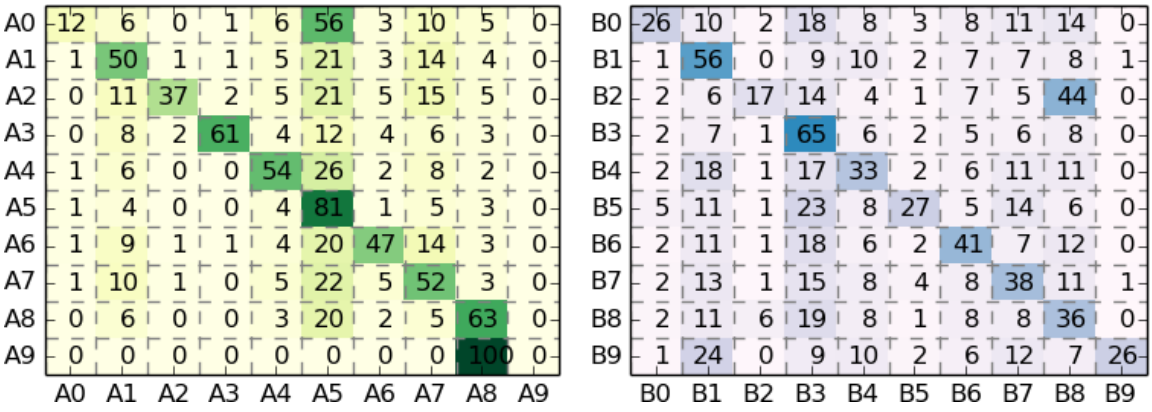
Com esses dados podemos concluir que os acessos aos tópicos do **Jornal Online A** são mais homogêneos do que os do **Jornal Online B**. Note que as probabilidades plotadas nas duas tabelas seriam as probabilidades $\mathbb{P}(T_n = l)$ do modelo **M-Independência** em (2.6) caso todas as sessões das bases fossem utilizadas no treino.

4.4 Transição de Tópicos entre Leitura

As leituras de uma sessão podem conter artigos de mais de um tópico. Como há uma ordem de leitura, há transições de um tópico para outro. Nesta seção, procuramos identificar como são as transições entre tópicos. Sabendo que a i-ésima leitura foi em um tópico especifico $T_i = l_i$, estimamos as probabilidades do tópico T_{i+1} da próxima leitura. Isto é, estimamos as probabilidades de transição entre leituras $\mathbb{P}(T_{n+1} = l \mid T_n = l_n)$ para diferentes valores de n.

Estimamos separadamente a transição do primeiro artigo para o segundo, a transição do segundo para o terceiro artigo, etc. Também obtivemos uma estimativa global de transição assumindo que a probabilidade $\mathbb{P}(T_{n+1} = l \mid T_n = l_n)$ não varia com n. Observamos que os valores das transições específicas por ordem de leitura n são parecidas com os valores da transição global, assim vamos mostrar a seguir somente a tabela da transição geral.

Na Figura 4.11(a) temos a informação das transições do **Jornal Online A**. Os números são porcentagens e somam 100% ao longo das linhas. Para o primeiro tópico, `A0`, temos os seguintes percentuais de transições: 12%, 6%, 0%, 1%, 6%, 56%, 3%,



(a) **Jornal Online A** (b) **Jornal Online B**

Figura 4.11: Distribuição geral da transição entre tópicos condicionada no tópico anterior. Tópico anterior no eixo Y e tópico posterior no eixo X.

10%, 5% e 0%. Logo, o usuário passa para o tópico `A5` na leitura seguinte em 56% dos casos em que a leitura em um dado instante estava no tópico `A0`. Esse é um dos tópicos onde há mais transição para um tópico diferente do que permanência nele.

Em geral, as leituras consecutivas tendem a permanecer no mesmo tópico. Isso é evidente a partir dos números na diagonal principal da tabela, que é carregada com as maiores probabilidades. O tópico `A5` é o que mais recebe transições dos demais tópicos, seguido pelos tópicos `A1` e `A7`. Com exceção desses três tópicos, nenhum outro recebe muitas transições oriundas de tópicos distintos. Uma anomalia dentre os tópicos é o tópico `A9`. Esse tópico possui todos as leituras seguintes para o tópico `A8`. Junto a esse fato, praticamente nenhuma leitura oriunda de outros tópicos chega a esse tópico, deixando assim os índices da tabela de transição praticamente zerados.

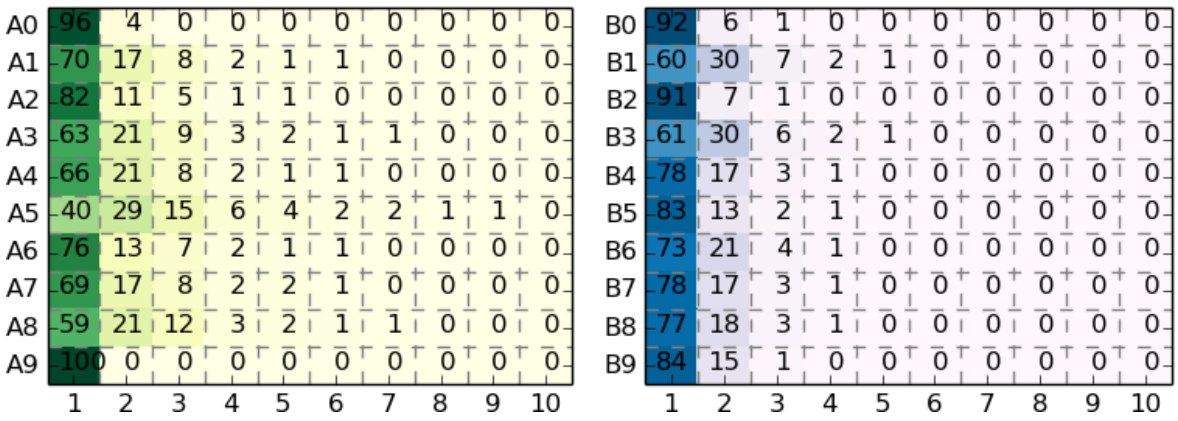
Na Figura 4.11(b) temos os percentuais gerais das transições do **Jornal Online B**. A diagonal principal também contém as maiores probabilidades, mostrando que os usuários tendem a ter leituras consecutivas num mesmo tópico. Se comparado com a diagonal principal do outro jornal, vemos que os usuários permanecem um pouco menos nos mesmos tópicos no caso do **Jornal Online B**. Nesse jornal, o tópico `B2` é aquele em que há mais transição para um tópico diferente. Um total de 44% das leituras que estavam em `B2` passaram para leituras do tópico `B8`. Os tópicos que mais recebem transições de outros tópicos são `B1`, `B3` e `B8`.

Se desconsiderarmos os tópicos e só analisarmos a informação se o usuário permaneceu no mesmo tópico ou mudou-se dele quando iniciou uma nova leitura teremos os valores de 67% para permanência no tópico e 23% para mudança (**Jornal Online A**) e 46% para permanência e 54% para mudança (**Jornal Online B**). Observe que os usuário do primeiro jornal permanecem mais nos tópicos do que os usuários do segundo

jornal. Esses valores mostram alta permanência, mas quantas leituras os usuários permanecem nesses tópicos em geral? Na próxima sessão, mostramos a resposta a essa pergunta.

4.5 Permanência nos Tópicos

A próxima análise considera o número de artigos que um usuário lê em sequência de um mesmo tópico. Vamos estudar a distribuição das leituras em que o usuário permanece em um tópico, independente da ordem de aparecimento do tópico na sessão, se na primeira leitura, na segunda, etc. Essa análise foi chamada de permanência geral. A Figura 4.12 mostra cada tópico numa linha da tabela. O eixo horizontal indica a duração, em número de artigos, da permanência no tópico a partir do momento de entrada no tópico. Os números no corpo da tabela são porcentagens que somam 100% ao longo das linhas.



(a) **Jornal Online A** (b) **Jornal Online B**

Figura 4.12: Permanência geral. Número de artigos no eixo X e tópicos no eixo Y.

Os índices de permanência são bem variados entre os tópicos mas dificilmente passam de quatro leituras. Em geral, os usuários permanecem pouco tempo em um mesmo tópico. A maioria dos usuários do jornal **Jornal Online A** permanece por duas ou três leituras em um mesmo tópico, e os usuários do **Jornal Online B**, no máximo, duas leituras. Particularmente no caso do **Jornal Online A**, a permanência no tópico `A0` é muito baixa, em geral não houve quase nenhum outro artigo do mesmo tópico lido em sequência. Pela Figura 4.12a, vemos que 96% das vezes os usuários leem somente um artigo desse tópico e os 4% restantes são de permanência por duas leituras. Já a permanência no tópico `A3` é bem maior. 63% dos casos permanecem

somente uma leitura, 21% permanecem duas leituras, 9% permanecem três leituras e o restante permanecem no mínimo 4 leituras nesse tópico.

O fim da permanência em um mesmo tópico pode ser ocasionado por dois fatores: por **mudança de tópico**, quando o usuário passa a ler artigo de outro tópico, ou por **saída da sessão**, quando o usuário não lê mais nenhum artigo, finalizando sua sessão de leitura.

Analisamos separadamente os casos de permanência até a mudança de tópico e até a saída da sessão. Também separamos a ordem de aparecimento dos tópicos na sessão: primeiro tópico, segundo tópico da sessão, etc. Os resultados mostraram um comportamento com valores bem próximos ao da permanência geral. Não plotamos esses casos especiais devido à sua similaridade com o gráfico de permanência geral. Assim, concluímos que o número de leituras que um usuário permanece em determinado tópico não se altera dependendo da ação futura, trocar de tópico ou finalizar a sessão, e aparenta fortemente não depender da ordem de aparição dos tópicos na sessão.

4.6 Os Principais Padrões de Trajetórias entre Tópicos

Vimos na última seção que os usuários tendem a fazer poucas leituras seguidas em um mesmo tópico. O que não é claro ainda é se essas poucas leituras seguidas em um tópico esgotam o interesse do usuário em uma sessão. Estamos interessados em verificar se, após sair de um tópico, um usuário tende a retornar ao mesmo. Se isso for verdade, a possibilidade de usar esse conhecimento para recomendar notícias é clara. Um sistema de recomendação deveria rastrear os tópicos já lidos e recomendar artigos de alguma maneira baseado na probabilidade de retorno ao tópico.

Dado esse contexto, tentamos identificar se há padrões no histórico de tópicos em uma sessão. Fizemos duas análises. A primeira será descrita nesta seção e a segunda, na seção seguinte. A primeira análise considerou somente as mudanças de tópicos. Tentamos identificar como são as mudanças entre os tópicos, independente de quantas leituras são feitas dentro de um tópico até que a transição ocorra. Para isto, contabilizamos as frequências de todas trajetórias distintas de tópicos encontradas nas bases de dados. Foram identificados 151.007 padrões de trajetórias de tópicos diferentes no **Jornal Online A** e 61.543 no **Jornal Online B**. Esses padrões de trajetórias de tópicos continham nenhuma, uma ou diversas mudanças de tópicos. No **Jornal Online A**, o padrão mais extremo continha 55 transições entre tópicos, e no **Jornal Online B**, foi encontrado um padrão que continha 74 mudanças de tópicos. Entretanto, esses são

padrões pouco frequentes. Tentando resumir e extrair informação útil desse grandes volume de padrões, apresentamos algumas estatísticas gerais na Tabela 4.1 e plotamos graficamente os principais padrões encontrados (Figuras 4.13, 4.14, 4.15, e 4.16).

Quantidade de Tópicos no Padrão	**Jornal Online A**		**Jornal Online B**	
	% das Sessões	Padrões Diferentes	% das Sessões	Padrões Diferentes
1	48,3%	9	37,2%	10
2	34,6%	73	48,0%	90
3	10,6%	562	9,1%	794
4	3,3%	3724	3,4%	5094
5	1,6%	14992	1,2%	14638
6 ou mais	1,5%	131647	1,0%	40917

Tabela 4.1: Estatísticas gerais dos padrões de trajetória de ambos os jornais.

A Tabela 4.1 mostra que, dentre mais de 150 mil padrões de trajetória do **Jornal Online A**, nós temos 9 padrões de trajetórias que possuem somente um único tópico e que juntos representam mais de 48% das sessões da base (o único tópico que não fez uma sessão sozinho é o `A9`). A seguir, 73 padrões de trajetórias onde houve uma mudança de tópico, constituindo aproximadamente 35% das sessões da base. Mais de 130 mil padrões de trajetórias contendo 6 ou mais tópicos representam menos de 2% das sessões da base de dados. Dessa forma, a maioria (73%) das sessões do **Jornal Online A** têm apenas um ou dois tópicos.

A mesma tabela também apresenta algumas estatísticas para o **Jornal Online B**. Para essa base, contabilizamos 10 padrões de trajetória com somente um tópico (um padrão por tópico possível), esses padrões representando 37% das sessões desse jornal. Em seguida, temos 90 padrões de trajetória que começaram em um tópico e terminaram em outro tópico. Esses padrões de uma mudança somam 48% das sessões. Temos 794 padrões onde houveram duas mudanças de tópicos. Esses últimos padrões somam praticamente 9% das sessões da base. O valor de 794 padrões se aproxima dos 810 padrões possíveis com somente duas mudanças, valor mais alto que o encontrado no outro jornal.

Como foram identificados muitos padrões de trajetória, selecionamos os mais frequentes de cada jornal. As Figuras 4.13 e 4.15 apresentam os principais padrões do **Jornal Online A** e as Figuras 4.14 e 4.16, os padrões do **Jornal Online B**. Em cada Figura, os padrões são separados por cores, em uma escala de vermelho a roxo, passando pelo verde. Os padrões de maior frequência são os vermelhos e os de menor frequência, os roxos. Na legenda de cada Figura, há entre colchetes uma sequência de números que designa as transições entre tópicos (os números representam os tópicos) e

entre parênteses o percentual de vezes que aquele padrão acontece em toda a base de dados.

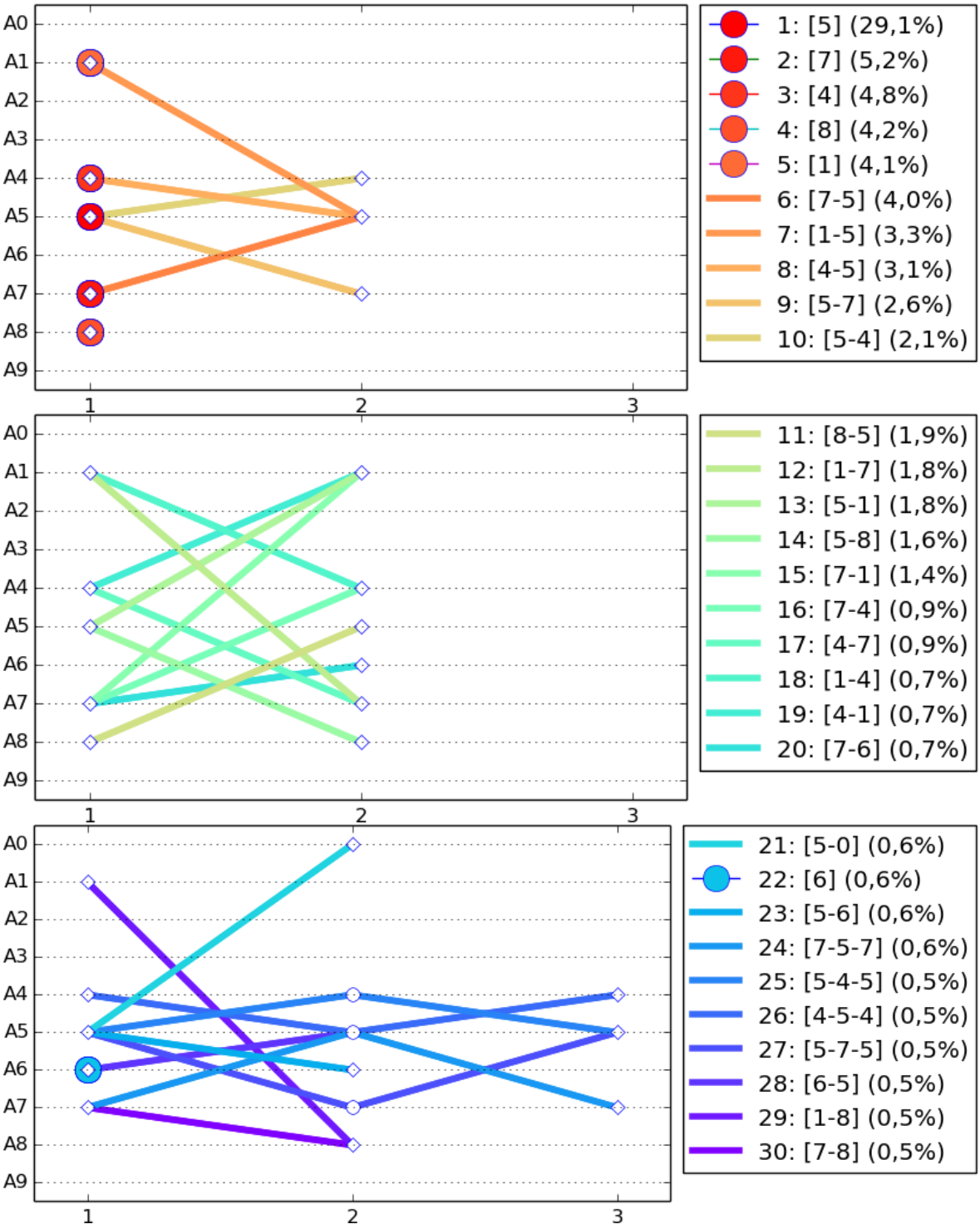


Figura 4.13: Top-30 padrões de trajetória das sessões do **Jornal Online A**.

Nas Figuras 4.13 e 4.14 temos os 30 maiores padrões de trajetória de ambos os

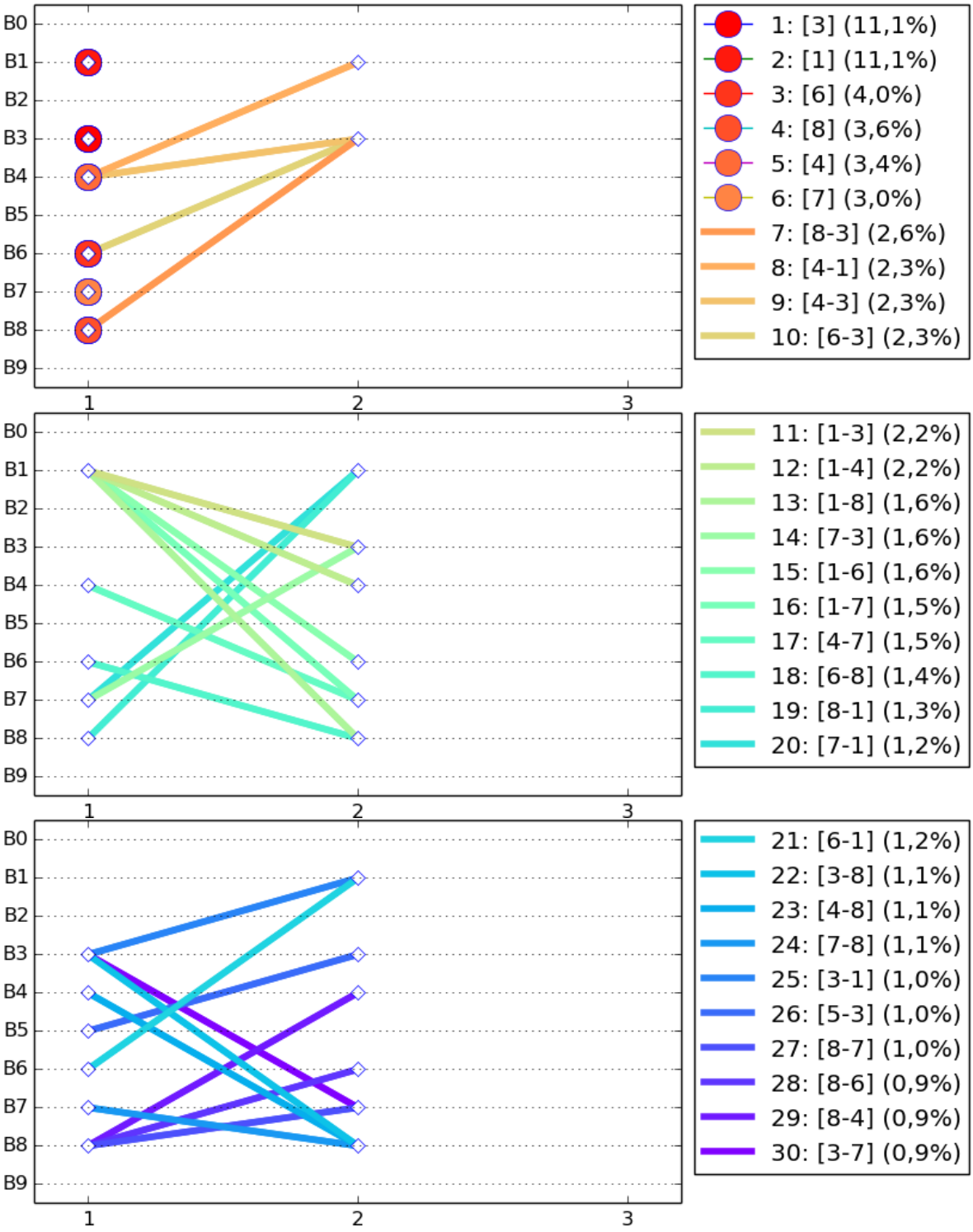


Figura 4.14: Top-30 padrões de trajetória das sessões do **Jornal Online B**.

jornais. Foram escolhidos apenas os 30 primeiros padrões para uma melhor visualização e também porque eles representam um alto percentual do conjunto de sessões. No caso do **Jornal Online A**, os top-30 padrões representam praticamente 80% de todas as

sessões dessa base. No caso do **Jornal Online B**, os top-30 padrões somam juntos mais de 70% das sessões.

Nos dois jornais, predominam os padrões de nenhuma ou apenas uma mudança de tópico. Na Figura 4.13 (**Jornal Online A**), 5 tópicos principais chamam a atenção, ora no padrão de leitura de tópico único, ora marcando presença nos padrões de dois tópicos. Também há a presença de 4 padrões de trajetória com duas mudanças de tópicos. Entretanto, existem apenas dois tópicos distintos entre esses três tópicos visitados. Esses padrões de três tópicos mostram que o usuário eventualmente *retorna* ao tópico no qual começou. No caso do **Jornal Online B** (Figura 4.14), são 6 tópicos distintos que se destacam dos demais. Eles constituem padrões de trajetória com apenas um ou dois tópicos.

Esta análise nos mostra dois pontos importantes. Primeiro, que as sessões com somente um tópico são bem frequentes. Segundo, que os tópicos mais lidos se revesam, ora como primeiro, ora como segundo tópico em uma sessão. Houve poucos padrões de 3 ou mais tópicos e os que apareceram nos resultados mostraram um comportamento de retorno ao primeiro tópico.

Como houve poucos casos de trajetórias com duas ou mais mudanças de tópicos, resolvemos plotar esses casos em particular. Porém com o número de sessões com k mudanças varia muito com k, decidimos analisar somente os 12 primeiros padrões de trajetórias de cada caso em separado. As Figuras 4.15 (referente ao **Jornal Online A**) e 4.16 (referente ao **Jornal Online B**) mostram os 12 principais padrões de trajetórias em cada um dos grupos de sessões para $k = 2, 3$ ou 4 mudanças de tópico.

O aspecto mais marcante desses gráficos é o padrão cíclico entre dois tópicos em praticamente todos as trajetórias nos dois jornais. Os padrões mais frequentes mostram um retorno ao primeiro tópico visitado. O usuário permanece algumas leituras em um tópico, muda para um segundo tópico aonde fica algum tempo e normalmente volta para o primeiro tópico. Em alguns poucos casos, há mudança para um terceiro tópico.

Esse padrão cíclico demonstra que os usuários não esgotam as leituras de um tópico e transitam para outro tópico sem chances de voltar ao primeiro tópico. O comportamento identificado é exatamente o oposto. Os usuários tendem a ler artigos de um tópico, transitam para outro tópico e muito provavelmente voltam para o primeiro tópico.

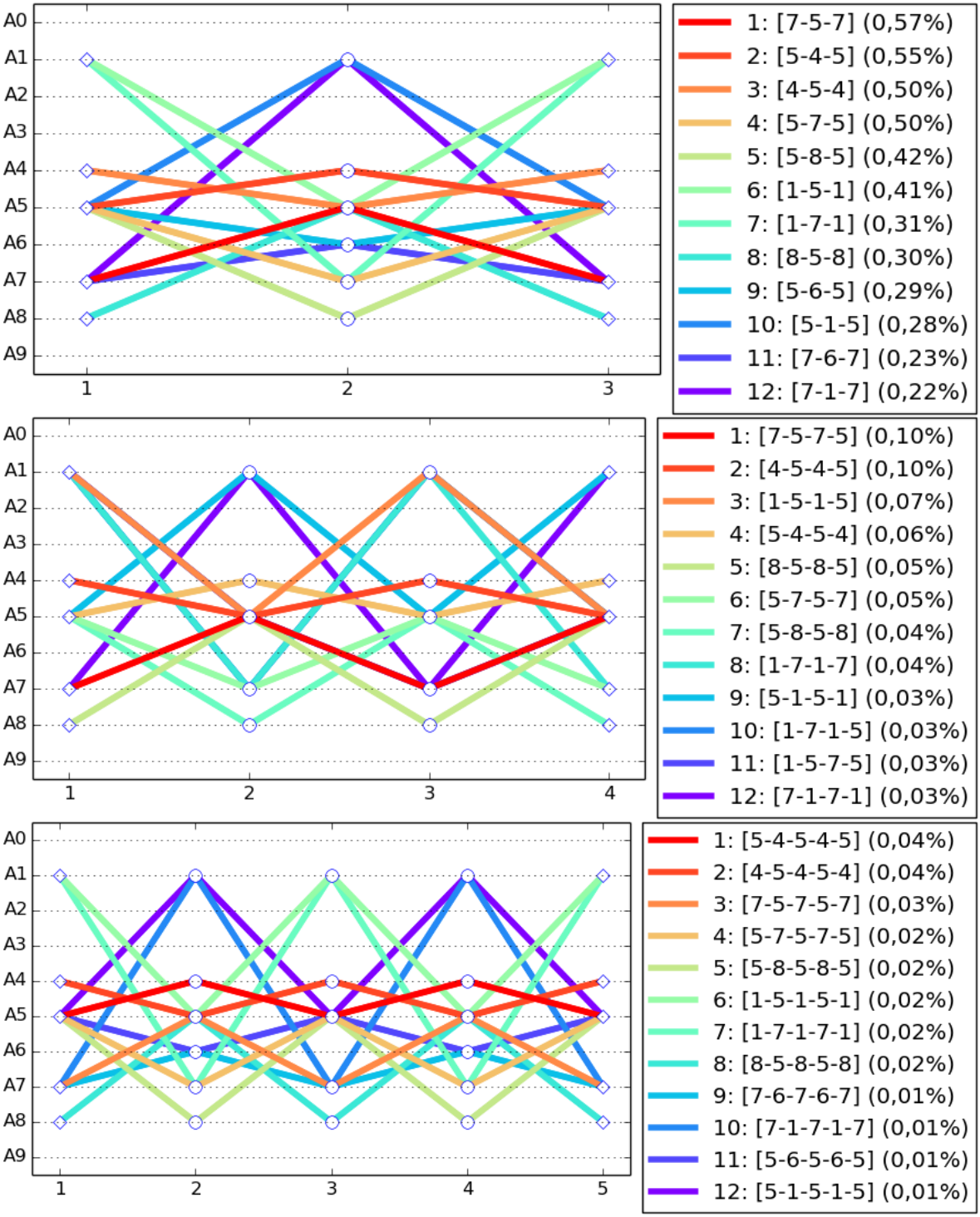


Figura 4.15: Top 12 padrões de trajetória de 2, 3 e 4 mudanças, **Jornal Online A**.

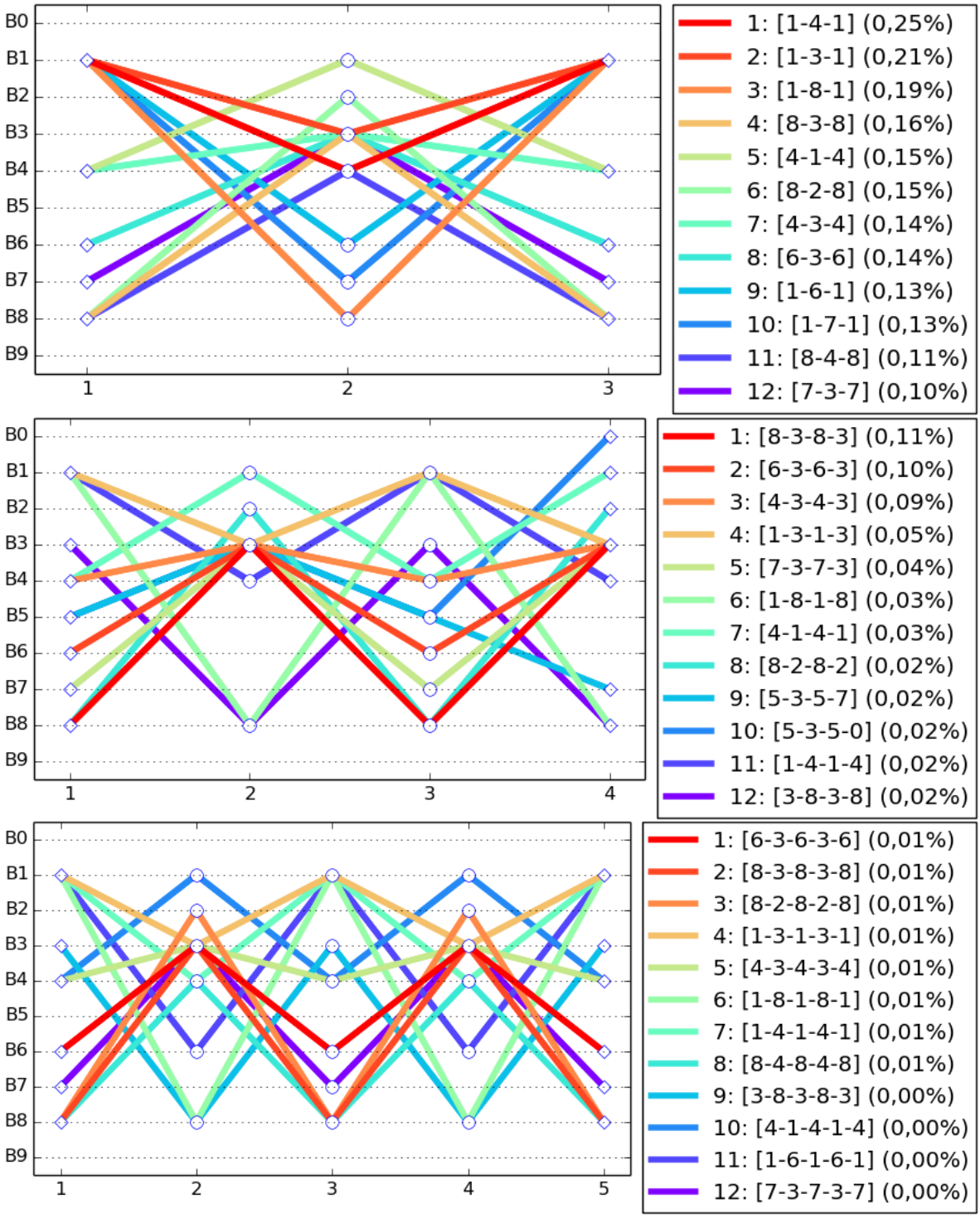


Figura 4.16: Top 12 padrões de trajetória de 2, 3 e 4 mudanças, **Jornal Online B**.

4.7 Os Principais Padrões de Trajetórias Leitura a Leitura

Na primeira análise de trajetória de tópicos nós ignoramos a duração em cada tópico. Nesta seção, queremos levar em conta também a permanência num dado tópico. Identificamos os padrões de troca de tópicos mais frequentes levando em conta a quantidade de artigos lidos em cada tópico. Logo, as trajetórias (`A5`, `A5`) e (`A5`, `A5`, `A5`) são consideradas trajetórias distintas pois a duração no (único) tópico é diferente. A seguir, veremos as top-30 trajetórias de cada jornal considerando o tópico instante a instante.

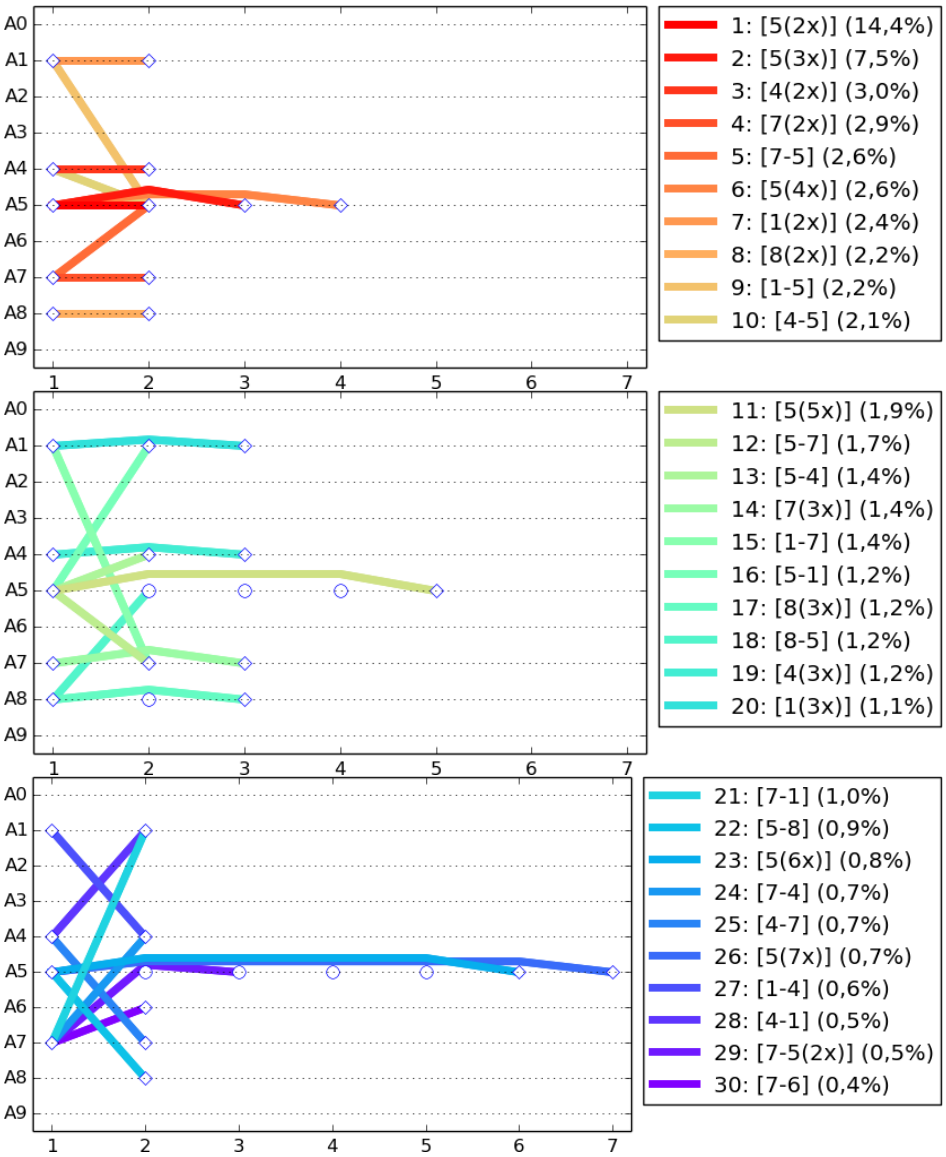


Figura 4.17: Os 30 maiores padrões de trajetórias entre tópicos do **Jornal Online A**, leitura a leitura.

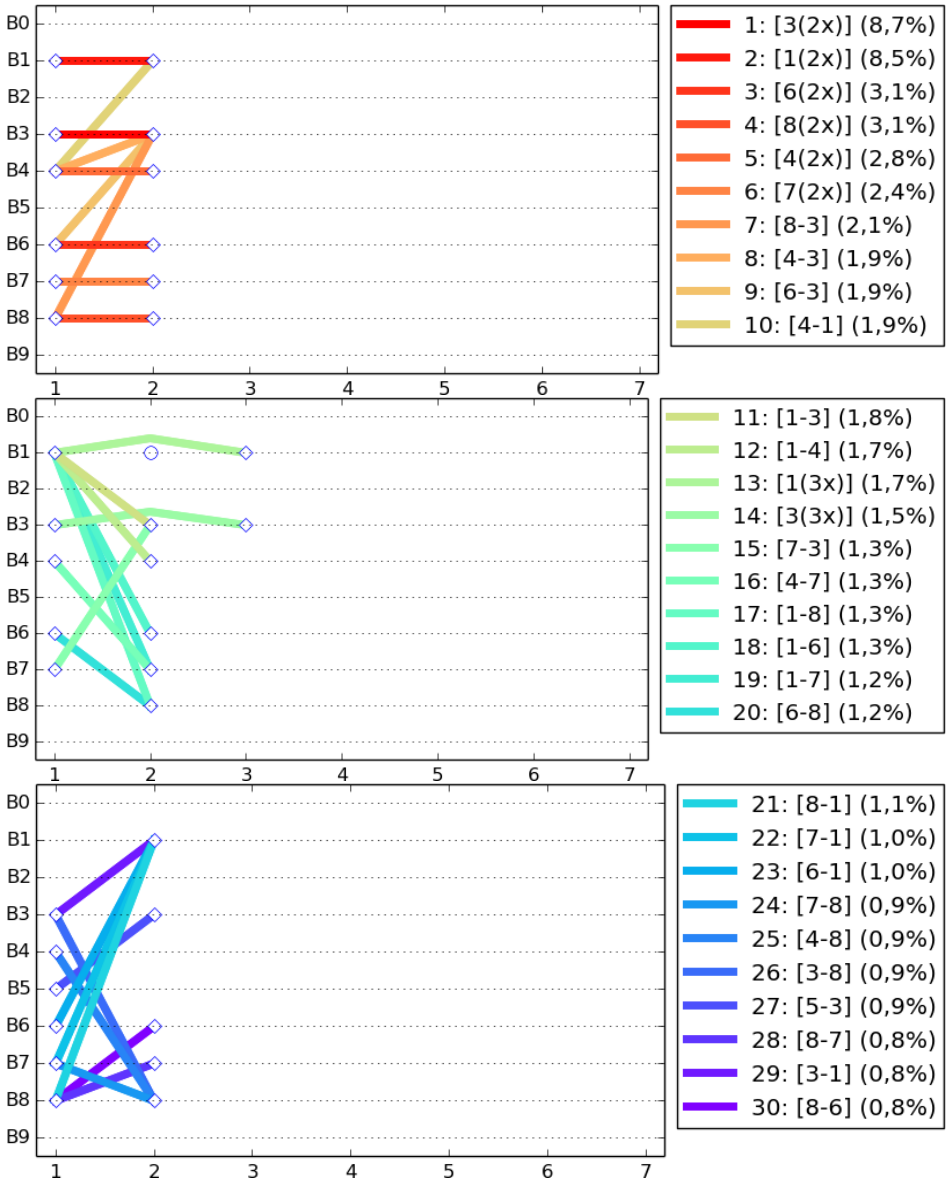


Figura 4.18: Os 30 maiores padrões de trajetórias entre tópicos do **Jornal Online B**, leitura a leitura.

No caso do **Jornal Online A** (Figura 4.17), vemos que são poucos os tópicos que duram muitas leituras. As sessões mais frequentes são as de duas leituras com dois tópicos distintos. Na Figura 4.18 referente ao **Jornal Online B**, vemos basicamente o mesmo padrão: uma baixa permanência no primeiro tópico, com troca de tópico seguida de rápida finalização da sessão.

Esses Top-30 padrões são uma pequena parcela de todos os padrões de trajetória que existem levando em conta a duração no tópico. Como é impossível mostrar todos os padrões, selecionamos os principais padrões pelo tópico inicial e analisamos os comportamentos. A descrição da análise e os resultados são apresentados a seguir.

4.8 Trajetórias por Tópico Inicial

Analisamos e identificamos as trajetórias de sessões que começam em um determinado tópico e somam no mínimo 70% dos casos dessas sessões. Essa análise foi chamada de **top 70 padrões de trajetórias**. Como em alguns casos encontramos muitos padrões, os resultados foram divididos em dois gráficos. O primeiro contém os top-60 padrões, as trajetórias de maior porcentagem que juntos somam 60%. E o segundo tem as trajetórias seguintes que, junto com as primeiras, atingem 70%. O esquema de cores nas trajetórias segue o mesmo esquema da seção anterior. Os padrões de maior frequência são os vermelhos e os de menor frequência os roxos. Na legenda, entre colchetes, há uma sequência de números que designam as transições entre tópicos (os números representam os tópicos). Entre parênteses, colocamos o percentual de vezes que aquele padrão aconteceu dentre as sessões que iniciam naquele tópico. Os gráficos da Figura 4.19 mostra um exemplo dos padrões identificados nessa análise. Os demais gráficos referente a cada um dos tópicos dos dois jornais estão presente no Apêndice A.

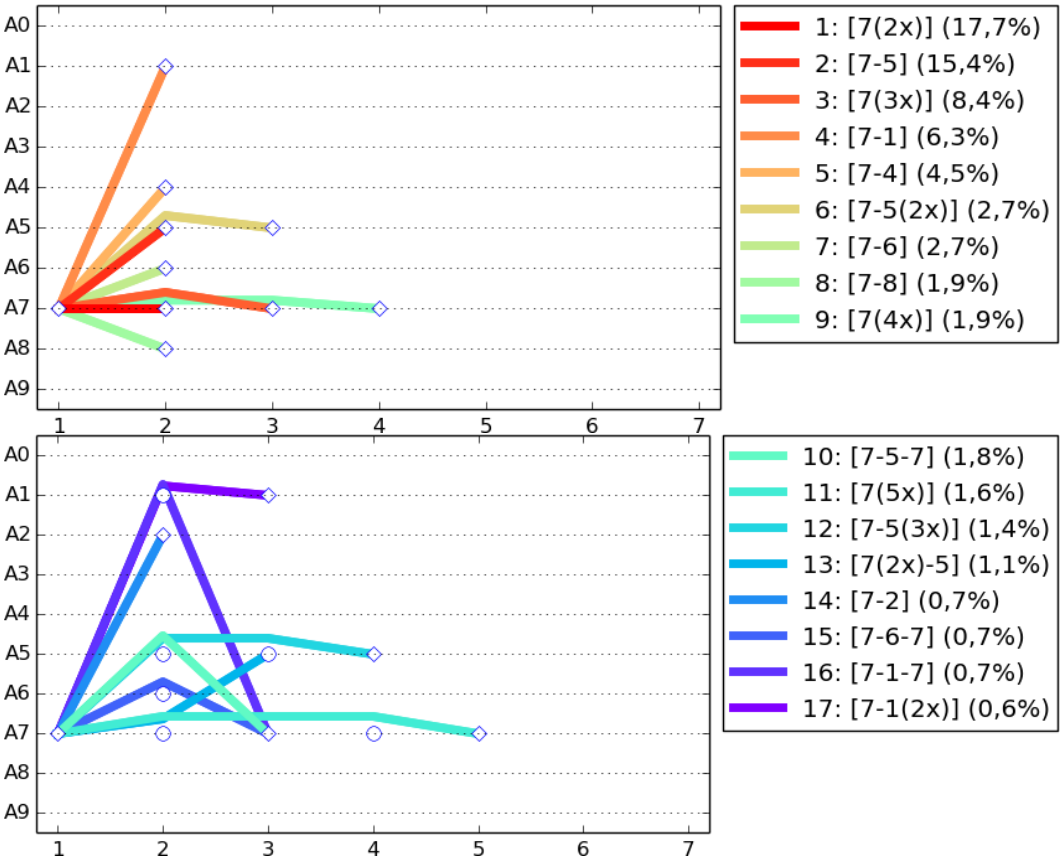


Figura 4.19: Exemplo da análise de top 70 padrões de trajetórias. O Top 70 padrões de trajetórias que começam no tópico `A7`.

O número de padrões de trajetórias que juntos somam 70% das sessões iniciadas

em um determinado tópico varia consideravelmente de tópico a tópico. No caso do **Jornal Online A**, o tópico que possui menos padrões é o tópico `A0` (7 no total) e o que possui mais padrões é o tópico `A2` (32 no total). Na média, são 16 padrões por tópico. No caso do **Jornal Online B**, esse número varia em torno de 9 padrões. O tópico `B3` é o que possui o menor número de padrões (5), e o tópico `B8` é o que possui o maior número de padrões (14).

Todo tópico tem um padrão relativamente mais longo que os demais. Para os tópicos do **Jornal Online A**, o tamanho desse padrão mais longo é normalmente de 5 ou 6 artigos lidos. E é sempre o padrão de permanência no mesmo tópico desde o início da sessão. Já para os tópicos do **Jornal Online B**, os padrões mais longos de cada tópico são de apenas 3 artigos em 90% dos casos. Somente em um tópico o padrão mais longo foi de 4 artigos. Outra diferença entre os jornais é que no **Jornal Online B** o padrão mais longo nem sempre é o de permanência no tópico inicial.

A quantidade de tópicos diferentes do tópico inicial que aparecem nos padrões top-70% varia substancialmente. Os tópicos `A0`, `A4`, `A5` e `A8` são os tópicos que possuem menos padrões com outros tópicos do **Jornal Online A**, com valores iguais a 2, 3, 3 e 4, respectivamente. Já os tópicos `A1`, `A3`, `A7`, `A2` e `A6` são os que possuem padrões com mais tópicos diferentes. Os valores para esse grupo são os seguintes: 5, 5, 6, 7 e 7. No caso do **Jornal Online B**, a maioria dos tópicos possuem padrões com 5, 6, 7 ou 8 tópicos nos seus principais padrões de trajetória. A única exceção é o tópico `B3` que possui no máximo um padrão com apenas 3 outros tópicos diferentes.

Resumindo, a quantidade de padrões diferentes por tópico é maior no **Jornal Online A**, fato que pode ser explicado pelo tamanho das sessões. Nos padrões que somam 70% das sessões, temos que os usuários do jornal **Jornal Online A** fazem leituras de 5, 6 ou 7 artigos, enquanto os usuários do outro jornal fazem sessões de 3 artigos. Mesmo fazendo sessões normalmente menores do que as sessões dos usuários do **Jornal Online A**, os usuário do **Jornal Online B** leem mais diversificadamente. Os top 70% dos padrões desses usuários apresentam mais tópicos (de 5 a 7) em comparação com os usuários do **Jornal Online A** (de 3 a 6).

4.9 Fluxos de Transições

A análise de **fluxo de transições** contabilizou como são os fluxos das leituras em cada tópico instante a instante. Das sessões que iniciaram a leitura em um determinado tópico, no instante seguinte, para qual tópico foi a leitura ou foi finalizada a sessão? O fluxo de transições foi medido nos instantes $1 \leq n \leq 10$. Nessa análise utilizamos além

das sessões relevantes, as sessões compostas por um único artigo lido. Essas sessões unas foram removidas da base e não entraram nas análises até este momento. Nós a utilizamos para avaliar os índices de saída da sessão desde a leitura do primeiro artigo.

Nos gráficos dessa análise temos no eixo X ordem da leitura e no eixo Y os rótulo dos 10 tópicos mais o rótulo `Saída` que denota o término da sessão. Os valores entre parênteses abaixo dos rótulos do eixo X mostram a porcentagem das leituras daquele tópico instante a instante. Há setas vermelhas em vários pontos dos gráficos. Elas existem para lembrar que os dados em cada instante recebem, além do percentual de continuação no tópico, transições oriundas de outros tópicos. Os fluxos que contabilizaram menos de 2% foram suprimidos para efeito de visualização das principais características. Portanto, mesmo que não haja na figura uma transição de um tópico a outro, existe a probabilidade dessa transição mas ela é menor que 2%. Um exemplo do resultado dos fluxos é apresentado na Figura 4.20. Os demais gráficos dessa análise podem ser vistos no Apêndice A.

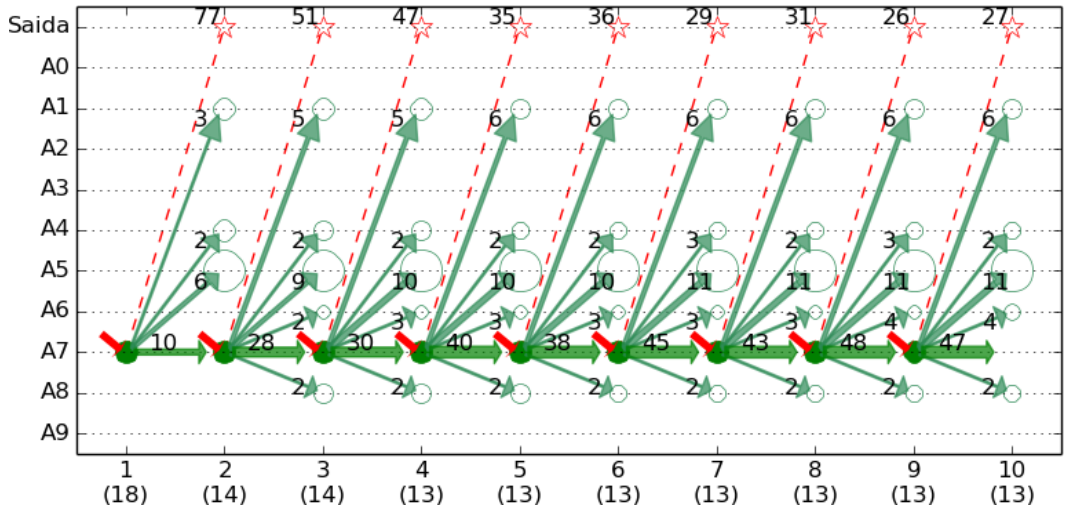


Figura 4.20: Exemplo da análise de Fluxo. No caso, o fluxo de transições centrado no tópico `A7`.

Em todos os tópicos do **Jornal Online A**, o fluxo de saída (fim de sessão) é bastante alto nas primeiras leituras e diminui a medida que os instantes de leitura aumentam. A distribuição das probabilidades dos fluxos vai se estabilizando com o passar das leituras. As probabilidades vão convergindo a certos valores com o passar do tempo. Cada tópico possui um grupo de outros tópicos na qual faz transições mais frequentes a esses. Esses tópicos são os destacados nos gráficos da análise top 70 padrões de trajetória. Os gráficos de ambas análises se encontram no Apêndice A para melhor visualização.

Os fluxos dos tópicos do **Jornal Online B** para a saída da sessão também são elevados. Os valores decaem com o passar do tempo mas são sempre os mais altos. A segunda maior probabilidade nos fluxos é normalmente a probabilidade de permanência no tópico.

Resumindo os resultados de fluxos e trajetórias, podemos dizer que a frequência de saída é alta em ambas as bases de dados e que ela diminui com o passar das leituras. O decaimento da probabilidade de saída com o passar dos instantes no **Jornal Online A** é maior do que no **Jornal Online B**. Os valores das probabilidades de transição entre tópicos variam bem ao longo dos instantes avaliados mas não mudam de ordem. A transição mais verossímil no início continua até o último instante analisado sendo a mais verossímil. Isso vale para todas as transições entre tópicos.

Nesse capítulo, foram apresentados os resultados das análises exploratórias dos dados. Nas primeiras análises, constatamos que as leituras no geral são normalmente rápidas, durando entre 1 e 3 minutos na média e que a primeira leitura é sempre a mais demorada. No **Jornal Online A** há um tópico que recebe muitas transições para ele. No **Jornal Online B**, não há esse padrão. Os usuários estão permanecendo poucas leituras num mesmo tópico, comportamento presente em ambos jornais. Contudo eles voltam com muita frequência a tópicos já visitados.

Uma última característica da base do **Jornal Online A** é que 59% das leituras a partir da segunda leitura apresentaram o mesmo tópico que o tópico da primeira leitura. Esse índice geral mostra que o tópico da primeira leitura está direcionando as demais leituras para esse mesmo tópico. Porém como vimos em algumas das análises, esse direcionamento não acontece só nas leituras que sucedem imediatamente, mas ao longo de toda a sessão, com voltas a esse tópico depois de mudança para outro tópico.

No caso do **Jornal Online B**, observamos que 43% das demais leituras apresentaram o mesmo tópico que o da primeira leitura, ou seja, 57% das leituras que sucedem a primeira são de outro tópico. Comparado com o índice do **Jornal Online A**, temos que os usuários desse jornal fizeram sessões mais diversificadas.

Agora que sabemos como os usuários se comportam, mostraremos no próximo capítulo quais os modelos estocásticos que conseguem capturar e prever bem os tópicos das leituras dos usuários.

Capítulo 5

Análise Experimental dos Modelos

"Essentially, all models are wrong, but some are useful."

— George E. P. Box

Neste capítulo, descreveremos os resultados dos experimentos com os modelos estocásticos propostos no Capítulo 2. Ao todo, foram testados 32 modelos. Eles foram comparados de acordo com o critério de informação de Akaike e com o escore de Brier. Todos os experimentos foram executados seguindo uma metodologia de avaliação baseada numa **validação cruzada de 5 partes** (*5-fold cross-validation*). Cada partição continha 80% das sessões para treino e 20% delas para teste. Cada sessão foi selecionada uma vez para teste, e quatro vezes para treino. A ordem das leituras em cada sessão foi preservada.

Alguns modelos permitem variações dependendo dos parâmetros utilizados. Testamos alguns valores para esses parâmetros e apresentaremos os resultados de cada variação separadamente, com exceção dos modelos de vantagem cumulativa e de permanência geométrica que utilizam um vetor de bônus β (modelos 2.20, 2.21 e 2.26). Para esses últimos modelos, só apresentamos os resultados obtidos com o vetor de melhor desempenho em cada modelo. Fizemos uma busca exaustiva em uma grade de valores para os parâmetros do bônus e selecionamos aqueles valores que geram o melhor resultado por modelo e por base de dados. Estes valores são diferentes para cada modelo. A Tabela 5.1 mostra os melhores β encontrados e que foram utilizados na avaliação desses modelos. β_A é o vetor de bônus maximizado na base do **Jornal Online A**, e β_B é o vetor de bônus maximizado na base do **Jornal Online B**.

Todos os modelos necessitam especificar um certo número de probabilidades. Essas probabilidades são estimadas a partir dos dados reais e nós limitamos a estimação aos instantes $n \leq 10$. Em alguns modelos a quantidade de probabilidades a ser estimada aumenta com os instantes avaliados e os dados ficam cada vez mais escassos com o passar dos instantes. Logo, se uma sessão continha mais de 10 artigos lidos, foram

Modelo Vantagem Cumulativa A (2.20)
$\beta_A = [2,0;\ 1,0;\ 1,0;\ 2,0;\ 2,0;\ 2,0;\ 1,0;\ 0,8;\ 1,0;\ 0,001]$
$\beta_B = [1,0;\ 1,0;\ 0,8;\ 2,0;\ 0,5;\ 0,4;\ 0,7;\ 0,5;\ 0,5;\ 0,3]$
Modelo Vantagem Cumulativa B (2.21)
$\beta_A = [10,0;\ 5,0;\ 10,0;\ 10,0;\ 10,0;\ 2,0;\ 10,0;\ 4,0;\ 10,0;\ 0,001]$
$\beta_B = [10,0;\ 4,0;\ 10,0;\ 5,0;\ 4,0;\ 10,0;\ 5,0;\ 4,0;\ 4,0;\ 10,0].$
Modelo Permanência Geométrica C (2.26)
$\beta_A = [0,2;\ 0,4;\ 0,3;\ 0,3;\ 0,5;\ 0,6;\ 0,3;\ 0,4;\ 0,4;\ 0,0]$
$\beta_B = [0,1;\ 0,3;\ 0,4;\ 0,4;\ 0,3;\ 0,2;\ 0,4;\ 0,3;\ 0,5;\ 0,2]$

Tabela 5.1: Os vetores de bônus maximizados.

consideradas nas avaliações dos modelos somente as 10 primeiras leituras. Na Seção 3.3, apresentamos a distribuição dos tamanhos das sessões. Há diversas sessões longas, contendo mais de 10 leituras. Contudo, a quantidade de sessões decai rapidamente quanto maior for a quantidade de leituras. Por exemplo, no caso do **Jornal Online A**, 51% tem no mínimo 3 artigos, 28% tem no mínimo 4 artigos, chegando a aproximadamente 3% das sessões com no mínimo 10 artigos lidos. No caso do **Jornal Online B**, 30% contém no mínimo 3 artigos, 15% contém no mínimo 4 artigos. A quantidade de sessões com no mínimo 10 artigos é aproximadamente 0,5% do total.

Os modelos tiveram seus nomes abreviados para uma melhor apresentação nos gráficos. A Tabela 5.2 mostra as abreviações de cada modelo por grupo. Relembrando os grupo, temos: Os **modelos sem influência do passado** são os modelos onde a informação de tópicos prévios é desconsiderada. O grupo dos **modelos de memória curta** é formado pelos modelos em que apenas as leituras recentes afetam o futuro. Já o grupo dos **modelos de preferência revelada** é o grupo dos modelos que condicionam o futuro somente à característica de um tópico por vez. Os **modelos de permanência geométrica** modelam as sessões de leituras como períodos de permanência ou de mudança de tópico. O último grupo, dos **modelos de vantagem cumulativa**, assume que as leituras prévias dos tópicos aumentam as chances de suas leituras no futuro.

Considerando a estimação até o instante 10, a coluna *df* da Tabela 5.2 mostra os valores dos graus de liberdade (*degrees of freedom* – df) de cada um dos modelos. O número de graus de liberdade de um modelo é o número de parâmetros independentes que necessitam ser estimados a partir dos dados para instanciar o modelo. Cada modelo possui um número de parâmetros independentes em cada instante de tempo. A soma desses números nos 10 instantes possíveis nos forneceram os valores apresentados na tabela. Os cálculos estão descritos no Apêndice B.

Modelos sem influência do passado	Abreviação	df
M-Uniforme	M0-U	0
M-Independência	M1-I	90
M-Independência Homogênea	M1-IH	9
Modelos de memória curta	Abreviação	df
M-Alta Permanência (p=0.91)	M2-AP 91	0
M-Alta Permanência (p=0.55)	M2-AP 55	0
M-Markov-I	M2-M1	819
M-Markov-I Homogêneo	M2-M1H	99
M-Markov-II	M2-M2	7.299
M-Markov-II Homogêneo	M2-M2H	999
M-Markov-III Homogêneo	M2-M3H	9.999
M-Markov-IV Homogêneo	M2-M4H	99.999
Modelos de preferência revelada	Abreviação	df
M-Histórico Visitas	M3-S	549
M-Ultimas M Visitas (m=2)	M3-S m2	269
M-Ultimas M Visitas (m=3)	M3-S m3	339
M-Duração no Estado	M4-D	549
M-Duração no Estado Últimos M Artigos (m=2)	M4-D m2	269
M-Duração no Estado Últimos M Artigos (m=3)	M4-D m3	339
M-Duração da Última Visita	M5-L	549
M-Leituras Pós Saída	M6-E	549
Modelos de permanência geométrica	Abreviação	df
M-PG-A (θ e π variantes a cada instante)	M7-PG A-I	180
M-PG-A (θ variante a cada instante e π fixo)	M7-PG A-II	99
M-PG-A (θ e π fixos)	M7-PG A-III	18
M-PG-B (θ e π variantes a cada instante)	M7-PG B-I	180
M-PG-B (θ variante a cada instante e π fixo)	M7-PG B-II	99
M-PG-B (θ e π fixos)	M7-PG B-III	18
M-PG-C (θ e π variantes a cada instante)	M7-PG C-I	180
M-PG-C (θ variante a cada instante e π fixo)	M7-PG C-II	99
M-PG-C (θ e π fixos)	M7-PG C-III	18
Modelos de vantagem cumulativa	Abreviação	df
M-Vantagem Cumulativa A ($\pi(l)$ variante a cada instante)	M8-VC A	90
M-Vantagem Cumulativa A ($\pi(l)$ homogêneo)	M8-VC A-H	9
M-Vantagem Cumulativa B ($\pi(l)$ variante a cada instante)	M8-VC B	90
M-Vantagem Cumulativa B ($\pi(l)$ homogêneo)	M8-VC B-H	9

Tabela 5.2: Os 32 modelos, suas abreviações e o número de graus de liberdade. Entre parênteses, mostramos os valores de alguns parâmetros de ajuste (*tuning parameters*) utilizados nos modelos.

5.1 Critério de Informação de Akaike

Como dito na Seção 2.12, usamos o critério de informação de Akaike (**AIC**) como medida de comparação dos modelos. Quanto maior o AIC resultante, melhor o modelo (Akaike [1998]). Após a experimentação da validação cruzada, uma média das 5 partições foi calculada para cada modelo e as discussões a seguir são guiadas por esses resultados. Os valores de cada partição não diferem muito da média e os desvios padrão são da ordem de 1% do valor das médias. Logo, não há valores em uma partição que difiram muito dos demais alterando o ranking final dos modelos. As tabelas completas com os valores das 5 partições em separado e a tabela com os valores da média e o desvio padrão de cada modelo podem ser vistas no apêndice C.

O Gráfico 5.1 apresenta os resultados do AIC dos modelos ajustados aos dois jornais. Os resultados foram plotados em conjunto para mostrar como os modelos se comportam dependendo do jornal. Como a base de dados do **Jornal Online B** possui praticamente um terço do volume da base do **Jornal Online A**, os valores de AIC estão em níveis diferentes. Já, os dois gráficos seguintes (5.2 e 5.3) mostram o ranking dos modelos pelo AIC nas bases de dados em separado.

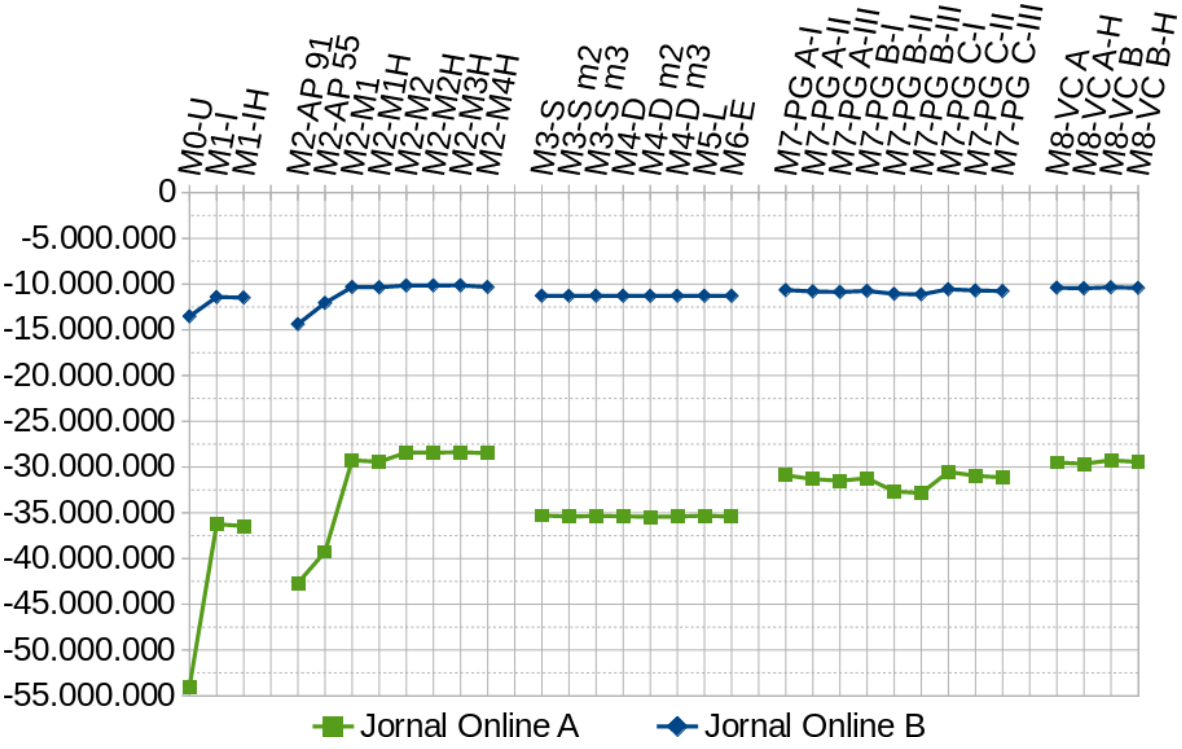


Figura 5.1: Resultado do AIC dos modelos com validação cruzada de 5 partes. Os modelos estão conectados por grupos: 1º grupo modelos sem influência do passado, 2º grupo modelos de memória curta, 3º grupo modelos de preferência revelada, 4º grupo de permanência geométrica e 5º grupo de vantagem cumulativa.

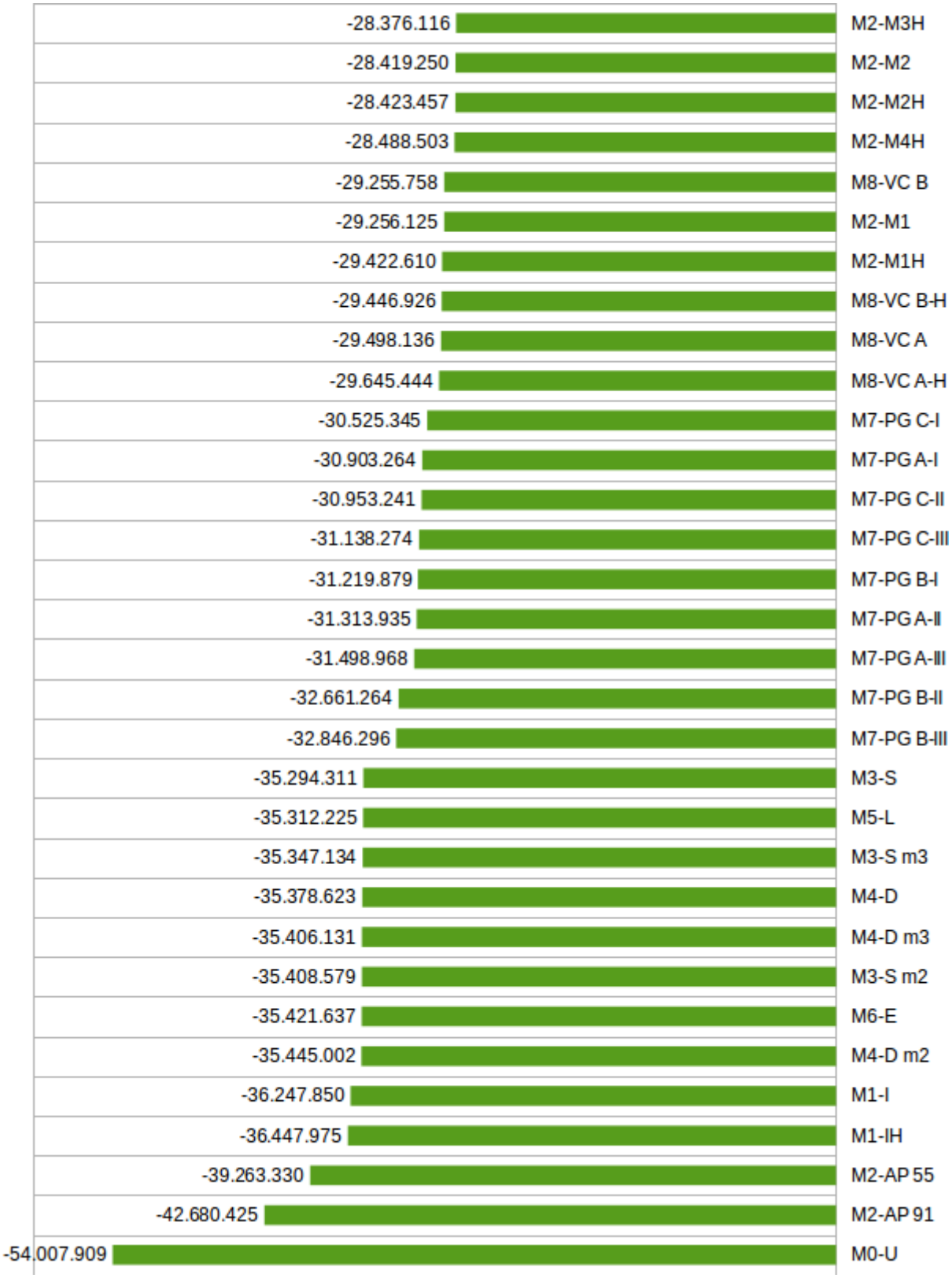


Figura 5.2: Ranking dos modelos pelo AIC médio, base do **Jornal Online A**.

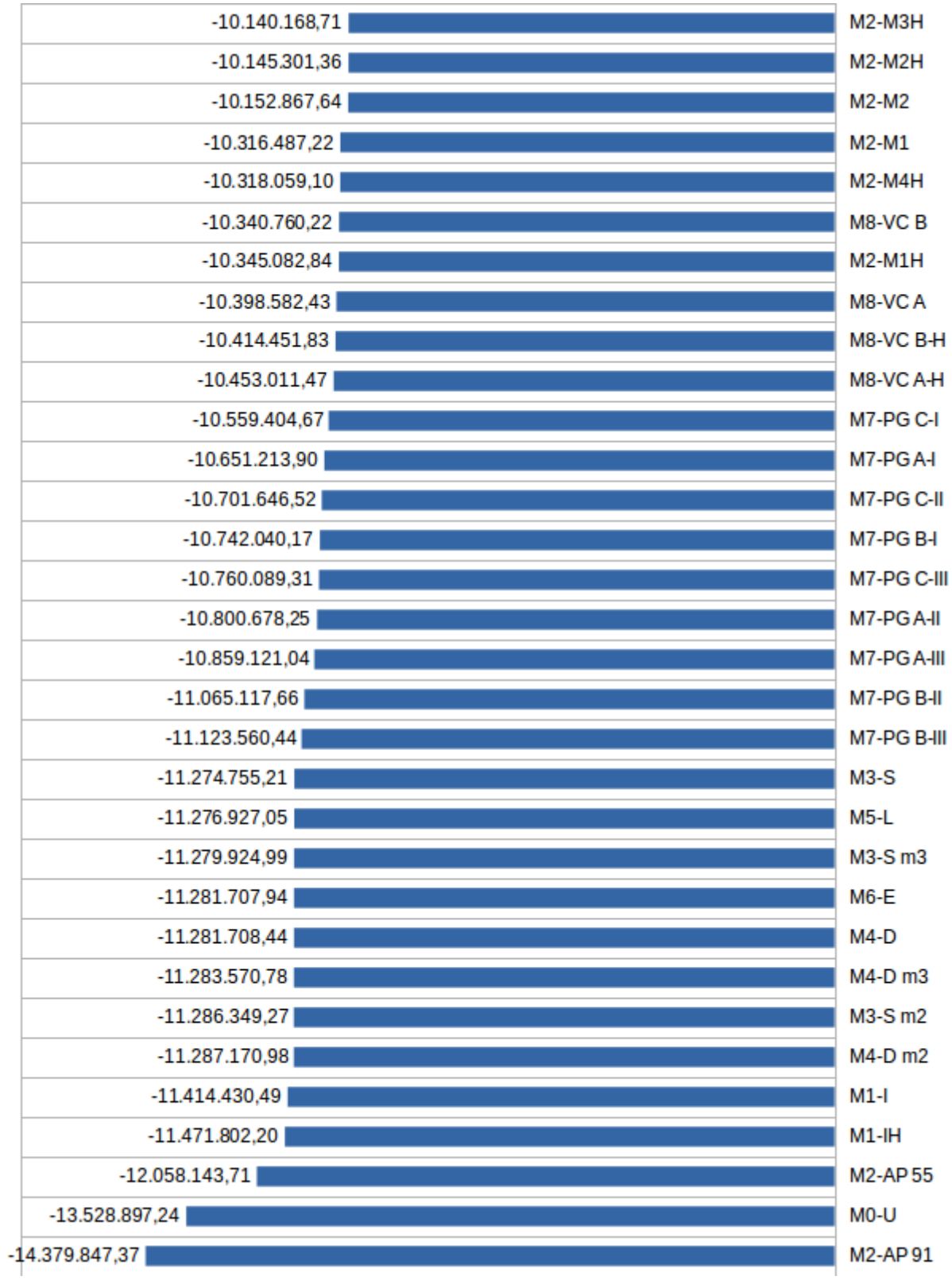


Figura 5.3: Ranking dos modelos pelo AIC médio, base do **Jornal Online B**.

O comportamento dos modelos é praticamente o mesmo em ambos os jornais, observe a Figura 5.1. Temos que os modelos mais ingênuos (**M0-U**, **M2-AP 91**, e **M2-AP 55**) são os que possuem os piores valores de AIC. O modelo **M0-U** obteve resultado pior que o modelo **M2-AP 91** na base do **Jornal Online A**, significando que assumir probabilidades iguais para todos os tópicos é pior que assumir uma permanência alta nesse jornal. O modelo **M2-AP 55** obteve melhores resultados que o modelo **M2-AP 91**, somente por assumir que a probabilidade de permanência é menor. Os modelos de independência (**M1-I** e **M1-IH**) são melhores que os três anteriores, mostrando que assumir independência é melhor que assumir uniformidade ou permanência simples.

Os modelos Markovianos de memória curta são os melhores modelos pelo AIC. Eles obtiveram os maiores valores de AIC em ambas as bases de dados. Há crescimento nos resultados com o aumento das ordens dos modelos: dos modelos de primeira ordem (**M2-M1**, **M2-M1H**) para o de segunda (**M2-M2**, **M2-M2H**), e de segunda para terceira ordem (**M2-M3H**). Porém, o modelo Markoviano de quarta ordem (**M2-M4H**) não manteve essa melhora nos resultados. Ele não é melhor que o modelos de segunda e terceira ordem. Esse comportamento pode ser melhor visto nos rankings das Figuras 5.2 e 5.3. No geral, os valores de AIC destes modelos nos mostram que modelos que guardam informação apenas do passado mais recente são normalmente os melhores modelos. A comparação entre eles mostrou que o tamanho deste passado pode ser pequeno: as 2 ou 3 últimas leituras são suficientes para fornecer bons resultados.

Os modelos de preferência revelada (**M3-S**, **M3-S m2**, **M3-S m3**, **M4-D**, **M4-D m2**, **M4-D m3**, **M5-L** e **M6-E**) possuem valores próximos de AIC entre si, e se alternam nas posições da classificação final comparando ambos os jornais. Estão todos acima dos modelos sem influência do passado, sendo melhores que o dois piores modelos de memória curta (**M2-AP 91**, e **M2-AP 55**), mas não superam o pior modelo Markoviano (**M2-M1H**). Esse resultado possivelmente se deve a simplicidade dos modelos em só observar características de um tópico por vez e por desprezar a ordem dos acontecimentos no histórico.

Os modelos de permanência geométrica obtiveram melhores resultados do que os modelos sem influência do passado, os modelos **M2-AP 91** e **M2-AP 55** e os modelos de preferência revelada. Esses modelos que assumem uma permanência geométrica e mudança de tópico por funções específicas, possuem queda no valor AIC quanto menos dados são estimados. Por exemplo, os valores do AIC decaem nessa ordem: **M7-PG A-I** > **M7-PG A-II** > **M7-PG A-III**. Sendo que o primeiro modelo adota valores diferentes de θ e ρ instante a instante, o segundo adota somente θ variante e o último adota as duas probabilidades constantes para todo instante. Quanto mais dados são utilizadas nos modelos, melhores são seus resultados. Se os compararmos pela

função que utilizam, os modelos **M7-PG C-**x são em geral melhores que os respectivos modelos **M7-PG A-**x e **M7-PG B-**x (x denota as variações I, II e III).

O último grupo, grupo dos modelos de vantagem cumulativa, é o segundo melhor grupo pelo valor de AIC. Todos os modelos desse grupo estão acima dos modelos sem influência do passado, de preferência revelada e de permanência geométrica. O modelo **M8-VC B** consegue superar o modelo **M2-M1H** em ambas as bases e o modelo **M2-M1** na base do **Jornal Online A**. Novamente entre eles quando assumimos probabilidades homogêneas os valores de AIC diminuem um pouco: **M8-VC A** > **M8-VC A-H** e **M8-VC B** > **M8-VC B-H**.

Pelo ranking gerado pelo AIC o modelo escolhido para ser utilizado em um sistema de recomendação seria o modelo Markoviano de 3ª ordem homogêneo (**M2-M3H**). Esse modelo mesmo assumindo que as probabilidades são iguais para os instantes $n \geq 4$ ficou melhor que os modelos de ordem menor. O modelo de 4ª ordem homogêneo (**M2-M4H**) ficou pior que os modelos de 3ª e 2ª ordem em ambos as bases de dados, mesmo considerando um passado mais longo que os demais.

Observe que o modelo (**M2-M3H**) possui praticamente 10 mil graus de liberdade ($df = 9999$), um grande número de parâmetros. Poderíamos escolher o modelo Markoviano de segunda ordem que, na versão simplificada (**M2-M2H**), possui uma ordem a menos de graus de liberdade ($df = 999$) e possui índice próximo ao do modelo de terceira ordem. Porém se o custo desse cálculo ainda for alto, ou a manutenção dessas probabilidades pelo recomendador for inviável, o modelo de vantagem cumulativa **M8-VC B** é uma alternativa pois tem índices próximos aos dos melhores modelos com um custo bem menor ($df = 90$).

Comparando os dois modelos: o modelo **M2-M2H** possui função de complexidade $O(L^3)$ e o modelo **M8-VC B** $O(L)$. Com L sendo o tamanho do conjunto dos tópicos.

5.2 Escore de Brier

Rodamos o escore de Brier sobre os dados de ambos os jornais em separado. O esquema de treino e teste dos modelos foi novamente a validação cruzada de 5 partes. Contudo, inicialmente analisamos o escore de Brier instante a instante separadamente. Calculamos o escore de Brier dos instantes $3 \leq n \leq 10$. Dessa forma podemos ver se a qualidade de predição dos modelos varia com a quantidade de informação do histórico da sessão do usuário. Quando $n = 3$, o modelo só tem a informação de 2 tópicos no

histórico e quando $n = 10$, o modelo tem informação de 9 tópicos no histórico. É de se esperar que a quantidade de histórico, caso o modelo utilize essa informação, altere as probabilidades calculadas e possivelmente o erro também.

As Figuras 5.4 e 5.5 mostram os resultados do escore de Brier de todos os modelos em ambas as bases de dados. Em cada figura há dois valores por modelo, conectados por duas linhas. A linha de cor mais escura representa a média dos escores de Brier, média tomada sobre todos os instantes de tempo. A outra linha, de cor mais clara, representa a média ponderada dos escores, com um peso associado ao volume de dados do instante n.

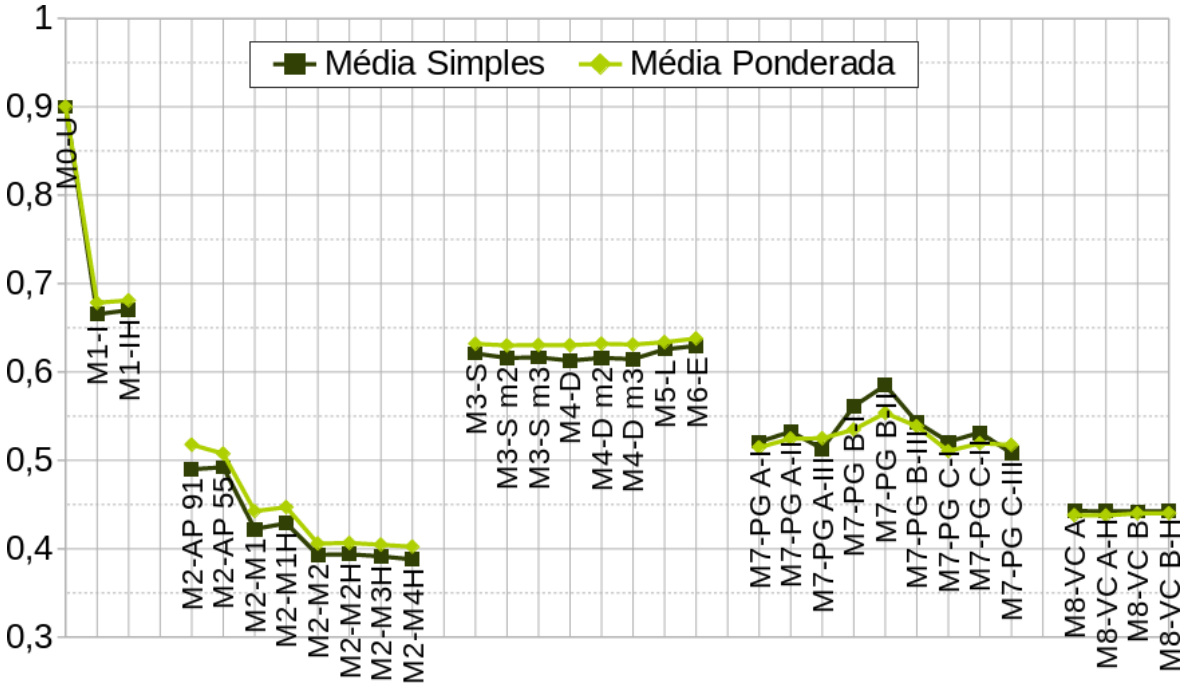


Figura 5.4: Escore de Brier dos modelos na base de dados do **Jornal Online A**. Os modelos estão conectados por grupos: 1º grupo modelos sem influência do passado, 2º grupo modelos de memória curta, 3º grupo modelos de preferência revelada e 4º grupo de permanência geométrica.

Em ambas as bases de dados, quanto maior o instante n avaliado, menor é o volume de dados. Assim, a média ponderada fornece maior peso para os instantes de maior volume de dados. Comparando as duas médias podemos observar certos comportamentos. Quando a média ponderada estiver acima da média simples, temos o caso de que quanto maior o instante avaliado (e menos dados), menor foi o erro resultante do escore. Porém quando a média ponderada estiver abaixo da média simples, temos a situação inversa: erro maior nos instante superiores. Os resultados avaliados instante a instante podem ser vistos no Apêndice D.

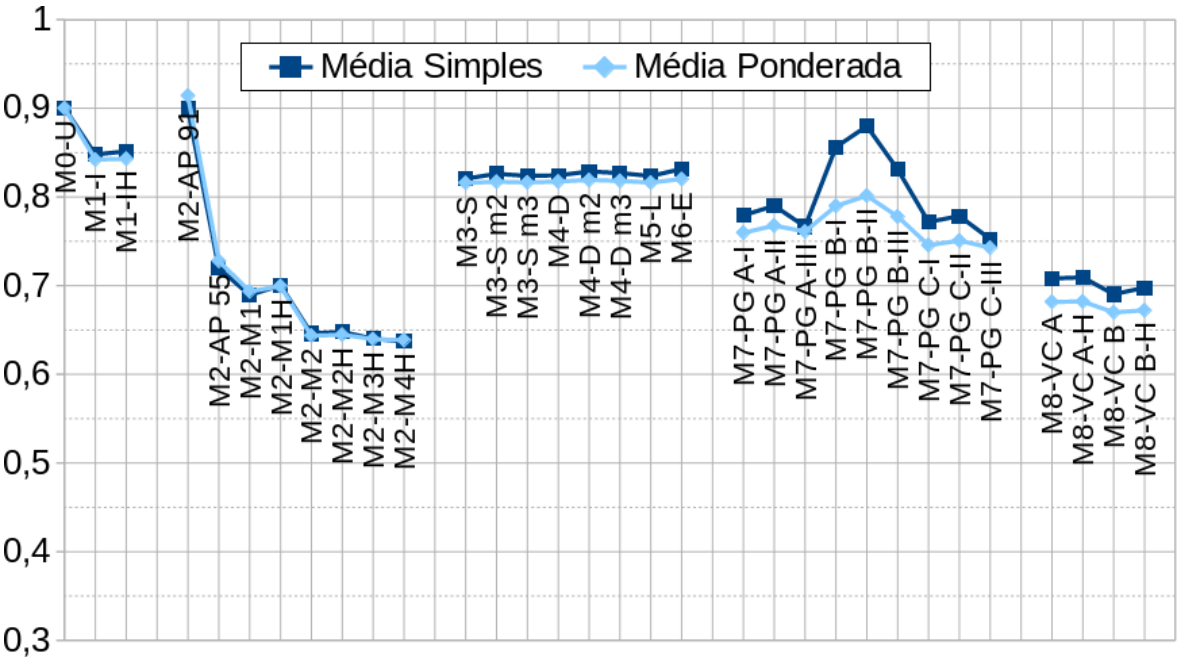


Figura 5.5: Escore de Brier dos modelos na base de dados do **Jornal Online B**. Os modelos estão conectados por grupos: 1º grupo modelos sem influência do passado, 2º grupo modelos de memória curta, 3º grupo modelos de preferência revelada e 4º grupo de permanência geométrica.

Os modelos têm praticamente o mesmo comportamento nas duas bases de dados. Novamente, os melhores modelos são os de memória curta, todos inclusive os modelos **M2-AP 91** (só na base do **Jornal Online A**) e **M2-AP 55** (em ambas as bases).

Os modelos de vantagem cumulativa formam o segundo melhor grupo de resultados. Eles se aproximam bem dos melhores modelos. Entre eles, não há muita diferença na base do **Jornal Online A**, com os escores de Brier muito similares. Na base do **Jornal Online B** o modelo **M8-VC B** obteve resultado claramente melhor que os demais modelos do seu grupo. Contudo, as médias simples estão acima das médias ponderadas de todos os modelos desse grupo, comportamento fraco na Figura 5.4 mas evidente na Figura 5.5. Essa é uma evidência que esses modelos são normalmente melhores nos instantes iniciais das sessões.

Os modelos de permanência geométrica foram novamente melhores que os modelos de preferência revelada e os sem influência do passado, no geral. Porém há uma variação considerável entre os modelos **M7-PG B-I**, **M7-PG B-II** e **M7-PG B-III**, observando as duas médias. Esses modelos também possuem a média ponderada abaixo da média simples. Logo, eles são melhores nos instantes iniciais e começam a errar mais no decorrer dos instantes. Os demais modelos desse grupo compartilham desse comportamento, mas de forma menos acentuada.

Os modelos de preferência revelada novamente possuem valores similares entre si e foram melhores somente aos modelos sem influência do passado. O valor do erro alto desses modelos reforça a hipóteses de que o foco em somente um tópico por vez na função que define os aspectos mais relevantes das sessões é algo ruim.

Uma constatação no escore de Brier diferente dos resultados do AIC foi que os modelos **M2-AP 55** (em ambas as bases) e **M2-AP 91** (somente na primeira base) obtiveram escores bons, melhores que todos os modelos de preferência revelada e de permanência geométrica. Ambos os modelos são modelos estacionários que atribuem uma alta probabilidade p para o caso de permanência e a probabilidade complementar a 1 é dividida igualmente como a probabilidade de mudança para os demais tópicos. O escore de Brier considera que esses dois modelos erram pouco, com valores próximos a 0,5 e a 0,7. Esse resultado é totalmente justificado pela característica das bases de dados de alta permanência no geral, vide Seções 4.4 e 4.5.

No Apêndice D.1 conferimos os resultados desses dois modelos em especial e mostramos como realmente eles possuem valores baixo de erro pelo escore de Brier. O índice de permanência é alto no **Jornal Online A**. Esse fato permitiu que ambos os modelos tivessem resultados bons nessa base. Já a base do **Jornal Online B** possui um índice de permanência razoável, menor que o do primeiro jornal. Esse índice menor fez com que o modelo que coloca a maior probabilidade de permanência errasse mais nessa base. O retrocesso do resultado do **M2-AP 91** na base do **Jornal Online B**, ficando até pior que o modelo **M0-U**, mostra instabilidade nesse tipo de modelo. Logo, acreditamos que o modelo **M2-AP 55** não se sairá tão bem em qualquer base de dados como os modelos Markovianos e os modelos de vantagem cumulativa apresentam ser.

Novamente, nossa escolha para o modelo ganhador estaria entre **M2-M2H** e **M8-VC B**. Esses são os modelos que possuem os menores custos dentre os melhores modelos identificados. Os modelos **M2-M4H** e **M2-M3H** são ligeiramente melhores que os escolhidos, mas possuem complexidade bem maiores.

Capítulo 6

Considerações Finais

Nesta dissertação, estudamos o comportamento de usuários em leituras sucessivas de artigos de notícias online. Foram analisadas mais de 20 milhões de sessões compostas por *clicks* sucessivos dos usuários, em notícias postadas em dois jornais online (o **Jornal Online A** e o **Jornal Online B**).

Inicialmente, foram estudadas características das bases tais como a quantidade de artigos por sessão, o tempo médio entre leituras, a popularidade dos tópicos dos artigos, a quantidade média de tópicos por sessão e a permanência e transição entre tópicos. Essas análises identificaram algumas características marcantes do comportamento dos leitores. As principais foram as seguintes:

- A base de leitores do **Jornal Online A** possui uma maior porcentagem de leitores assíduos do que a base de leitores do **Jornal Online B**. Dos usuários do **Jornal Online A**, 26% geraram mais de uma sessão, totalizando 63% das sessões desse jornal. No entanto, somente 20% dos usuários do **Jornal Online B** geraram mais de uma sessão, totalizando 45% das sessões dessa base. (Seção 3.4);

- Não houve viés de leitura para os tópicos mais publicados. Os tópicos mais lidos não são os tópicos com maior número de artigos disponíveis. Esse comportamento está presente em ambas as bases. Isso deveria servir para reorientar a produção de artigos de maneira a torná-la mais adequada ao consumo preferido dos leitores. No **Jornal Online A**, embora 50% dos artigos lidos sejam de *Famosos*, apenas 3% dos artigos disponíveis cobrem esse tópico. Aparentemente, um maior esforço para gerar notícias desse tópico poderia atrair mais sessões de leitura dos leitores atuais. (Seção 3.6);

- Os usuários do **Jornal Online B** leram mais diversificadamente (tópicos diferentes em uma mesma sessão) comparados com os usuários do **Jornal Online A**.

Na Seção 3.7, ao compararmos a quantidade de tópicos distintos para as sessões de tamanho 10, podemos ver que os valores normais do **Jornal Online A** estão distribuídos entre 2 e 3 tópicos enquanto os do **Jornal Online B** ficam entre 3 e 6 tópicos;

- Os usuários do **Jornal Online A** gastaram mais tempo para ler os artigos do que os usuários do outro jornal. A transição entre artigos no **Jornal Online A** é entre 1,5 a 2 minutos em média, contra 2 a 2,5 minutos em média para o **Jornal Online B**. (Seção 4.1);

- Os usuários de ambos os jornais demoraram mais nas leituras dos primeiros artigos das sessões. Contudo, os usuários do **Jornal Online A** leram cada vez mais rápido com o passar das leituras, enquanto os usuários do **Jornal Online B** oscilaram no tempo gasto nas demais leituras. (Seção 4.1);

- Há indícios da necessidade de foco em poucos tópicos em uma sessão para capturar bem o comportamento de leitura. Os principais padrões de trajetória apresentam poucas mudanças de tópicos nas duas bases de dados. A soma dos padrões que representam cerca de 95% das sessões contém no máximo 3 mudanças de tópicos (Tabela 4.1). Porém, os padrões mais frequentes na base de dados continham somente 2 tópicos *distintos* por sessão, mesmo havendo mais de duas mudanças de tópico. Em geral, essas mudanças se comportam como um ciclo entre dois tópicos. Contudo, não há uma única dupla de tópicos dominando todos os tópicos possíveis. Vários arranjos de tópicos se apresentam como prováveis. (Seções 4.6 e 4.7);

- Ao comparar o tópico da primeira leitura com os tópicos das demais leituras, observamos que no **Jornal Online A** 59% das demais leituras apresentaram o mesmo tópico que o da primeira leitura. No caso do **Jornal Online B**, temos um índice menor: 43% das leituras seguintes são do mesmo tópico que o da primeira leitura. (Seção **??**);

- Ambos os jornais possuem alto índice de retenção nos tópicos. Quando comparamos as leituras sucessivas que os usuários fizeram, observamos que uma grande parte das leituras se mantêm no mesmo tópico que o da leitura anterior. Os valores no **Jornal Online A** são 67% para permanência no tópico e 33% para mudança. No caso do **Jornal Online B**, temos um índice menor: 46% para permanência e 54% para mudança. Logo podemos ver que os usuários do **Jornal**

Online A permanecem mais em um tópico que os usuários do **Jornal Online B**. (Seção 4.4);

- Os usuários do **Jornal Online A** fazem em geral sessões maiores em número de leituras do que os usuários do outro jornal. Normalmente as sessões dos usuários do primeiro jornal são de 5 a 7 leituras contra sessões de 3 a 4 leituras dos usuários do **Jornal Online B**. Assim, foi identificado um número maior de padrões de trajetória no **Jornal Online A** do que no **Jornal Online B**. Entretanto, ao observar os padrões de trajetória condicionados pelo tópico inicial, vemos que as mudanças de tópicos nos padrões de trajetória do **Jornal Online A** têm menos variabilidade de tópicos. Os padrões começam em um determinado tópico e mudam para menos de 5 dos 9 tópicos disponíveis. Já os padrões do **Jornal Online B** tendem a mudar em média para 7 dos 9 possíveis outros tópicos, formando padrões mais variados. (Seção **??**).

Após a análise inicial, foram estudados 32 modelos estocásticos para recomendação de próximo tópico de leitura, cada um deles procurando capturar a essência do comportamento do usuário. Alguns desses modelos foram inspirados nos resultados das análises exploratórias. O modelo de independência foi pensado após o experimento da Seção 4.3. Os modelos Markovianos foram motivados pelo experimento da Seção 4.4. Os modelos de alta permanência foram inspirados pelos índices de permanência obtidos na mesma Seção 4.4. Os modelos de permanência geométrica foram elaborados após os resultados dos experimentos das Seções de 4.5 a **??**.

Todos os modelos estocásticos foram ajustados por máxima verossimilhança e comparados com base no critério de informação de Akaike e no escore de acurácia de predição de Brier. Os melhores modelos são aqueles em que o usuário move-se pelos tópicos influenciado apenas pelas suas leituras mais recentes, os modelos de memória curtaMarkovianos. Os modelos de vantagem cumulativa ficaram logo atrás com valores satisfatórios, mostrando que as primeiras escolhas influenciam sim as escolhas futuras. Em seguida vêm os modelos de permanência geométrica e de preferência revelada. Já os modelos sem influência do passado foram os que obtiveram os piores índices.

Os dois principais modelos, o de Markov com ordem pequena e o de vantagem cumulativa, parecem capturar aspectos muito distintos e até contraditórios entre si. Afinal, se as primeiras leituras impactam o futuro mais longínquo, como um modelo que considera apenas o passado mais recente pode ser também um bom descritor dos mesmos dados? Não temos uma explicação para esse fato. Porém há algumas hipóteses que poderiam justificá-lo. Primeiramente a possibilidade do fato só ocorreu por causa

das sessões serem normalmente de poucas leituras. Outra hipotese é que os modelos são complementares, possuem característica complementares.

Todos os modelos que utilizam de dependência do passado com ordem (os de memória curta e os de permanência geométrica) são melhores que os que desprezam a ordem do passado ou a sua existência (os de preferência revelada e os sem influência do passado, respectivamente). Em geral, os modelos tiveram resultados de predição melhores no escore de Brier na base do **Jornal Online A** do que no **Jornal Online B**. Os menores erros de predição obtidos no **Jornal Online A** possivelmente se devem a dois aspectos desse jornal já mencionados: usuários mais assíduos e tópicos com transições para poucos tópicos diferentes.

Apesar do trabalho envolver bases de dois jornais online específicos, os modelos podem ser utilizados em qualquer outro jornal online que utilize tópicos únicos para categorizar os artigos. Uma granularidade maior de tópicos podem ser explorados, e para recomendação na prática seriam mais viáveis. Contudo quanto maior o número de tópicos L, pior é a fase de estimar certos modelos.

Dentre todos os modelos testados, o modelo Markoviano de segunda ordem e o de terceira ordem são as melhores escolhas ponderadas pelos índices nas duas métricas estudadas apesar do alto índice de graus de liberdade. Contudo, pode haver melhores modelos que não foram estudados ainda. Assim, um trabalho futuro seria estudar outros modelos, como o modelo de urna de Pólya (Blackwell & MacQueen [1973]) ou os modelos Markovianos de cadeia de memória variável (Bühlmann [2000]). Outra proposta seria mesclar características dos modelos já estudados, por exemplo adicionar aos modelos de preferência revelada a informação de ordem dos k-últimos tópicos. Outros trabalhos futuros incluem:

- Utilizar LDA para caracterização dos tópicos. Não utilizamos LDA neste trabalho por consideramos os rótulos dos *groups* das notícias informativos e acertados. Como eles não eram sempre únicos, optamos por avaliá-los manualmente e criar os 10 tópicos finais de cada base.

- Experimentar modelos totalmente personalizados por usuários, modelos específicos por usuários ou por grupos de usuários que compartilham perfis de leituras similares. No Capítulo 2, assumimos que as probabilidades $\mathbb{P}(T_{u,1} = l), \text{com } l \in \mathcal{L}$ e $\mathbb{P}(T_{u,i} = l_i \mid T_{u,1} = l_1,\ ..,\ T_{u,i-1} = l_{i-1}), \text{com } l_i \in \mathcal{L}$ e $1 > i \geq n_u$ são aplicáveis a todos os usuários e sessões, ou seja, são probabilidades independentes do usuário u que fez a sessão. Para experimentar modelos totalmente personalizados por usuários, podemos selecionar os usuários que fizeram mais de uma sessão, treinando os modelos com algumas de suas sessões e testando com outras. Porém o

volume de dados se reduz drasticamente. Na Seção 3.4 constatamos que o percentual de usuários que fizeram mais de uma sessão é baixo em ambos os jornais. Os usuários que fizeram no mínimo 10 sessões (quantidade mínima razoável de sessões para gerar modelos) são 2,5% dos usuários do **Jornal Online A** e 0,6% dos usuários do **Jornal Online B**. Com esses índices se reduz muito a quantidade de dados para experimentação. Logo, uma alternativa seria identificar usuários com perfis de acessos similares, e treinar modelos para cada perfil separadamente. Porém essa abordagem gera o problema de identificar o perfil para escolher qual modelo ajustado utilizar.

- Utilizar outro esquema teste/treino, observando a ordem temporal dos dados, para testes de adequação dos modelos quanto ao surgimento das notícias e o histórico dos usuários a decorrer dos dias. Os experimentos que fizemos não observaram a ordem cronológica das sessões por se tratar de testes *offline*. Porém, poderíamos treinar os modelos simulando um teste *online*, onde somente as sessões feitas pelos usuários antes daquele momento estão presentes no histórico de sessões utilizado para ajustar os modelos.

- Testar se a variável intervalo de leitura adiciona informação útil aos modelos melhorando as predições. Os usuários leem de forma variada as notícias, ora rápido, ora demorando mais na leitura. Poderíamos tentar identificar se existem padrões entre tópicos e a quantidade de tempo que usuário gastou nas leituras prévias. Também poderíamos testar se adotando as leituras super rápidas como leituras não relevantes, altera consideravelmente o padrão de leituras. Nesse caso, os modelos não deveriam utilizar essas sessões para ajuste das suas probabilidades, ou pelo menos as leituras não relevantes.

- E por último, o trabalho futuro mais almejado seria implementar de fato um sistema de recomendação que utilize todas as informações acerca dos hábitos dos leitores de jornais online. Já identificamos algumas características dos usuários de jornais online que podem ser utilizadas em sistemas de recomendação, como listado acima. Porém, implementar, testar, analisar e por em prática um sistema de recomendação é um trabalho longo e que não foi contemplado nesta dissertação.

Através desta dissertação, mostramos que os usuários de jornais onlines se comportam de forma diferente dependendo do jornal. Apesar de comportamentos diversificados, modelos estocásticos que capturam o comportamento recente são bem úteis para modelar as preferências dos usuários ao longo de uma sessão.

Apêndice A

Fluxos de transições e Principais Trajetórias por Tópico

Os gráficos das análises de **fluxo de transições** e dos **top 70 padrões de trajetórias** são mostradas conjuntamente para cada tópico a seguir. Primeiramente são mostrados os resultados dos tópicos do **Jornal Online A** e depois os resultados dos tópicos do **Jornal Online B**.

Nos gráficos da análise de **fluxo de transições** temos no eixo X ordem da leitura e no eixo Y os rótulo dos 10 tópicos mais o rótulo `Saída` que denota o término da sessão. Abaixo dos rótulos do eixo X, há valores entre parênteses que mostram a porcentagem das leituras daquele tópico, instante a instante. Também há setas vermelhas em vários pontos dos gráficos para lembrar que os dados em cada instante recebem, além do percentual de continuação no tópico, transições oriundas de outros tópicos. Os fluxos que contabilizaram menos de 2% foram suprimidos para efeito de visualização das principais características.

Abaixo dos gráficos de fluxo de cada tópico, temos os resultados da análise dos **top 70 padrões de trajetórias**. Nessa segunda análise identificamos as trajetórias de sessões que começam em um determinado tópico e somam no mínimo 70% dos casos dessas sessões. Como em alguns casos encontramos muitos padrões, os resultados foram divididos em duas figuras. A primeira contém os top-60 padrões, as trajetórias de maior porcentagem que juntos somam 60%. E a segunda tem as trajetórias seguintes que, junto com as primeiras, atingem 70%. Os padrões de maior frequência são os vermelhos e os de menor frequência os roxos. Na legenda, entre colchetes, há uma sequência de números que designam as transições entre tópicos (os números representam os tópicos). Entre parênteses, colocamos o percentual de vezes que aquele padrão acontece na base.

Tópico A0 – Jornal Online A

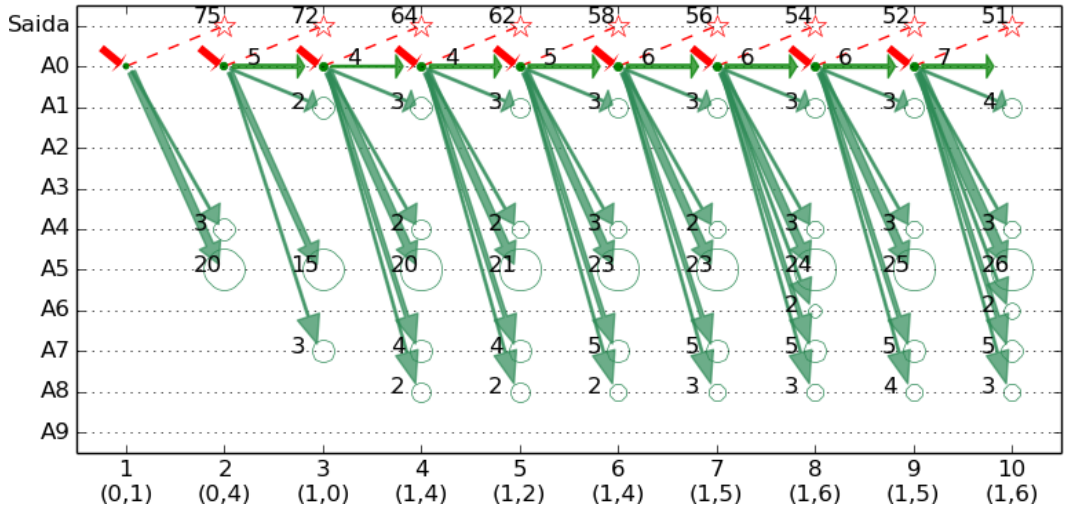


Figura A.1: O fluxo de transições centrado no tópico `A0`.

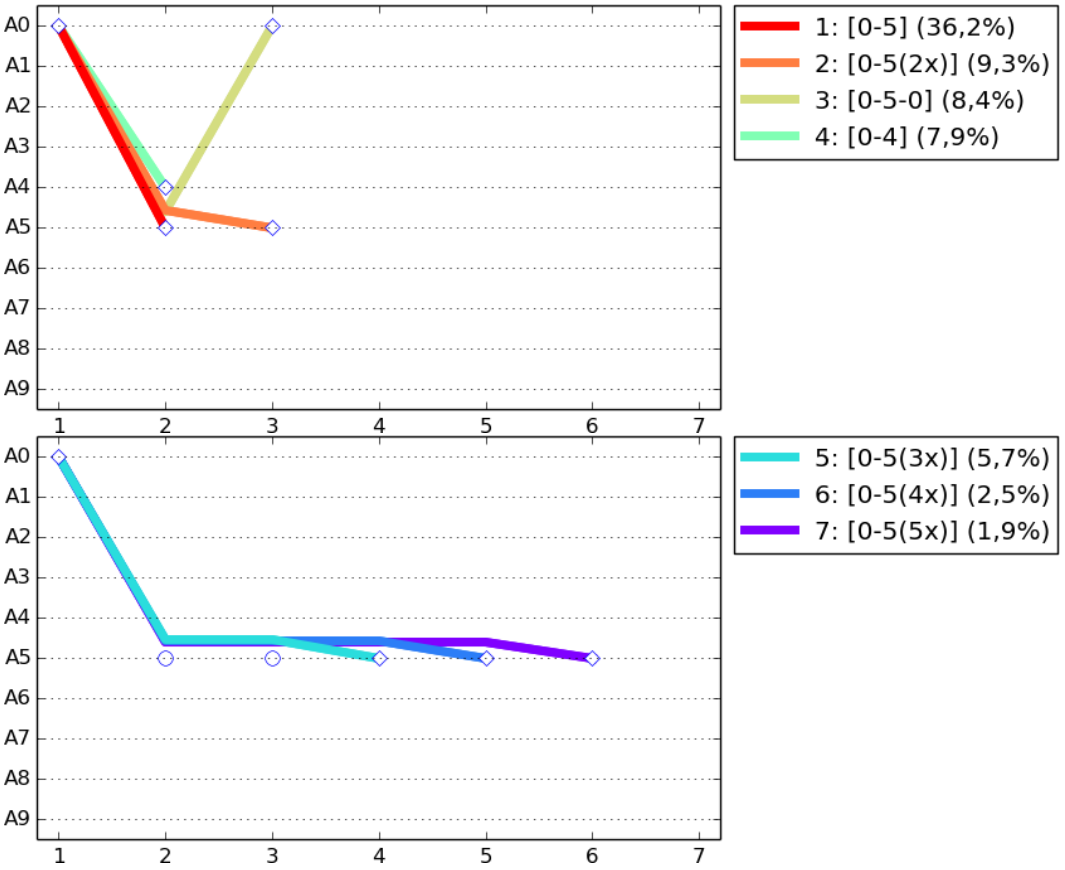


Figura A.2: Os Top-60-70% padrões de trajetórias que começam pelo tópico `A0`.

Tópico A1 – Jornal Online A

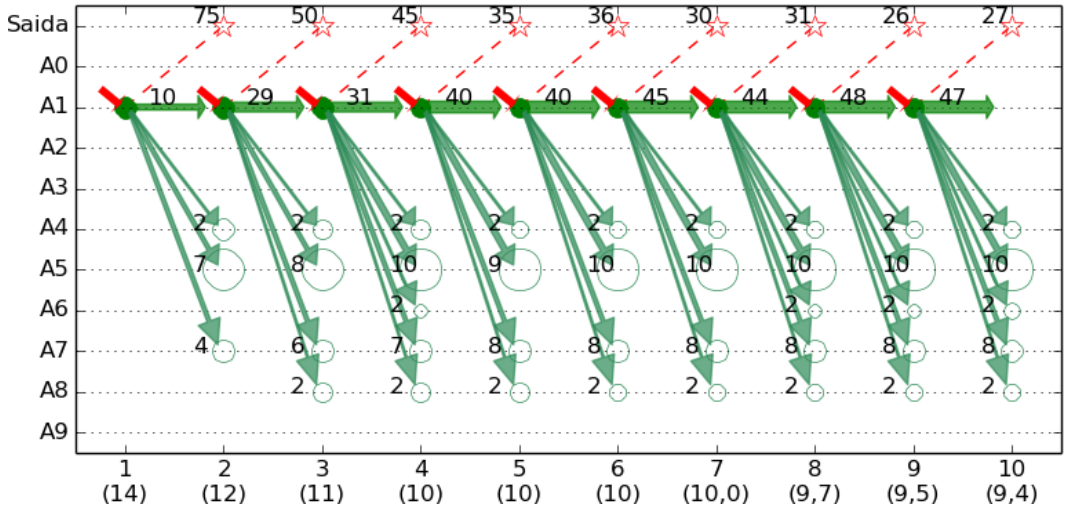


Figura A.3: O fluxo de transições centrado no tópico `A1`.

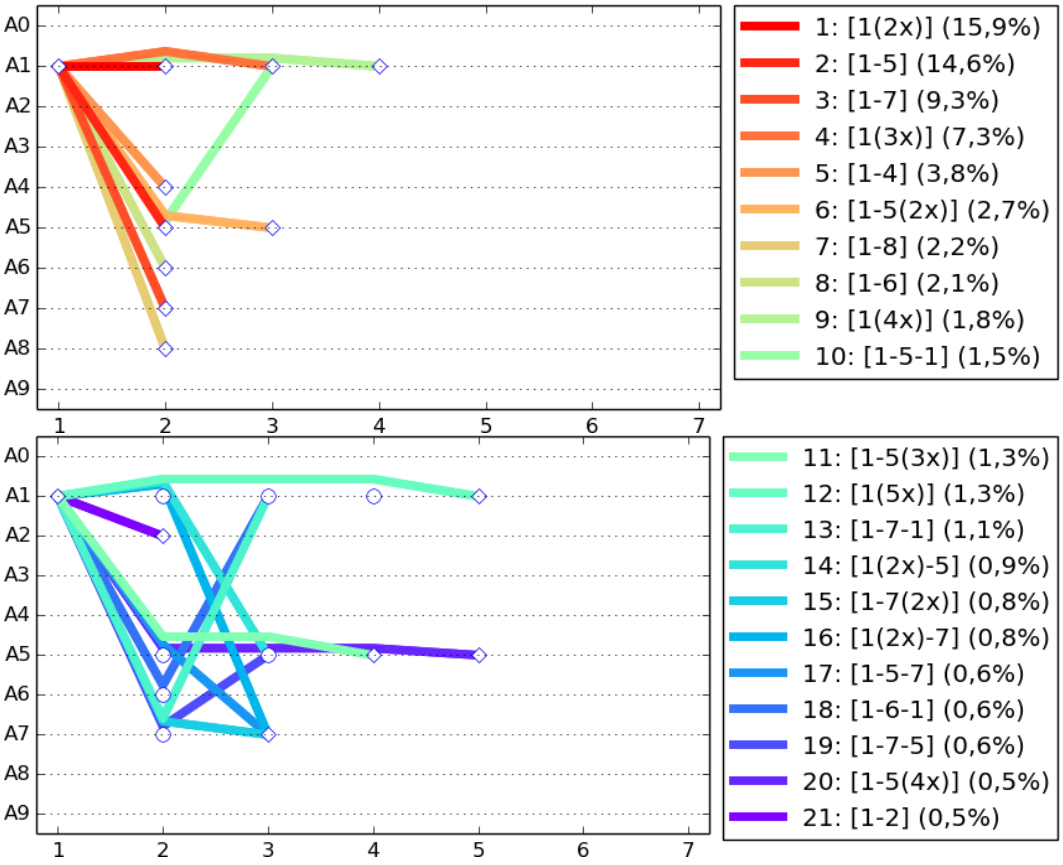


Figura A.4: Os Top-60-70% padrões de trajetórias que começam pelo tópico `A1`.

Tópico A2 – Jornal Online A

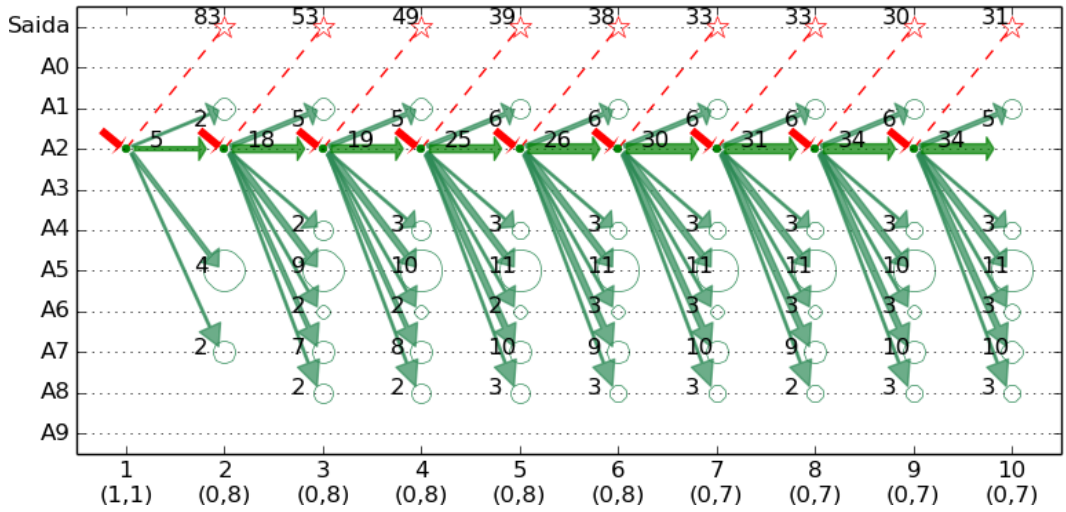


Figura A.5: O fluxo de transições centrado no tópico `A2`.

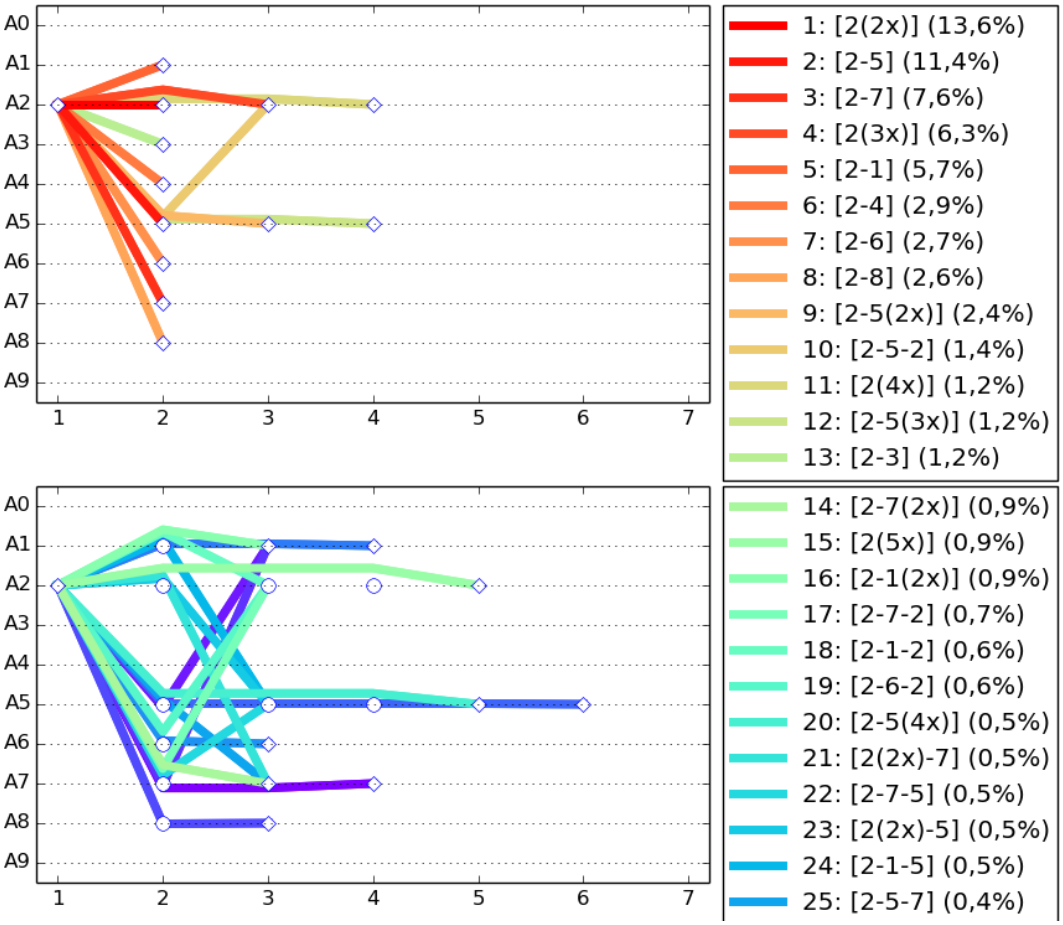


Figura A.6: Os Top-60-70% padrões de trajetórias que começam pelo tópico `A2`.

Tópico A3 – Jornal Online A

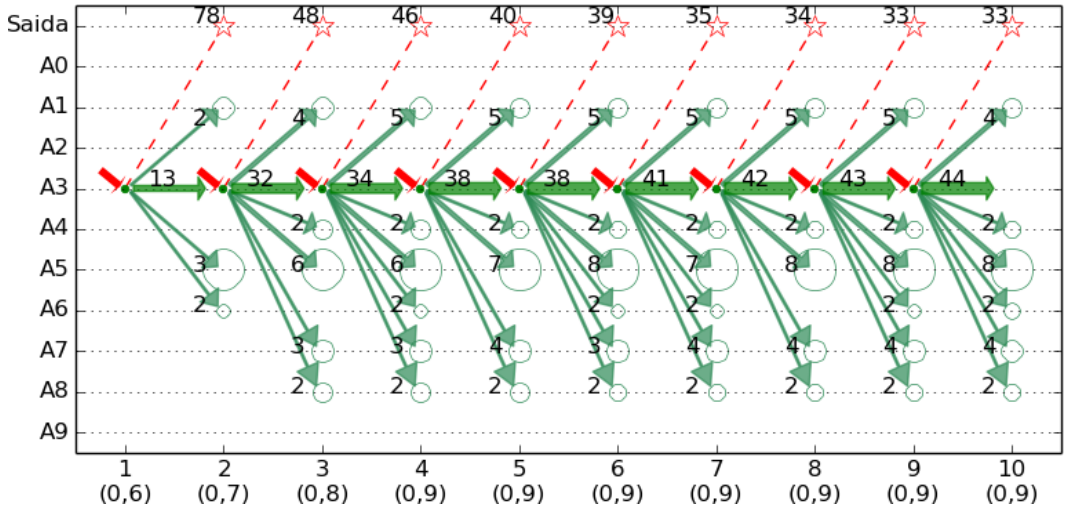


Figura A.7: O fluxo de transições centrado no tópico `A3`.

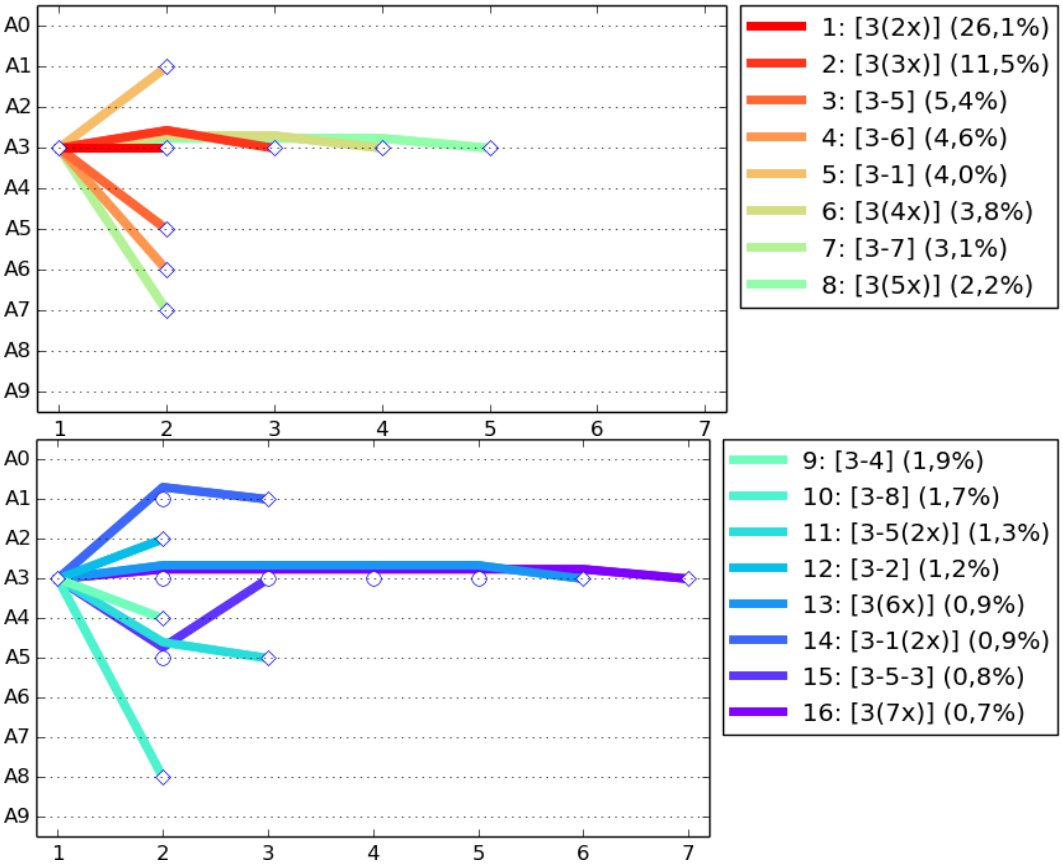


Figura A.8: Os Top-60-70% padrões de trajetórias que começam pelo tópico `A3`.

Tópico A4 – Jornal Online A

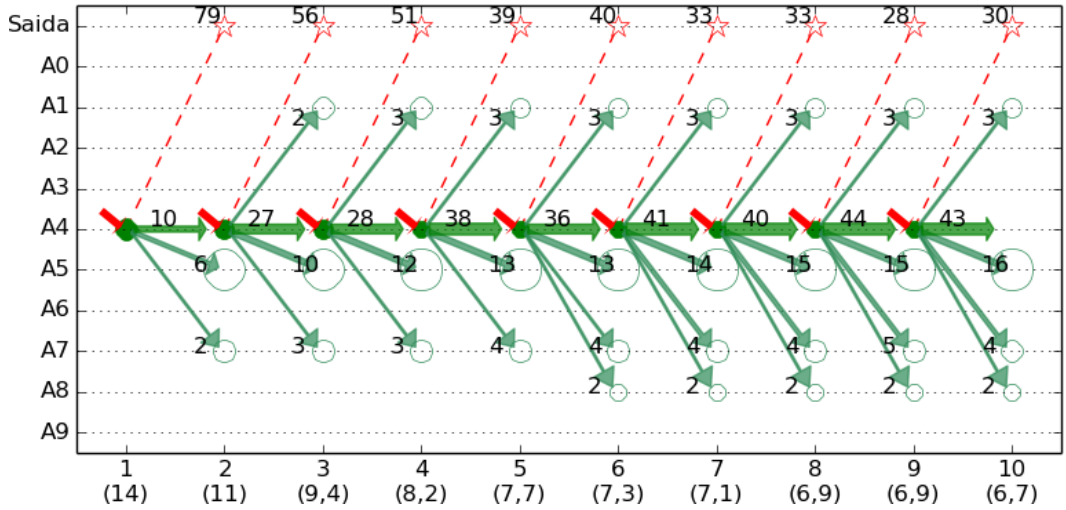


Figura A.9: O fluxo de transições centrado no tópico `A4`.

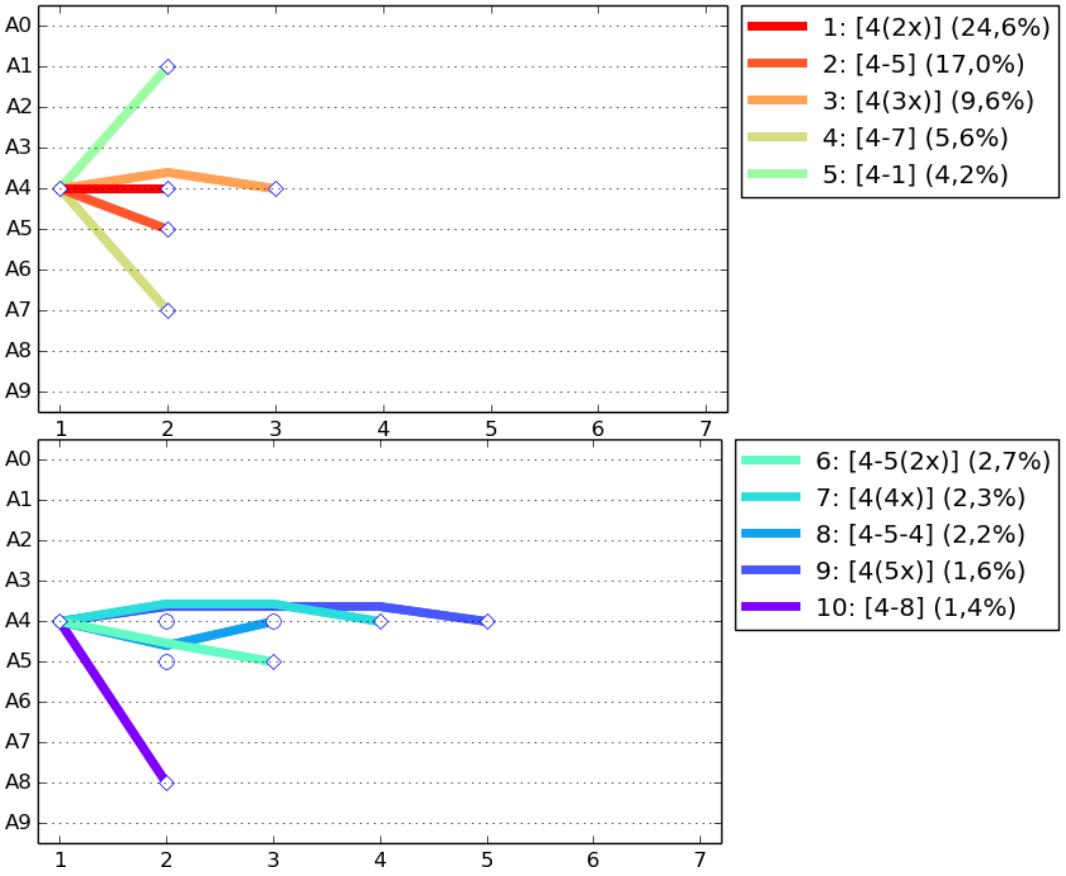


Figura A.10: Os Top-60-70% padrões de trajetórias que começam pelo tópico `A4`.

Tópico A5 – Jornal Online A

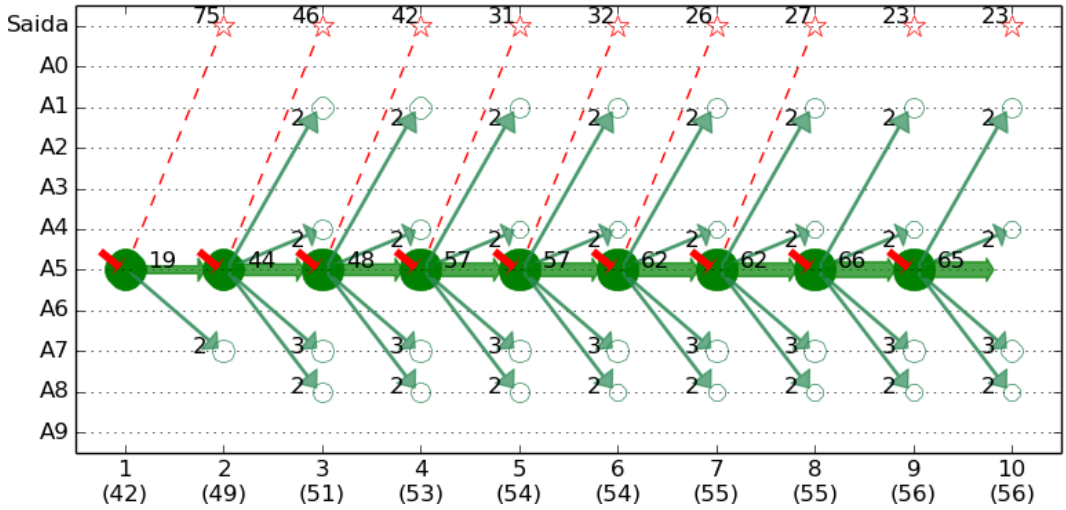


Figura A.11: O fluxo de transições centrado no tópico `A5`.

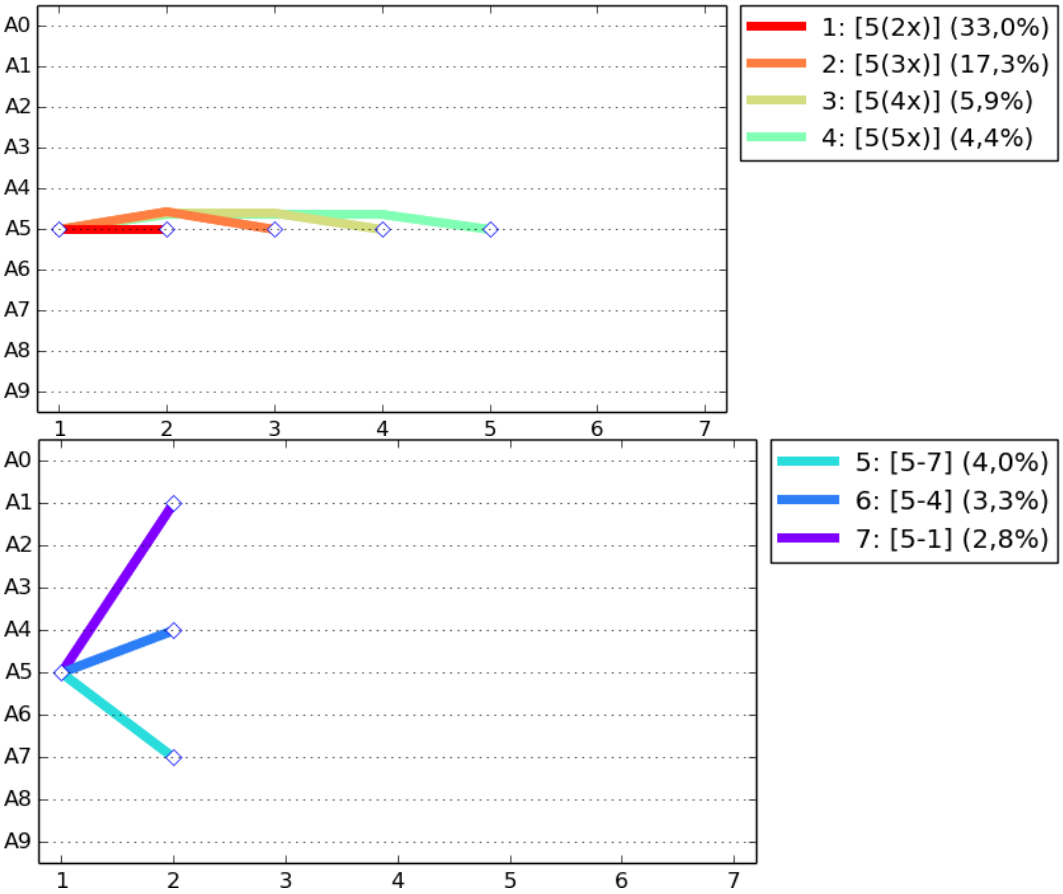


Figura A.12: Os Top-60-70% padrões de trajetórias que começam pelo tópico `A5`.

Tópico A6 – Jornal Online A

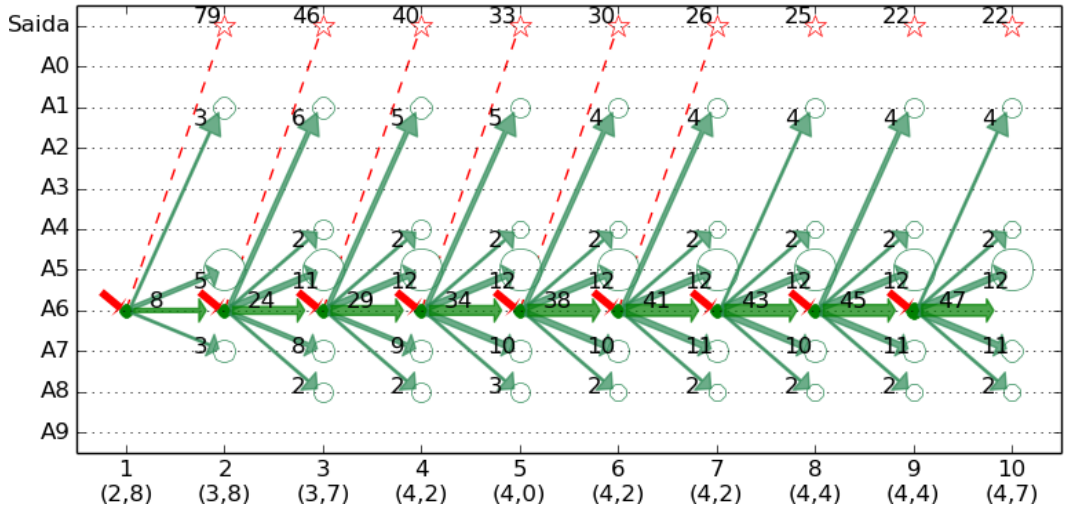


Figura A.13: O fluxo de transições centrado no tópico `A6`.

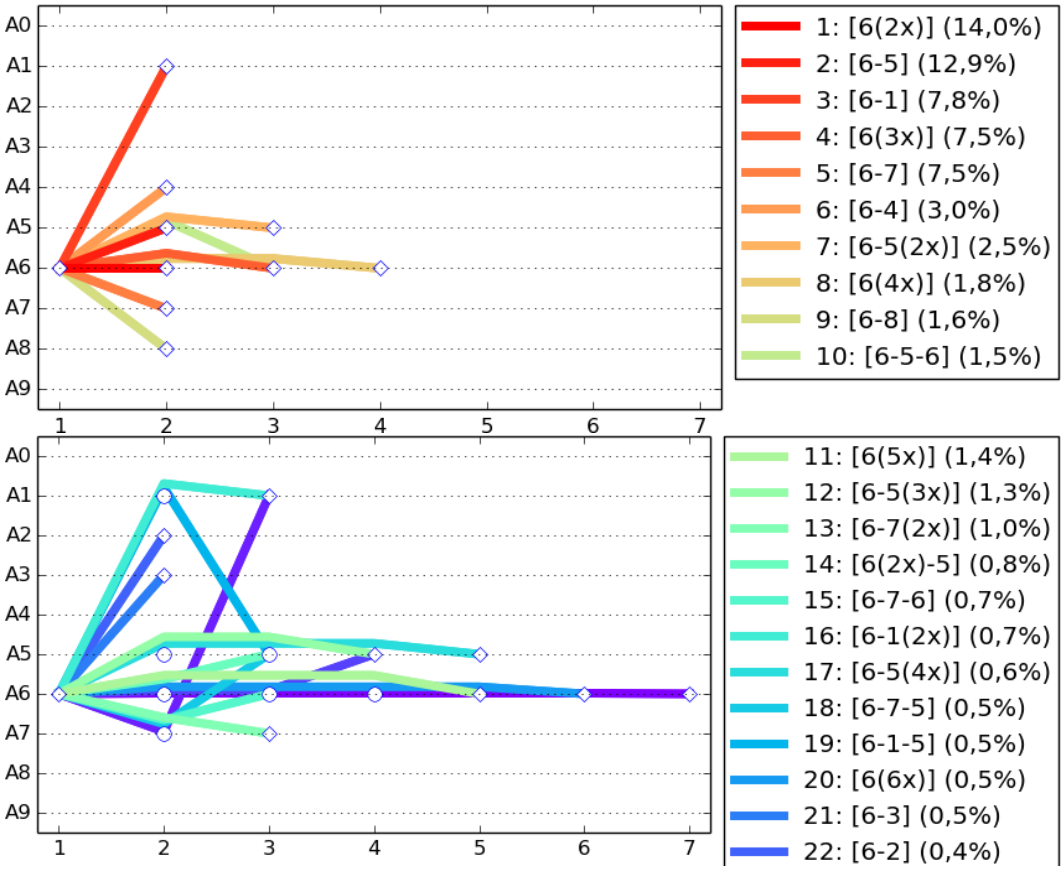


Figura A.14: Os Top-60-70% padrões de trajetórias que começam pelo tópico `A6`.

Tópico A7 – Jornal Online A

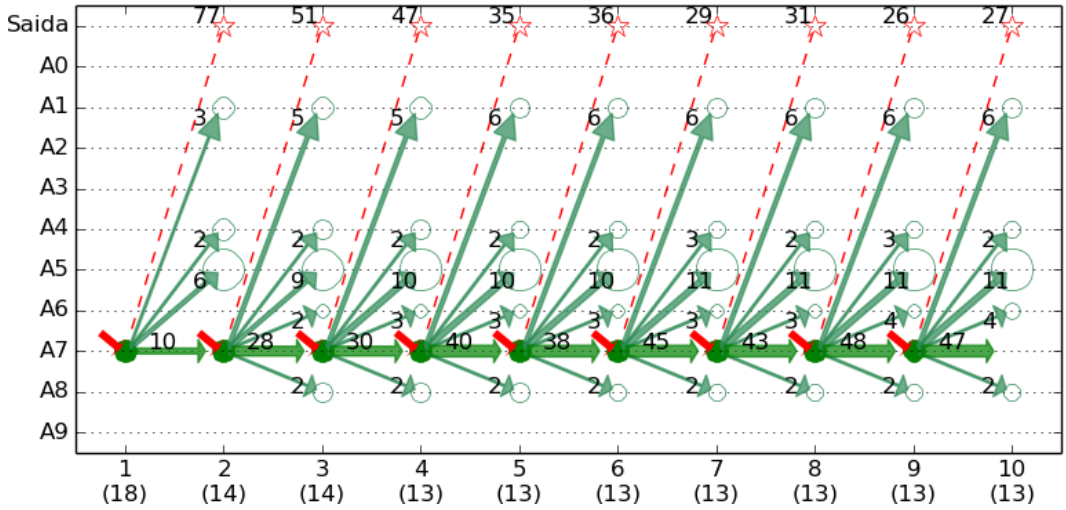


Figura A.15: O fluxo de transições centrado no tópico A7.

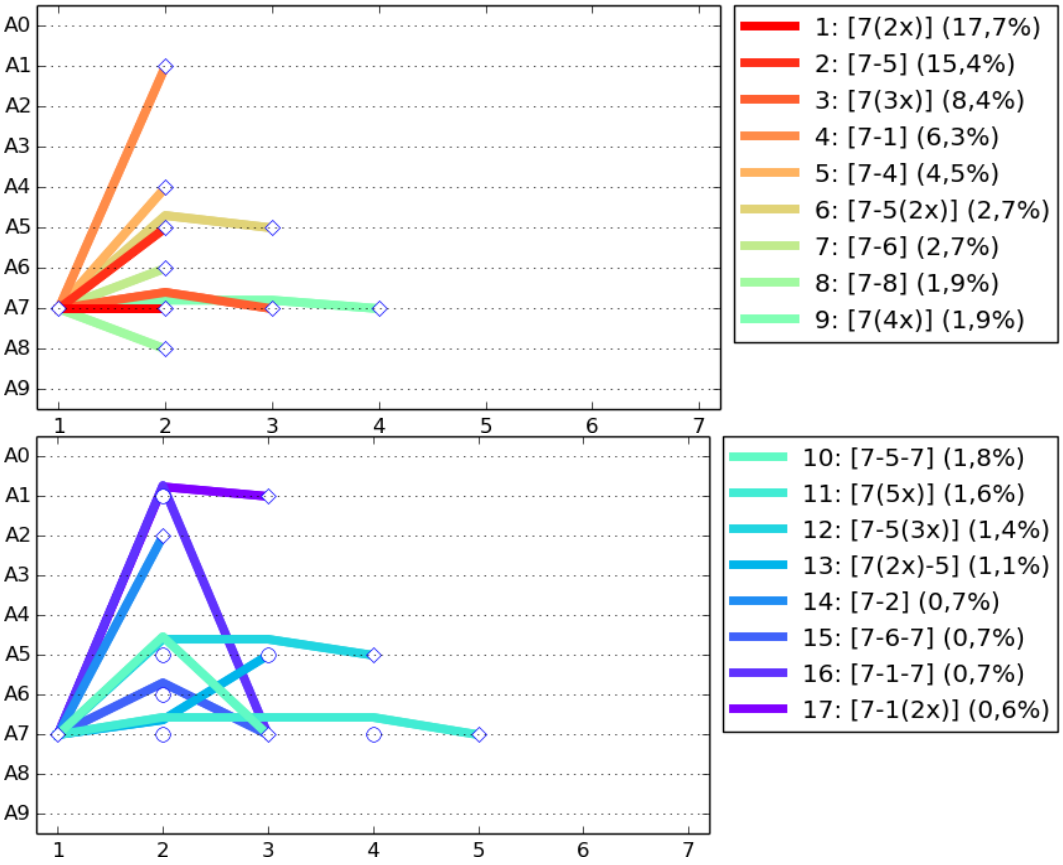


Figura A.16: Os Top-60-70% padrões de trajetórias que começam pelo tópico A7.

Tópico A8 Jornal Online A

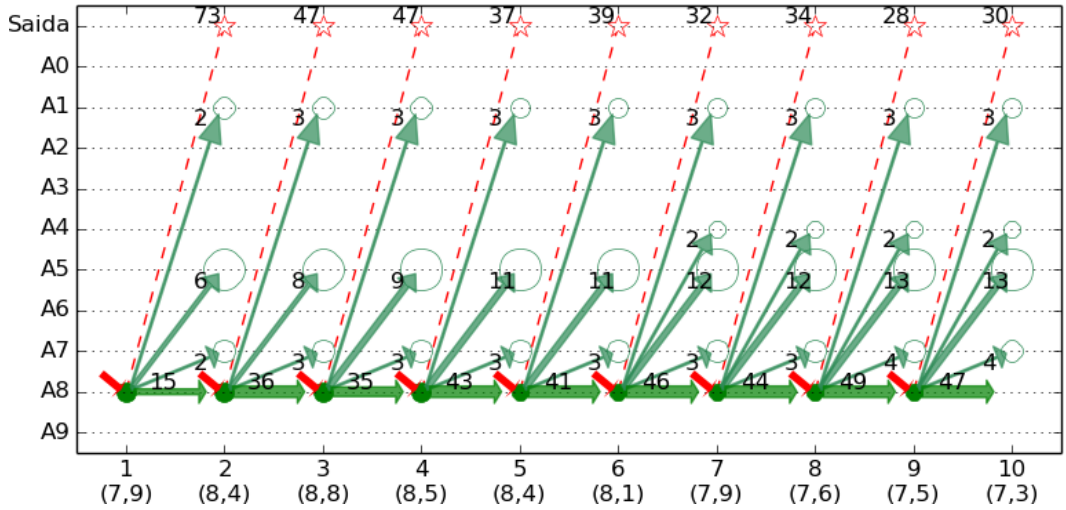


Figura A.17: O fluxo de transições centrado no tópico A8.

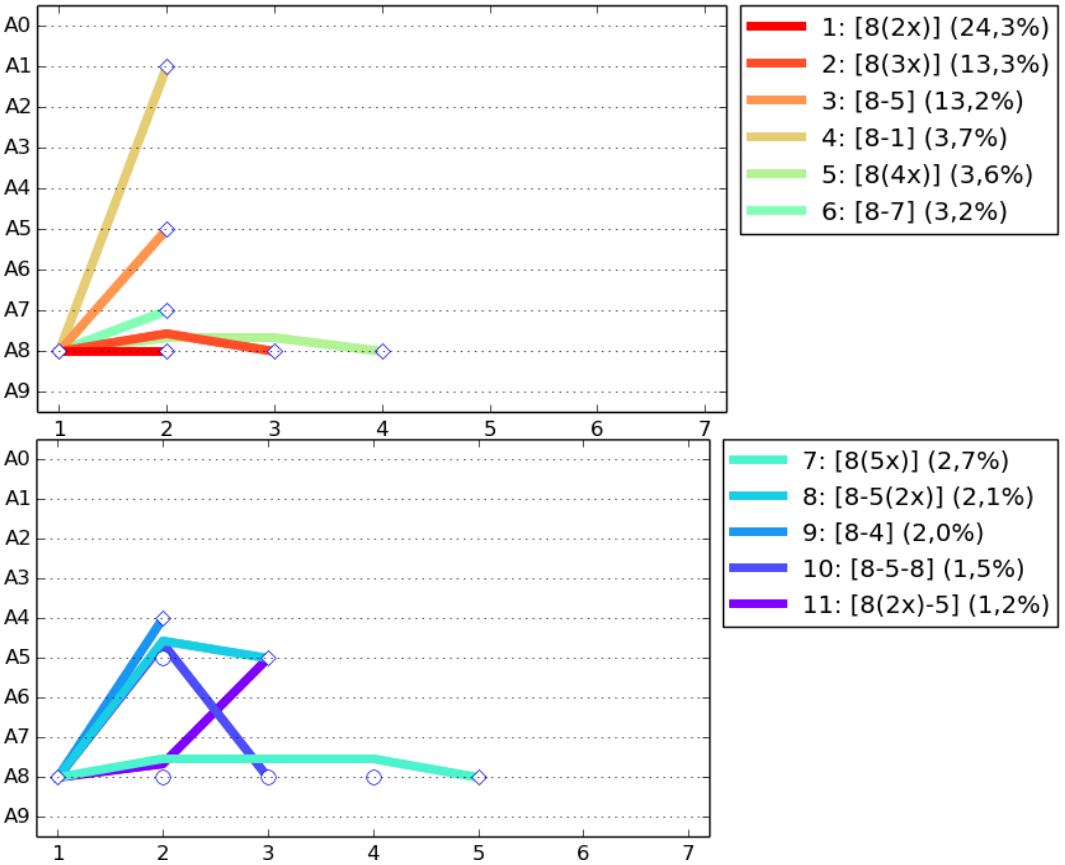
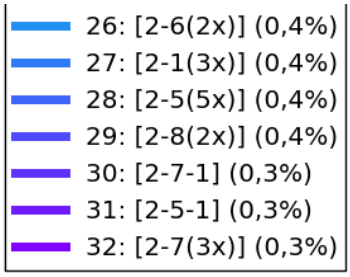
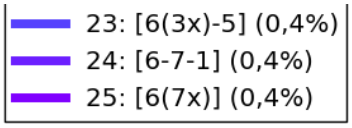


Figura A.18: Os Top-60-70% padrões de trajetórias que começam pelo tópico A8.

As Figuras A.6 e A.14 tiveram a legenda cortada para obtermos uma melhor apresentação. Os últimos padrões dessas legendas podem ser vistos abaixo:



(a) Corte da legenda A.6.



(b) Corte da legenda A.14.

Figura A.19: Os padrões cortados das legendas das Figuras A.6 e A.14.

Os resultados do tópico `A9` foram omitidos por ser este um tópico com poucos acessos. A seguir os resultados dos tópicos do **Jornal Online B**.

Tópico B0 – Jornal Online B

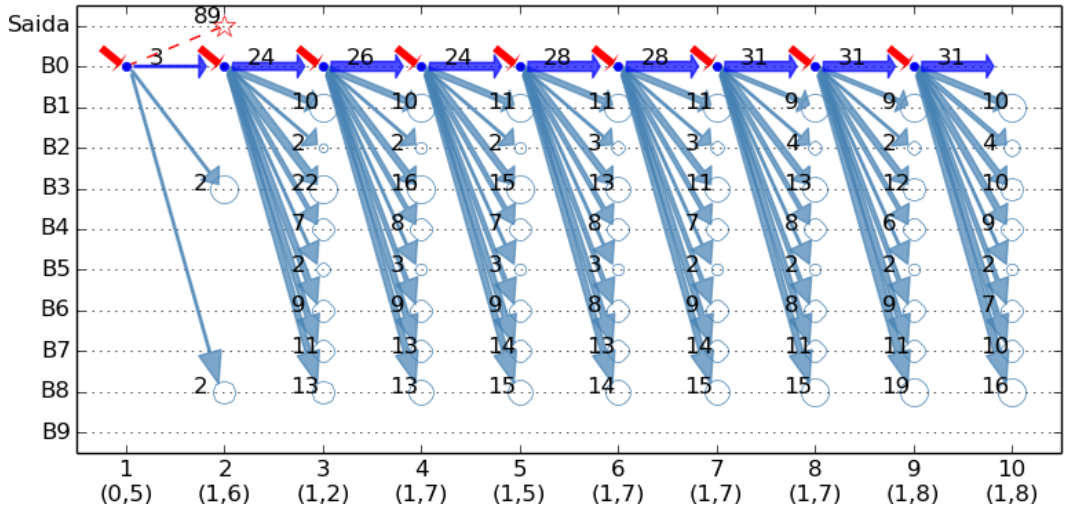


Figura A.20: O fluxo de transições centrado no tópico B0.

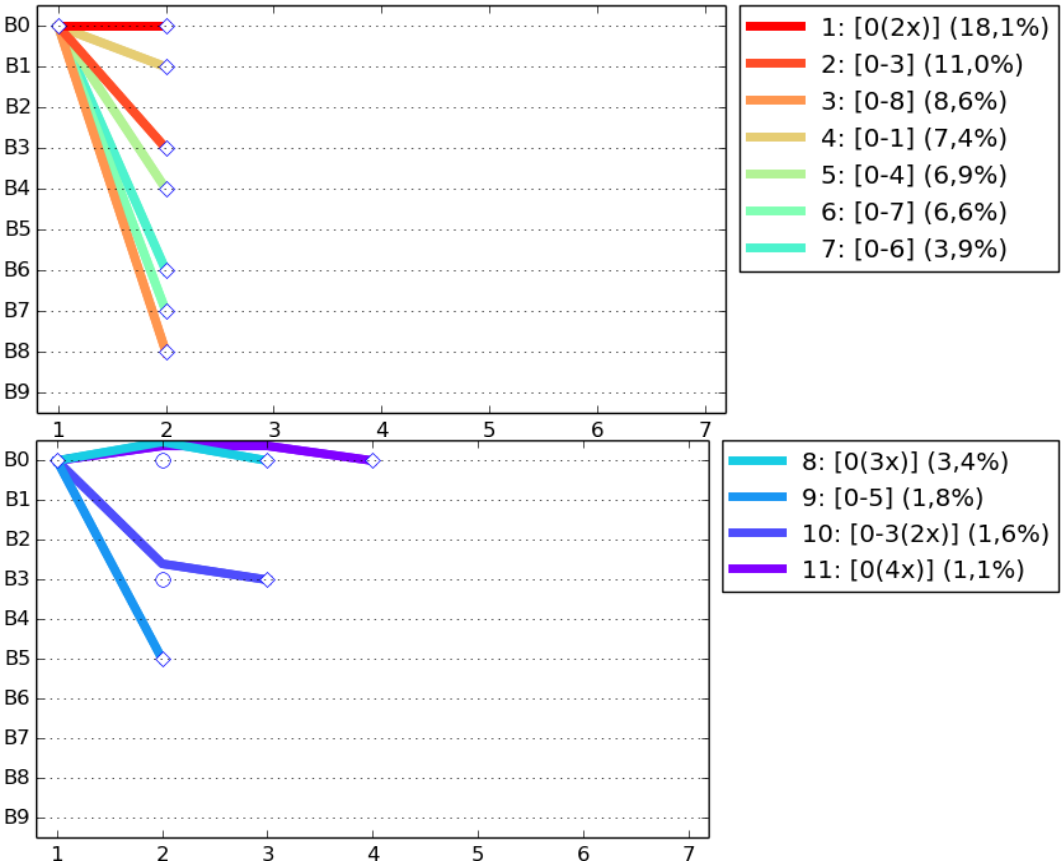


Figura A.21: Os Top-60-70% padrões de trajetórias que começam pelo tópico B0.

Tópico B1 – Jornal Online B

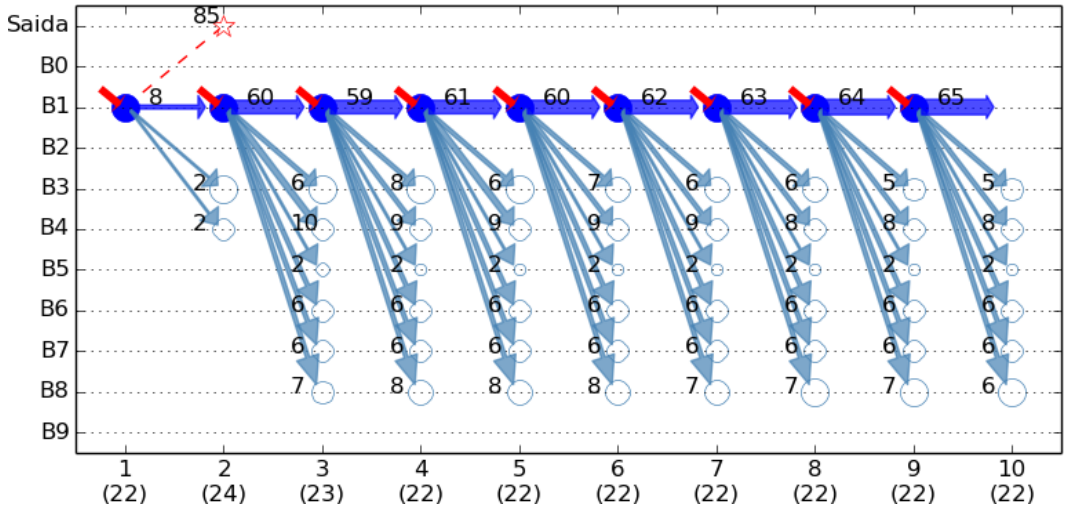


Figura A.22: O fluxo de transições centrado no tópico `B1`.

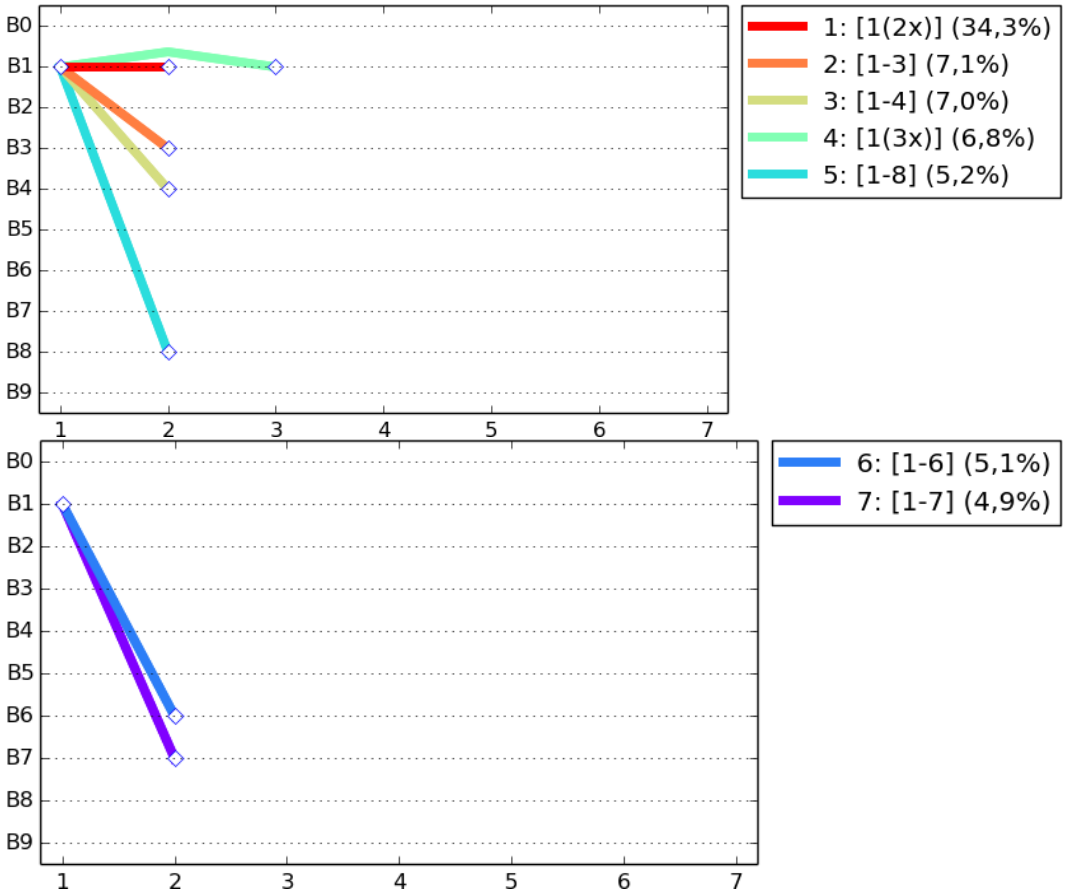


Figura A.23: Os Top-60-70% padrões de trajetórias que começam pelo tópico `B1`.

Tópico B2- Jornal Online B

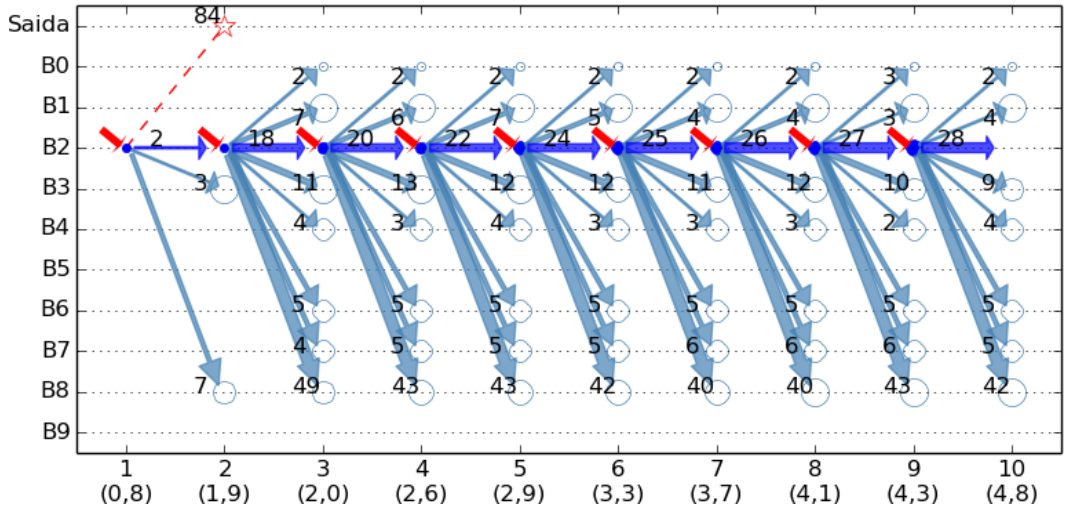


Figura A.24: O fluxo de transições centrado no tópico B2.

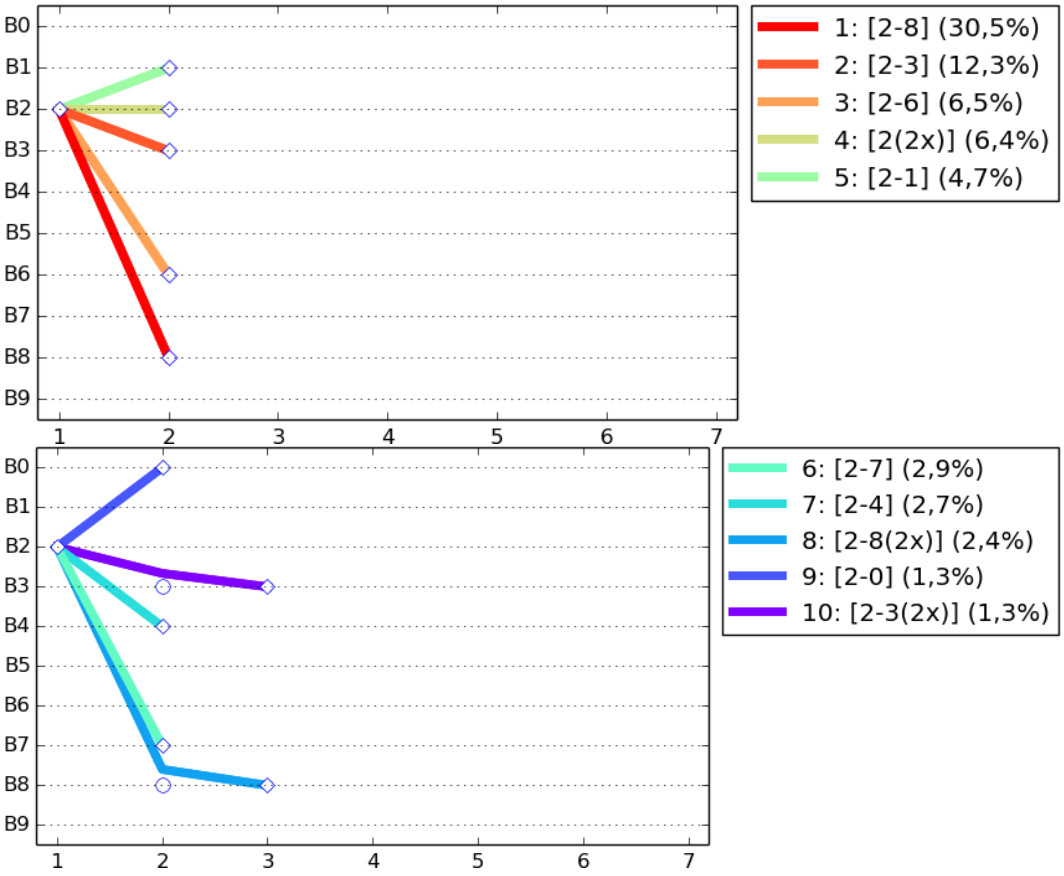


Figura A.25: Os Top-60-70% padrões de trajetórias que começam pelo tópico B2.

Tópico B3 – Jornal Online B

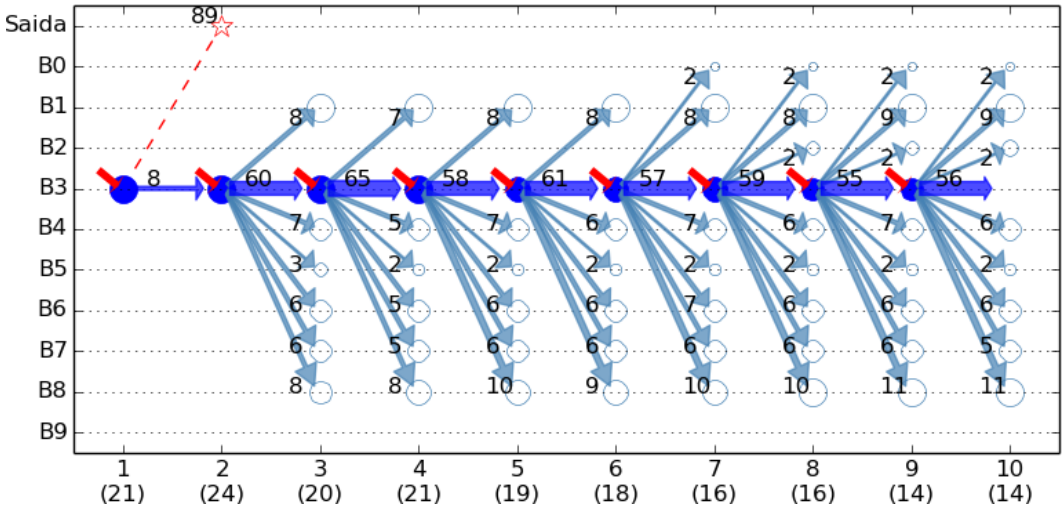


Figura A.26: O fluxo de transições centrado no tópico B3.

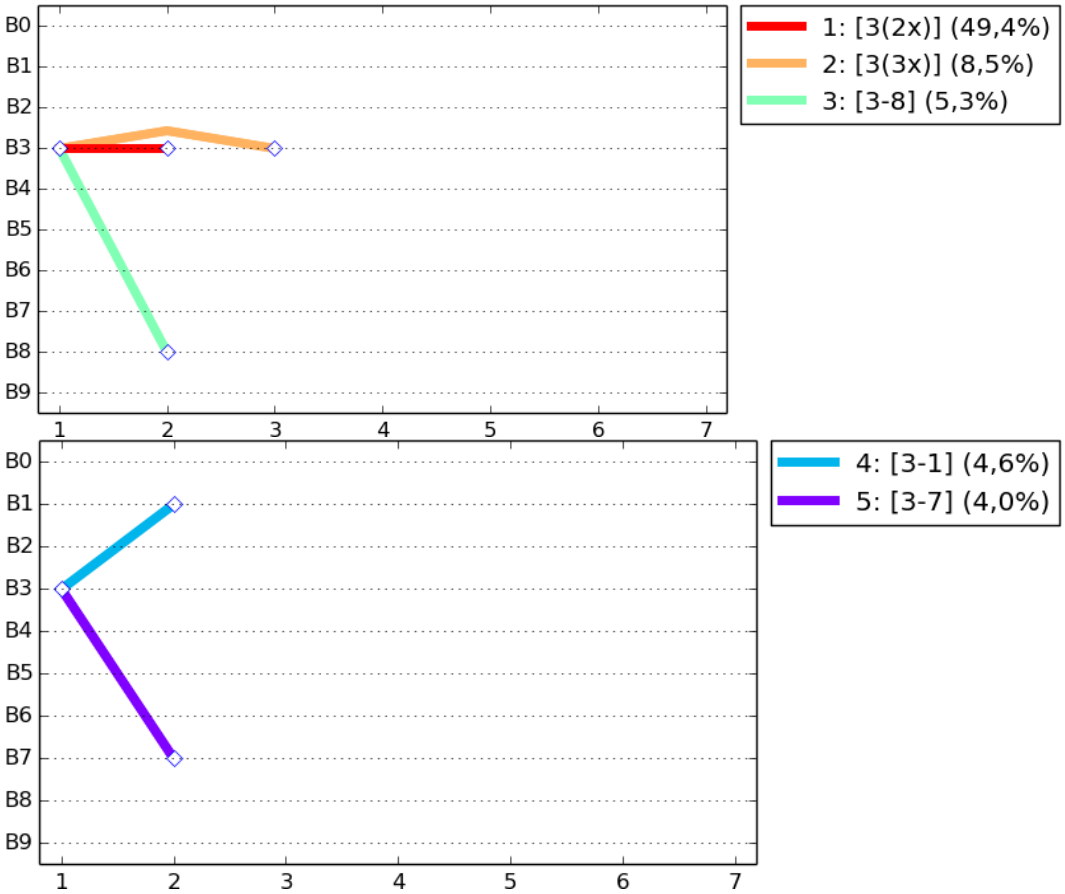


Figura A.27: Os Top-60-70% padrões de trajetórias que começam pelo tópico B3.

Tópico B4 – Jornal Online B

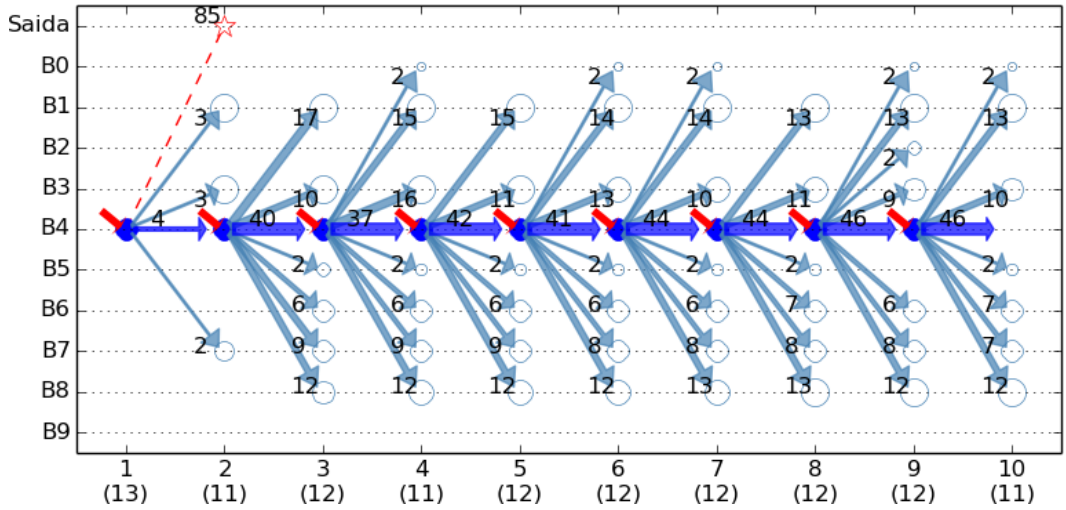


Figura A.28: O fluxo de transições centrado no tópico B4.

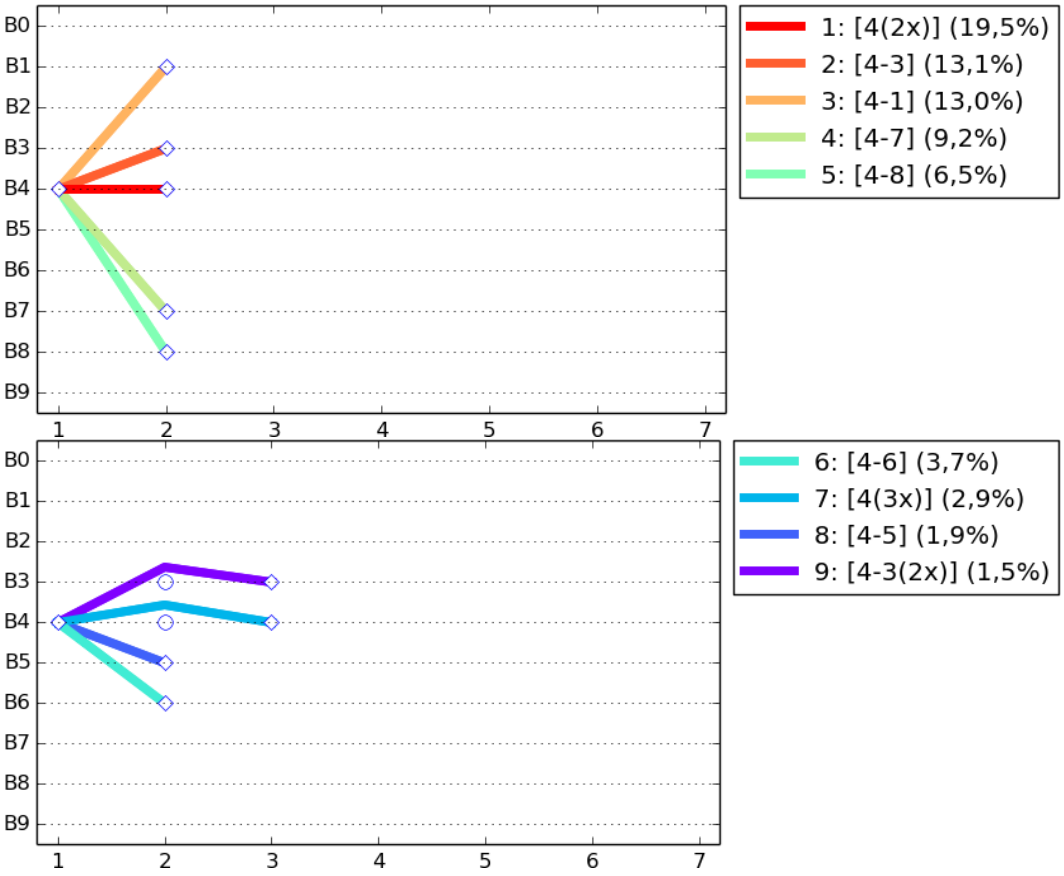


Figura A.29: Os Top-60-70% padrões de trajetórias que começam pelo tópico B4.

Tópico B5 – Jornal Online B

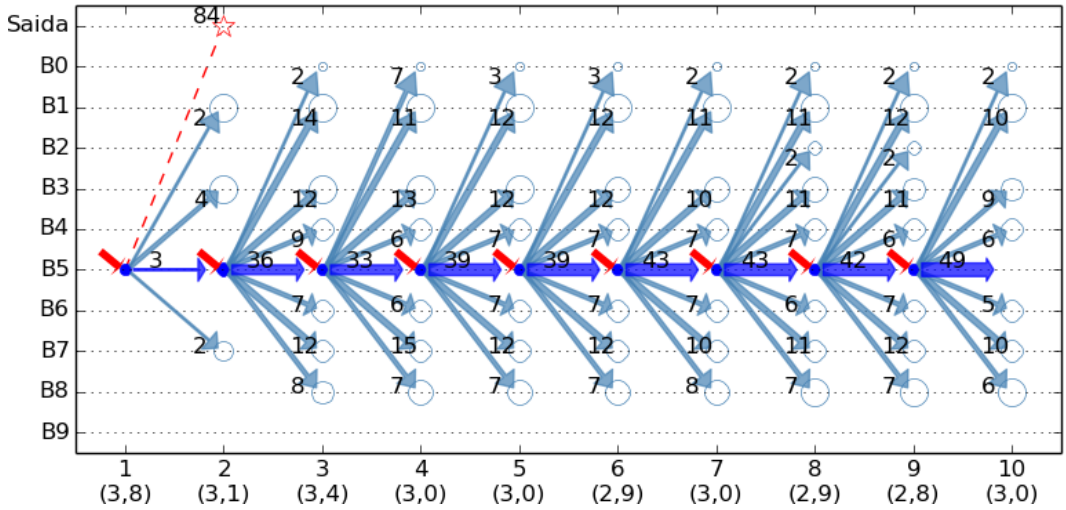


Figura A.30: O fluxo de transições centrado no tópico B5.

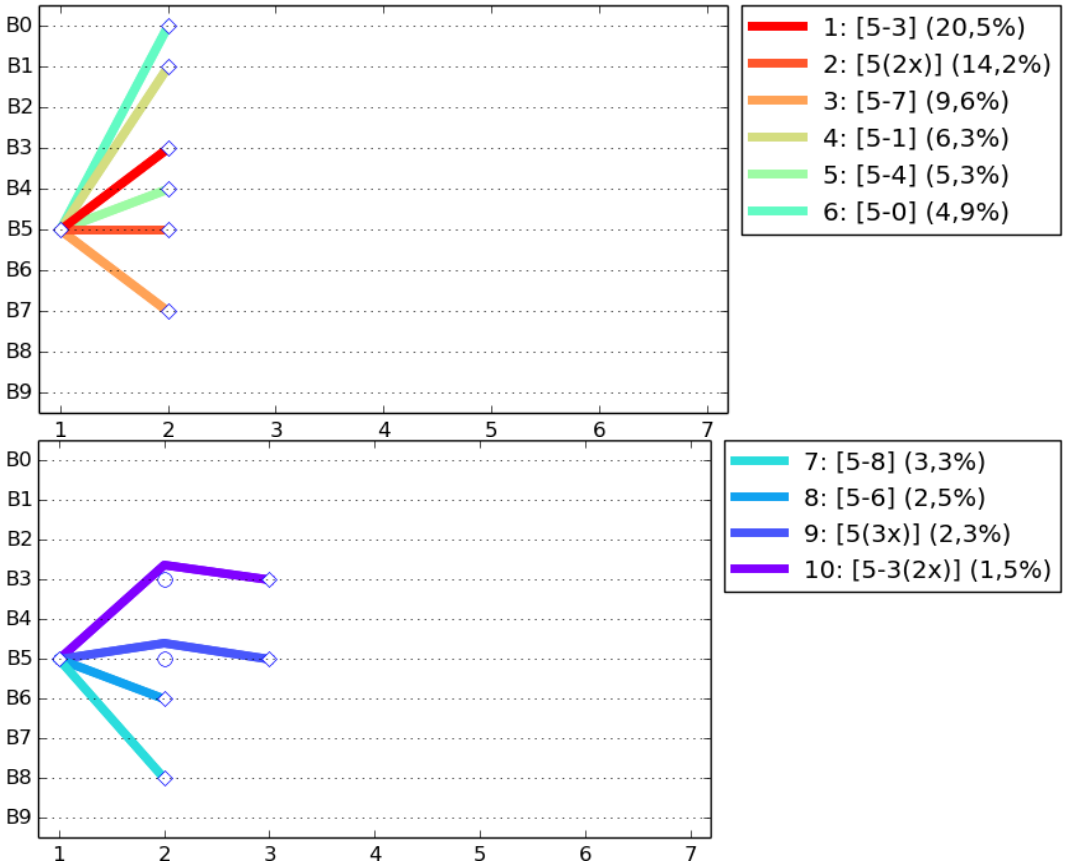


Figura A.31: Os Top-60-70% padrões de trajetórias que começam pelo tópico B5.

Tópico B6 – Jornal Online B

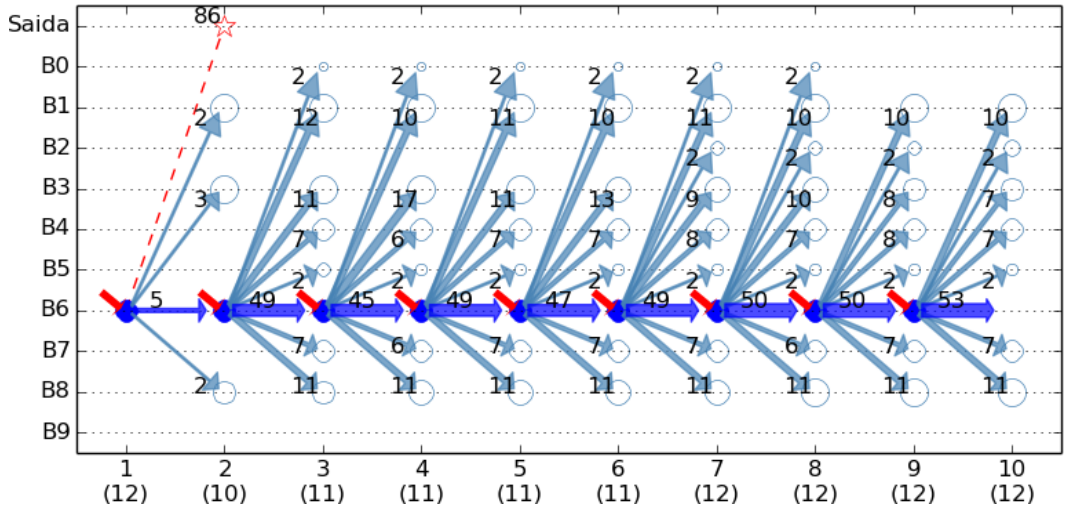


Figura A.32: O fluxo de transições centrado no tópico B6.

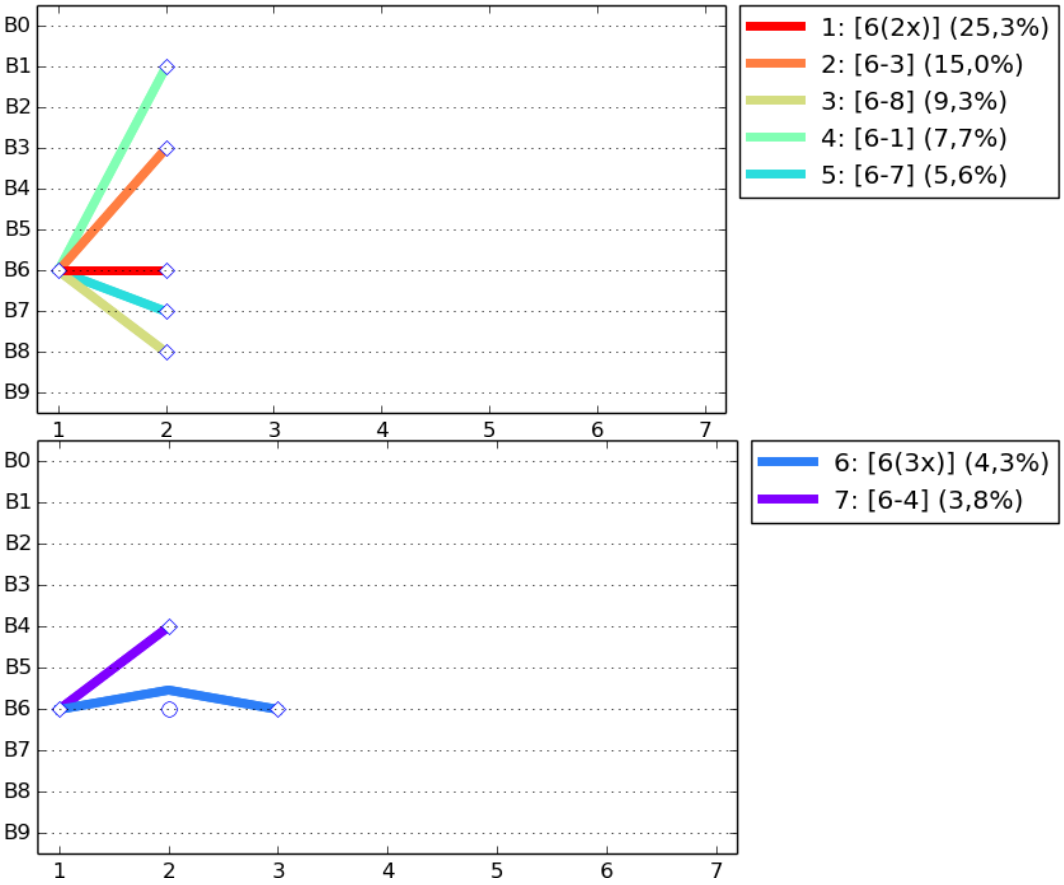


Figura A.33: Os Top-60-70% padrões de trajetórias que começam pelo tópico B6.

Tópico B7 – Jornal Online B

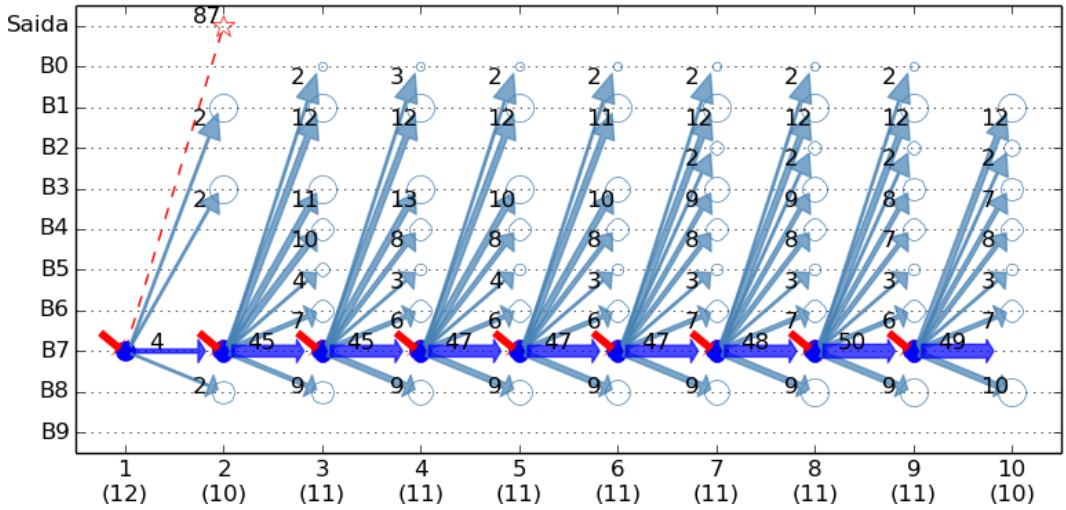


Figura A.34: O fluxo de transições centrado no tópico B7.

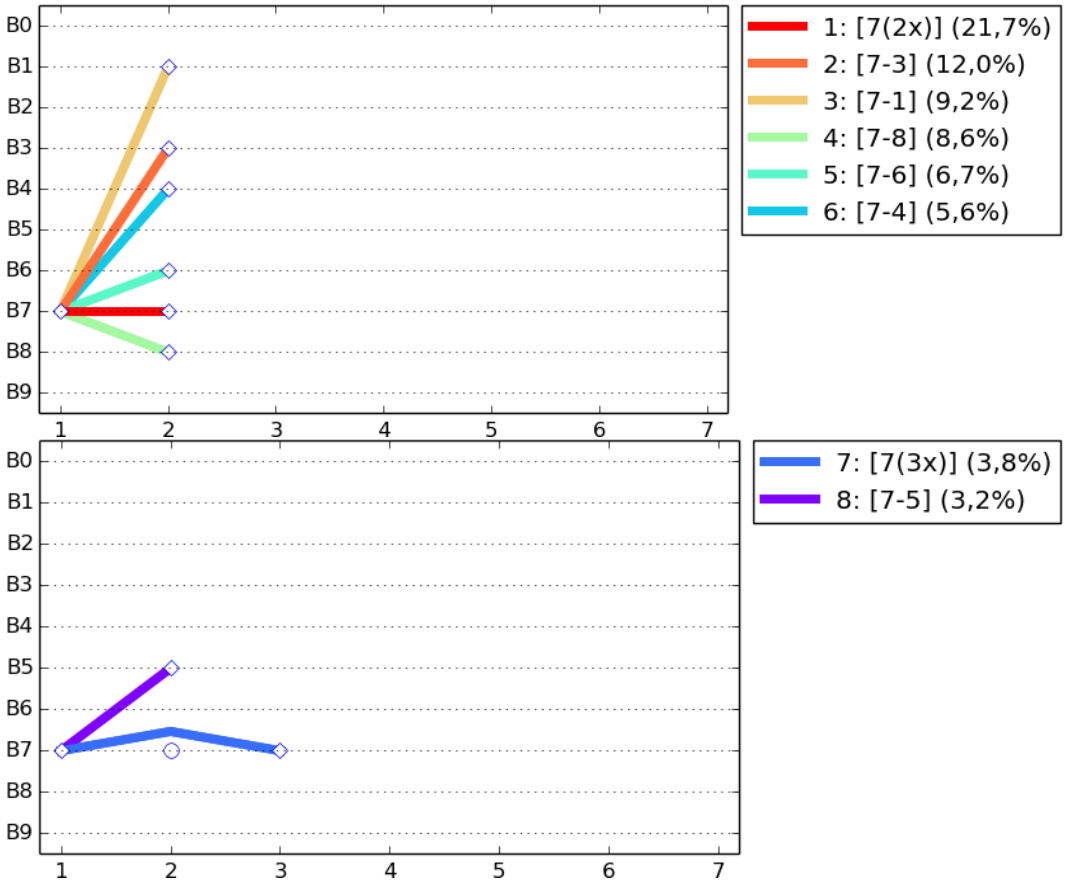


Figura A.35: Os Top-60-70% padrões de trajetórias que começam pelo tópico B7.

Tópico B8 – Jornal Online B

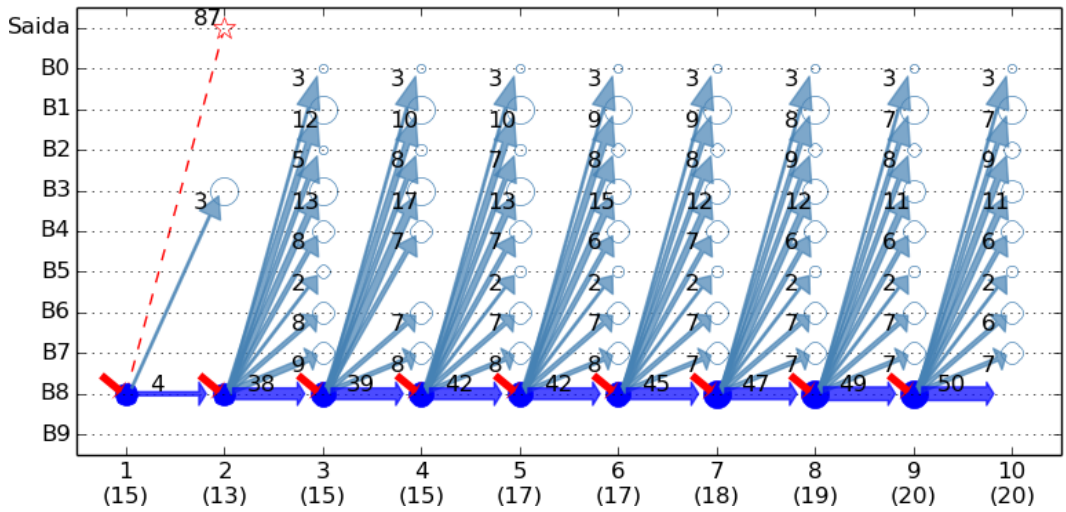


Figura A.36: O fluxo de transições centrado no tópico B8.

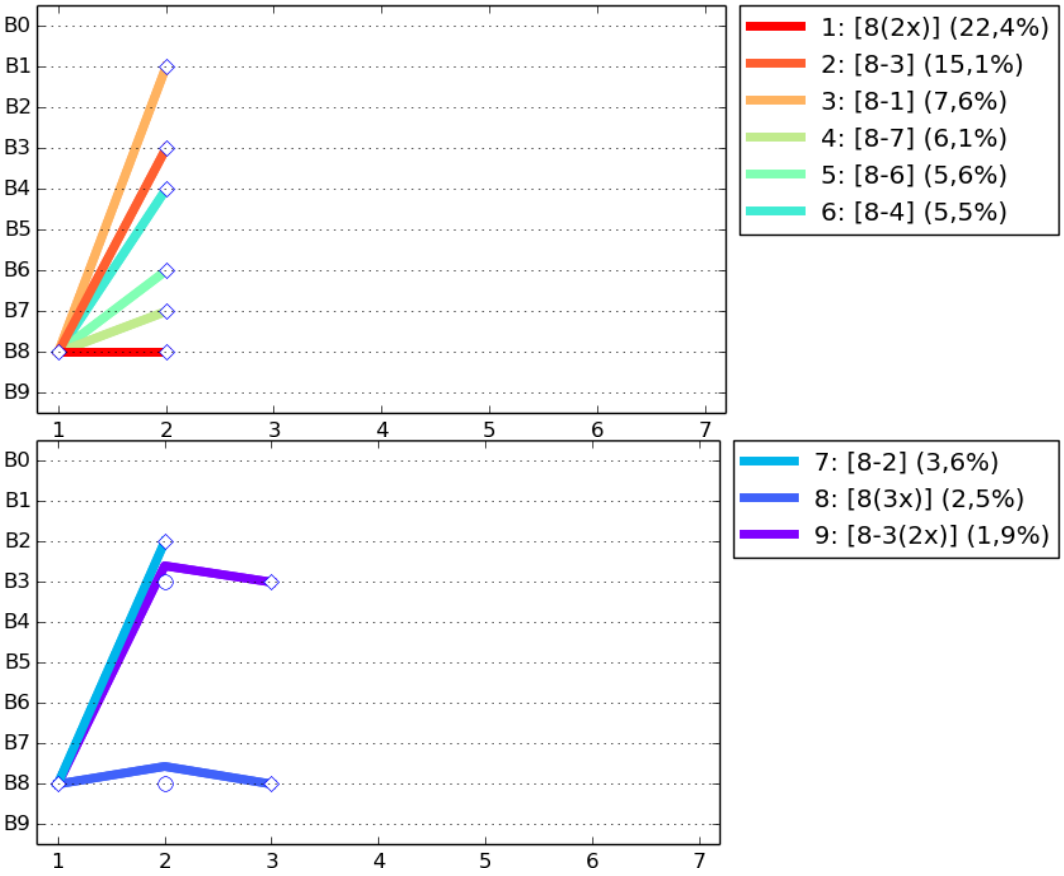


Figura A.37: Os Top-60-70% padrões de trajetórias que começam pelo tópico B8.

Tópico B9 – Jornal Online B

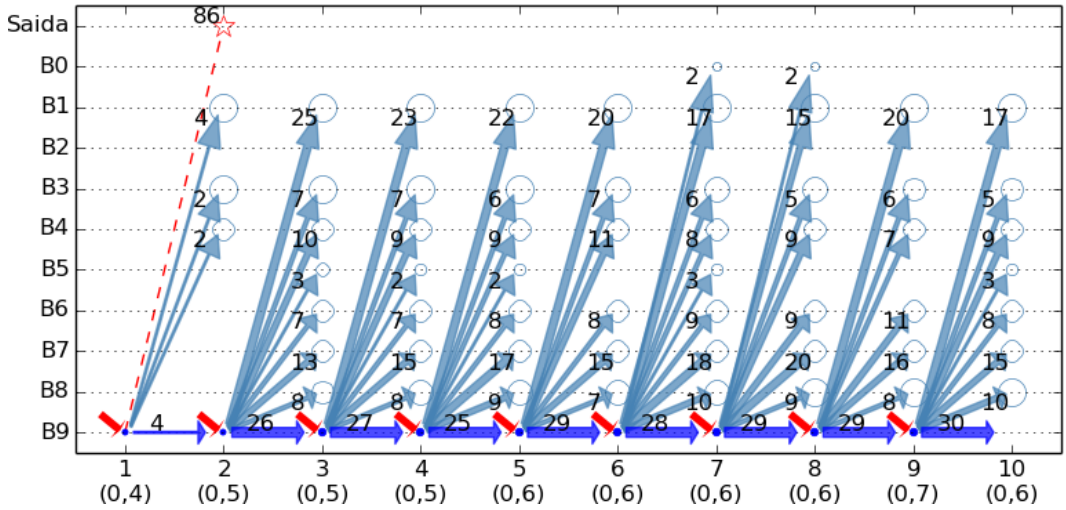


Figura A.38: O fluxo de transições centrado no tópico B9.

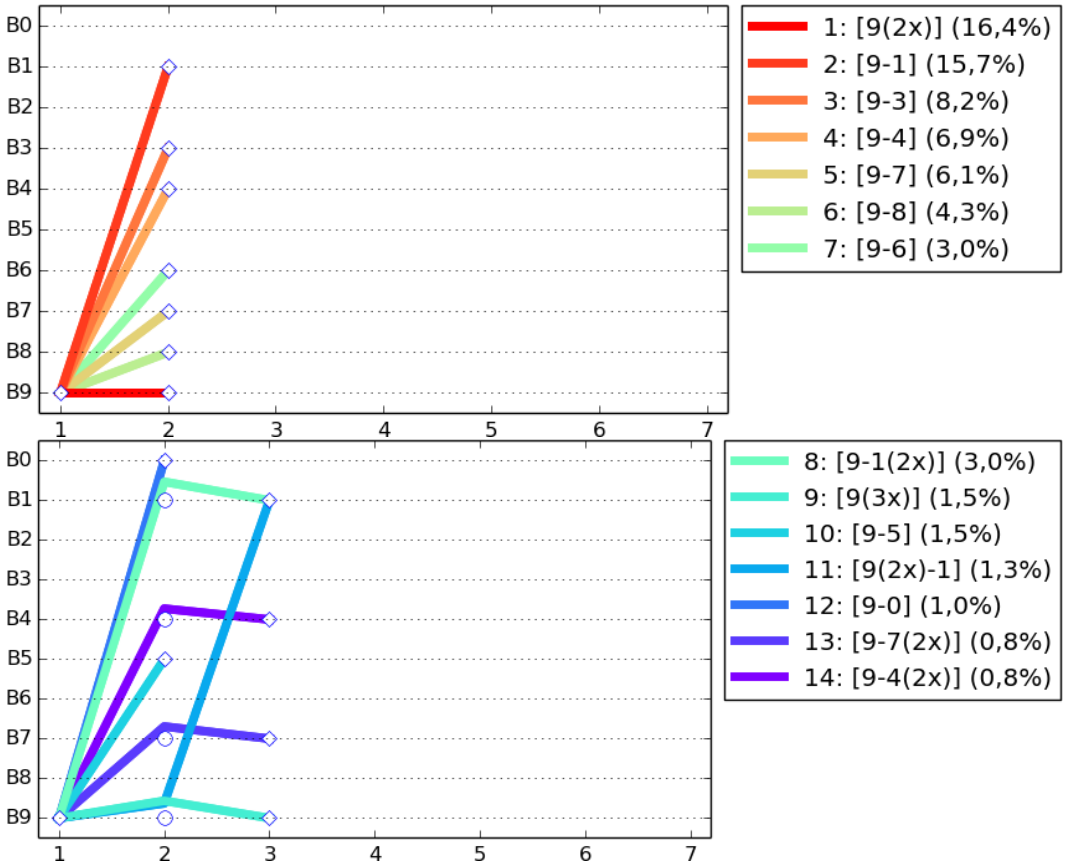


Figura A.39: Os Top-60-70% padrões de trajetórias que começam pelo tópico B9.

Apêndice B

Graus de Liberdade dos Modelos

A demonstração dos cálculos de *df* de cada modelo é apresentada a seguir. Para os cálculos foram considerados os instantes $1 \leq n \leq 10$ e o tamanho do conjunto de tópicos $L = 10$.

Os modelos **M0-U**, **M2-AP 91**, e **M2-AP 55** não estimam probabilidades pelos dados, logo possuem $df = 0$.

O modelo de independência **M1-I** calcula 10 probabilidades (9 variáveis independentes – v.i.) por instante, logo $df = 9 * 10 = 90$. Já o modelos **M1-IH**, que só calcula 10 probabilidades para todos os instantes, tem $df = 9$.

O modelo Markoviano de primeira ordem **M2-M1** calcula para cada instante $n \geq 2$, 100 probabilidades (10 para cada tópico, logo 90 v.i.) e para o instante inicial somente 10 probabilidades, logo $df = 90 * 9 + 9 = 819$. Já o modelo **M2-M1H** que só calcula 100 probabilidades para todos os instantes $n \geq 2$, e além das 10 iniciais, totaliza $df = 90 + 9 = 99$. Semelhantemente, o modelo Markoviano de segunda ordem **M2-M2** calcula para cada instante $n \geq 3$, 1000 probabilidades (100 para cada tópico, logo 900 v.i.) e 100 probabilidades (90 v.i.) no instante $n = 2$ e 10 probabilidades (9 v.i.) no instante $n = 1$, logo $df = 900 * 8 + 90 + 9 = 7299$. Já o modelo **M2-M2H** que só calcula 1000 probabilidades para todos os instantes $n \geq 3$, e além das 110 nos dois instantes iniciais, totaliza $df = 900 + 90 + 9 = 999$. Analogamente ao último, os modelos **M2-M3H** e **M2-M4H** tem respectivamente $df = 9.000 + 900 + 90 + 9 = 9.999$ e $df = 90.000 + 9.000 + 900 + 90 + 9 = 99.999$.

Os modelos **M3-S**, **M4-D**, **M5-L** e **M6-E** para os instantes $n \geq 2$ estimam cada um n probabilidades por tópico, porém são todos independentes entre si essas probabilidades. Assim cada instante temos $10 \times n$ v.i., totalizando $df = 100 + 90 + 80 + 70 + 60 + 50 + 40 + 30 + 20 + 9 = 549$. Observação, no primeiro instante são calculados 9 v.i..

Para os modelos **M3-S m2** e **M4-D m2** há menos variáveis sendo estimadas por instantes, no máximo 30 para qualquer instante $n \geq 2$, totalizando $df = 30+30+30+30+30+30+30+30+30+20+9 = 269$. Já para os modelos **M3-S m3** e **M4-D m3**, para os instantes $n \geq 3$ são estimadas 40 probabilidades independentes, o que totaliza com as demais probabilidades dos instantes anteriores: $df = 40+40+40+40+40+40+40+30+20+9 = 339$.

Para **M7-PG A-I**, **M7-PG B-I** e **M7-PG C-I** são estimados os valores das variáveis θ e π para cada tópico instante a instante. Logo os graus de liberdade desses dois modelos é igual a $df = 10*9+10*9 = 180$. Os modelos **M7-PG A-II**, **M7-PG B-II** e **M7-PG C-II** estimam somente os valores de θ variante nos instantes, logo $df = 10*9+9 = 99$. E os modelos **M7-PG A-III**, **M7-PG B-III** e **M7-PG C-III** só estimam as probabilidades θ e π uma vez por tópico gerando $df = 9+9 = 18$.

Para **M8-VC A** e **M8-VC B** são estimados os valores das probabilidades $\pi(l)$ para cada tópico l instante a instante. Logo os graus de liberdade desses dois modelos são igual a $df = 10*9 = 90$. E os modelos **M8-VC A-H** e **M8-VC B-H** só estimam as probabilidades $\pi(l)$ uma vez por tópico l gerando $df = 9$.

Apêndice C

Resultados AIC das 5 partições

A seguir as tabelas completas com os resultados do AIC para cada partição de dados. A tabela C.1 mostra os resultados dos modelos na base do **Jornal Online A** e a tabela C.2 os resultados na base **Jornal Online B**. A terceira tabela C.3 mostra os valores das médias das partições e o desvio padrão (DP). Nessa última tabela os modelos estão ranqueados pela média+1DP. O resultado de ambas bases de dados são apresentadas em conjunto.

Modelos	AIC Fold-1	AIC Fold-2	AIC Fold-3	AIC Fold-4	AIC Fold-5
M0-U	-54003697,5	-54003140,2	-54019336,6	-54003508,7	-54009863,8
M1-I	-36239310,4	-36244391,9	-36252769,5	-36246221,8	-36256556,4
M1-IH	-36439873,1	-36445717,8	-36451796,1	-36446730,9	-36455758,2
M2-AP 91	-42668887,0	-42691120,7	-42691784,1	-42674820,1	-42675513,5
M2-AP 55	-39256194,8	-39267881,4	-39272086,6	-39259302,8	-39261183,1
M2-M1	-29247439,6	-29260590,3	-29259226,3	-29253962,9	-29259408,0
M2-M1H	-29413764,0	-29426331,4	-29426273,9	-29420497,1	-29426183,3
M2-M2	-28411091,7	-28420151,9	-28423106,3	-28416747,2	-28425150,7
M2-M2H	-28414945,5	-28423830,8	-28427762,4	-28421227,9	-28429516,0
M2-M3H	-28368482,0	-28375899,5	-28380302,3	-28374387,9	-28381509,4
M2-M4H	-28481072,6	-28487302,7	-28492142,7	-28487312,5	-28494682,9
M3-S	-35283583,1	-35293629,8	-35297531,5	-35295181,8	-35301630,8
M3-S m2	-35398700,7	-35405456,6	-35413404,7	-35409381,2	-35415950,3
M3-S m3	-35336901,4	-35345214,5	-35351244,2	-35347955,7	-35354355,7
M4-D	-35367654,1	-35377554,0	-35381389,3	-35378342,6	-35388174,4
M4-D m2	-35435155,7	-35441888,5	-35449499,3	-35445090,5	-35453373,5
M4-D m3	-35395720,6	-35404116,7	-35409981,0	-35406197,6	-35414636,9
M5-L	-35301137,4	-35311368,2	-35315279,0	-35312954,7	-35320385,4
M6-E	-35412265,5	-35418013,2	-35426112,7	-35422194,3	-35429597,1
M7-PG A-I	-30897025,6	-30902921,7	-30911260,7	-30900634,5	-30904479,9
M7-PG A-II	-31306316,0	-31313807,9	-31321724,6	-31313008,1	-31314818,7
M7-PG A-III	-31492322,2	-31499485,9	-31503117,1	-31497859,9	-31502052,8
M7-PG B-I	-31217466,6	-31215605,5	-31225383,4	-31216447,8	-31224493,0
M7-PG B-II	-32655619,5	-32658854,0	-32668509,6	-32658861,3	-32664473,3
M7-PG B-III	-32841625,7	-32844532,0	-32849902,1	-32843713,1	-32851707,4
M7-PG C-I	-30520660,4	-30522366,7	-30532190,6	-30523101,0	-30528407,9
M7-PG C-II	-30946916,0	-30950716,0	-30960345,2	-30952315,7	-30955913,9
M7-PG C-III	-31132922,2	-31136394,0	-31141737,6	-31137167,6	-31143147,9
M8-VC A	-29491466,9	-29493055,4	-29506885,5	-29495241,3	-29504031,6
M8-VC A-H	-29639548,0	-29639679,0	-29653758,0	-29643021,2	-29651213,5
M8-VC B	-29249723,3	-29252539,6	-29262932,1	-29251512,4	-29262080,3
M8-VC B-H	-29441494,1	-29443838,7	-29453162,7	-29443097,2	-29453037,2

Tabela C.1: AIC das 5 partições na base do **Jornal Online A**.

Modelos	AIC Fold-1	AIC Fold-2	AIC Fold-3	AIC Fold-4	AIC Fold-5
M0-U	-13530247,5	-13527226,5	-13533130,3	-13533296,1	-13520585,8
M1-I	-11416036,3	-11412648,6	-11417067,6	-11417862,5	-11408537,5
M1-IH	-11474010,8	-11469810,7	-11474111,2	-11475191,7	-11465886,5
M2-AP 91	-14381160,2	-14381104,2	-14386786,2	-14382525,7	-14367660,5
M2-AP 55	-12059163,6	-12058414,2	-12062838,9	-12060613,6	-12049688,2
M2-M1	-10317287,1	-10318089,6	-10318822,4	-10318420,4	-10309816,5
M2-M1H	-10345353,2	-10346618,5	-10347915,0	-10346825,8	-10338701,9
M2-M2	-10151500,5	-10154770,9	-10155459,1	-10154732,8	-10147874,8
M2-M2H	-10144049,0	-10147095,2	-10147773,9	-10147264,5	-10140324,3
M2-M3H	-10139285,3	-10141944,5	-10142725,6	-10141646,5	-10135241,7
M2-M4H	-10316697,8	-10320219,5	-10321131,0	-10319489,6	-10312757,5
M3-S	-11276352,2	-11272779,0	-11277122,7	-11278764,7	-11268757,6
M3-S m2	-11287499,2	-11284432,7	-11288894,1	-11290453,7	-11280466,6
M3-S m3	-11281348,7	-11278039,5	-11282258,4	-11283914,7	-11274063,7
M4-D	-11282822,1	-11280032,5	-11283564,1	-11286017,7	-11276105,9
M4-D m2	-11288210,5	-11285435,4	-11289273,5	-11291488,5	-11281447,0
M4-D m3	-11284666,7	-11281935,7	-11285523,9	-11287869,7	-11277857,9
M5-L	-11278380,3	-11274978,1	-11279279,0	-11281051,3	-11270946,6
M6-E	-11282610,8	-11279492,0	-11284603,9	-11285612,2	-11276220,9
M7-PG A-I	-10649727,1	-10651295,0	-10656748,6	-10654000,1	-10644298,7
M7-PG A-II	-10800425,4	-10801337,9	-10805877,9	-10801876,7	-10793873,5
M7-PG A-III	-10860665,0	-10858670,9	-10862318,7	-10860857,8	-10853092,9
M7-PG B-I	-10737929,9	-10740175,8	-10748100,5	-10746080,6	-10737914,0
M7-PG B-II	-11064098,0	-11064069,4	-11070081,3	-11068514,3	-11058825,4
M7-PG B-III	-11124337,6	-11121402,4	-11126522,1	-11127495,4	-11118044,8
M7-PG C-I	-10556473,5	-10559421,3	-10564962,4	-10562592,0	-10553574,1
M7-PG C-II	-10699673,5	-10702597,1	-10706949,3	-10703024,4	-10695988,3
M7-PG C-III	-10759913,0	-10759930,1	-10763390,1	-10762005,5	-10755207,8
M8-VC A	-10396788,9	-10399796,8	-10403064,1	-10400607,9	-10392654,5
M8-VC A-H	-10451600,4	-10454110,6	-10457581,8	-10455045,1	-10446719,4
M8-VC B	-10339928,3	-10341319,3	-10344277,4	-10342866,9	-10335409,3
M8-VC B-H	-10414086,0	-10414746,6	-10417924,2	-10416733,2	-10408769,2

Tabela C.2: AIC das 5 partições na base do **Jornal Online B**.

	Jornal Online A			**Jornal Online B**		
	Modelos	Média	D.P.	Modelos	Média	D.P.
#1	M2-M3H	-28.376.116,2	5.194,3	M2-M3H	-10.140.168,7	3.038,5
#2	M2-M2	-28.419.249,6	5.552,6	M2-M2H	-10.145.301,4	3.143,5
#3	M2-M2H	-28.423.456,5	5.762,0	M2-M2	-10.152.867,6	3.186,5
#4	M2-M4H	-28.488.502,7	5.232,3	M2-M1	-10.316.487,2	3.771,5
#5	M8-VC B	-29.255.757,5	6.249,8	M2-M4H	-10.318.059,1	3.395,2
#6	M2-M1	-29.256.125,4	5.487,4	M8-VC B	-10.340.760,2	3.407,6
#7	M2-M1H	-29.422.609,9	5.539,8	M2-M1H	-10.345.082,8	3.681,3
#8	M8-VC B-H	-29.446.926,0	5.699,6	M8-VC A	-10.398.582,4	4.000,6
#9	M8-VC A	-29.498.136,1	6.891,7	M8-VC B-H	-10.414.451,8	3.527,6
#10	M8-VC A-H	-29.645.444,0	6.638,5	M8-VC A-H	-10.453.011,5	4.117,5
#11	M7-PG C-I	-30.525.345,3	4.798,3	M7-PG C-I	-10.559.404,7	4.572,6
#12	M7-PG A-I	-30.903.264,5	5.276,7	M7-PG A-I	-10.651.213,9	4.701,9
#13	M7-PG C-II	-30.953.241,4	5.119,8	M7-PG C-II	-10.701.646,5	4.087,5
#14	M7-PG C-III	-31.138.273,9	4.157,9	M7-PG B-I	-10.742.040,2	4.755,2
#15	M7-PG B-I	-31.219.879,2	4.675,5	M7-PG C-III	-10.760.089,3	3.100,5
#16	M7-PG A-II	-31.313.935,1	5.486,0	M7-PG A-II	-10.800.678,3	4.338,1
#17	M7-PG A-III	-31.498.967,6	4.254,2	M7-PG A-III	-10.859.121,0	3.611,3
#18	M7-PG B-II	-32.661.263,5	5.153,6	M7-PG B-II	-11.065.117,7	4.413,2
#19	M7-PG B-III	-32.846.296,0	4.297,7	M7-PG B-III	-11.123.560,4	3.871,0
#20	M3-S	-35.294.311,4	6.713,1	M3-S	-11.274.755,2	4.003,4
#21	M5-L	-35.312.224,9	7.073,2	M5-L	-11.276.927,0	4.006,8
#22	M3-S m3	-35.347.134,3	6.672,9	M3-S m3	-11.279.925,0	3.914,7
#23	M4-D	-35.378.622,9	7.423,3	M6-E	-11.281.707,9	3.855,6
#24	M4-D m3	-35.406.130,6	7.064,2	M4-D	-11.281.708,4	3.789,8
#25	M3-S m2	-35.408.578,7	6.811,7	M4-D m3	-11.283.570,8	3.834,3
#26	M6-E	-35.421.636,6	6.794,5	M3-S m2	-11.286.349,3	3.966,2
#27	M4-D m2	-35.445.001,5	7.017,1	M4-D m2	-11.287.171,0	3.870,3
#28	M1-I	-36.247.850,0	6.847,2	M1-I	-11.414.430,5	3.847,5
#29	M1-IH	-36.447.975,2	6.071,7	M1-IH	-11.471.802,2	3.894,2
#30	M2-AP 55	-39.263.329,7	6.502,5	M2-AP 55	-12.058.143,7	5.018,8
#31	M2-AP 91	-42.680.425,1	10.393,4	M0-U	-13.528.897,2	5.266,9
#32	M0-U	-54.007.909,4	6.968,7	M2-AP 91	-14.379.847,4	7.196,2

Tabela C.3: Ranking dos modelos pelo resultado do AIC pela média das 5 partições para ambos jornais. DP é o desvio padrão das 5 partições.

Apêndice D

Resultados Escore de Brier

Resultados do escore de Brier dos modelos instante a instante nas bases de dados.

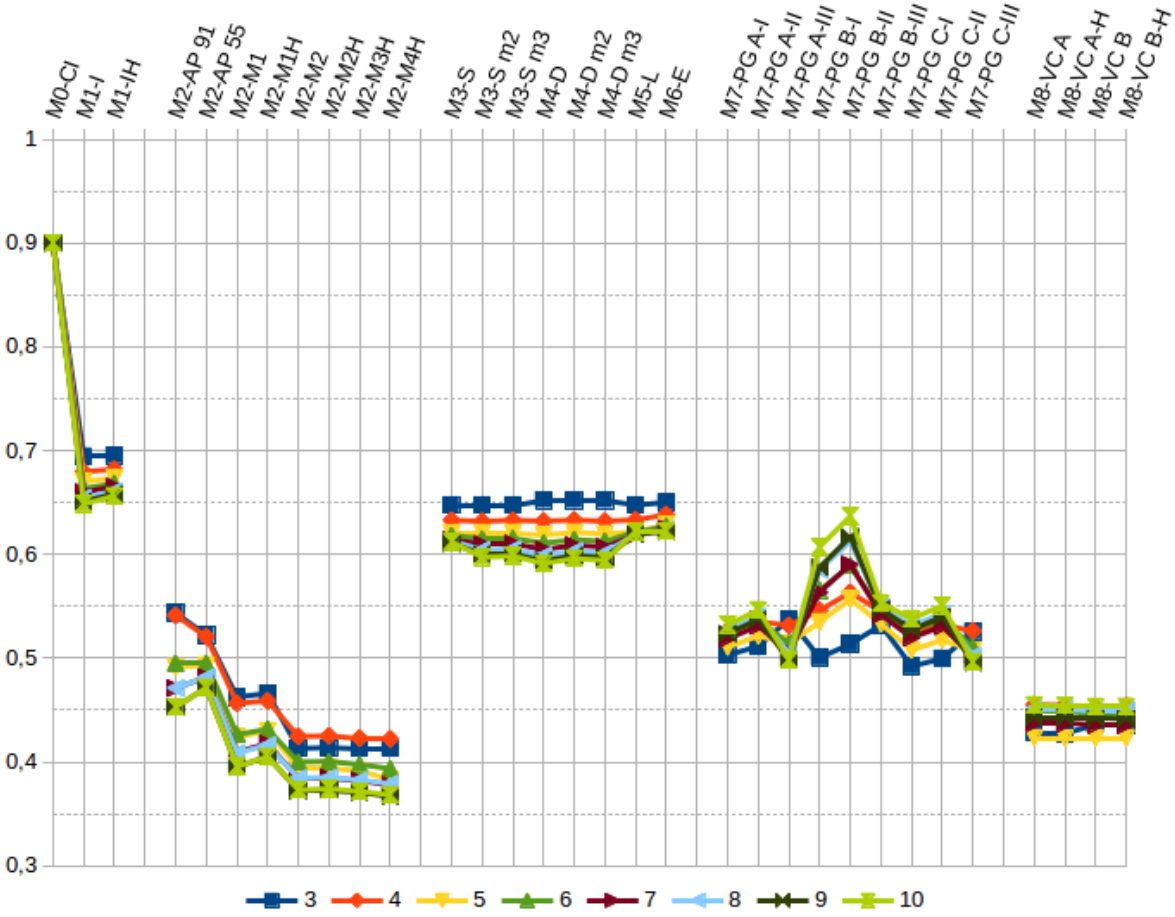


Figura D.1: Resultados do escore de Brier dos modelos na base de dado do **Jornal Online A**, avaliados instante a instante.

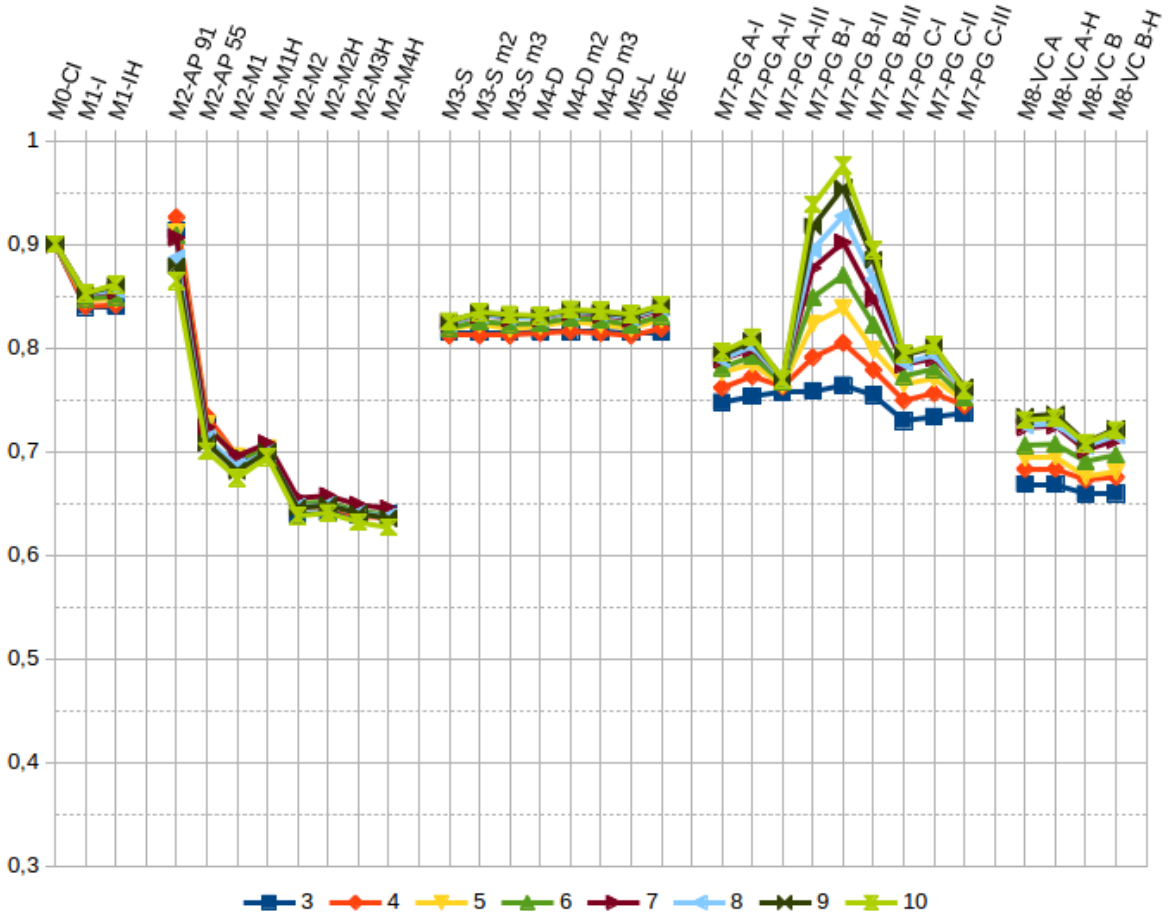


Figura D.2: Resultados do escore de Brier dos modelos na base de dado do **Jornal Online B**, avaliados instante a instante.

D.1 Escore de Brier dos Modelos de Alta Permanência

Os modelos de alta permanência, **M2-AP 91** e **M2-AP 55** obtiveram valores de erro baixos pelo escore de Brier, ficando bem colocados. Esse resultado contrasta com o resultado do critério de informação de Akaike, onde mostra que esses modelos são muito simples e não se ajustam bem aos dados. Para confirmar os últimos resultados vamos explicitar o cálculo do escore de Brier desses dois modelos nas duas bases de dados.

O modelo **M2-AP 91** assume que a probabilidade do próximo tópico só depende do tópico atual, $\mathbb{P}(T_n = l \mid T_1 = l_1, \ldots, T_{n-1} = l_{n-1}) = \mathbb{P}(T_n = l \mid T_{n-1} = l_{n-1})$ e nos fornece as probabilidades $\mathbb{P}(T_n = l \mid T_{n-1} = l_{n-1}) = 0,91$, se $l = l_{n-1}$ e $\mathbb{P}(T_n = l \mid T_{n-1} = l_{n-1}) = 0,01$, caso contrário. A última probabilidade independe para qual tópico está havendo a mudança, logo basta o próximo tópico ser diferente do anterior que a probabilidade é assumida nesse modelo como 0,01. Assim, as 10 probabilidades possíveis, uma probabilidade de permanência no mesmo tópico e 9 probabilidades de mudança para os outros tópicos, somam um: $0,91 + 0,01 * 9 = 1,0$.

O escore de Brier é facilmente calculado nesse modelo pois só há dois valores de probabilidades para as sessões: Para uma sessão do tipo $S = (\ldots, T_{i-1} = x, T_i = x)$ com $x \in \mathcal{L}$, ao avaliarmos o instante i, a parte mais interna do cálculo do escore de Brier é igual a:

$$\begin{aligned}
&\sum_{l \in \mathcal{L}} (f_M(S, i, l) - o(S, i, l))^2 \\
&= (0,91 - 1)^2 + 9 \times (0,01 - 0)^2 \\
&= 0,009
\end{aligned}$$

Agora quando o instante avaliado i é do caso de mudança de tópico, $S = (\ldots, T_{i-1} = x, T_i = y)$, com x e $y \in \mathcal{L}; x \neq y$, o cálculo mias interno do escore de Brier é igual a:

$$\begin{aligned}
&\sum_{l \in \mathcal{L}} (f_M(S, i, l) - o(S, i, l))^2 \\
&= (0,91 - 0)^2 + (0,01 - 1)^2 + 8 \times (0,01 - 0)^2 \\
&= 1,809
\end{aligned}$$

Assim o escore de Brier desse modelo, abusando um pouco da notação, é dado

por:

$$\begin{aligned} BS(i, M) &= \frac{1}{N_i} \sum_{S \in \mathcal{S}; n \geq i} \sum_{l \in \mathcal{L}} (f_M(S, i, l) - o(S, i, l))^2 \\ BS(i \in \{3..10\}, M2 - AP91) &= 0,009\alpha_1 + 1,809\alpha_2 \end{aligned} \tag{D.1}$$

Onde α_1 é a porcentagem dos instantes nas sessões que são do tipo permanência, e α_2 é a porcentagem de instantes avaliados que são do tipo mudança.

No teste do escore de Brier avaliamos os instantes $i \in \{3..10\}$. Para o **Jornal Online A** nesses instantes contabiliza-se o índice de permanência no mesmo tópico em leituras consecutivas igual a 71,7%. Para o **Jornal Online B** o índice de permanência é igual a 49,7%. Logo, o escore de Brier para o modelo **M2-AP 91** na base do **Jornal Online A** é $Brier = 0,009 \times 71,7\% + 1,809 \times 28,3\% =$ **0,5184**. Já o escore de Brier para do modelo na base do **Jornal Online B** é $Brier = 0,009 \times 49,7\% + 1,809 \times 50,3\% =$ **0,914**.

Analogamente, o modelo **M2-AP 55** assume que a probabilidade de permanência é $\mathbb{P}(T_n = l \mid T_{n-1} = l_{n-1}) = 0,55$ e a probabilidade de mudança independente do tópico é $\mathbb{P}(T_n = l \mid T_{n-1} = l_{n-1}) = 0,05$. Logo para esse modelo, o escore de Brier é calculado como:

$$\begin{aligned} BS(i, M) &= \frac{1}{N_i} \sum_{S \in \mathcal{S}; n \geq i} \sum_{l \in \mathcal{L}} (f_M(S, i, l) - o(S, i, l))^2 \\ BS(i \in \{3..10\}, M2 - AP55) &= 0,225\alpha_1 + 1,225\alpha_2 \end{aligned} \tag{D.2}$$

E o escore de Brier para o modelo **M2-AP 55** na base do **Jornal Online A** é $Brier = 0,225 \times 71,7\% + 1,225 \times 28,3\% =$ **0,508**. Já o escore de Brier para do modelo na base do **Jornal Online B** é $Brier = 0,225 \times 49,7\% + 1,225 \times 50,3\% =$ **0,728**.

Todos esses resultados conferem com os obtidos no experimento na seção 5.2. E concluímos do mesmo modo: os modelos são simples e obtiveram erros relativamente baixo por causa dos índices de permanência serem altos. Entretanto são instáveis, quando o índice de permanência diminui, como no caso do segundo jornal, o modelo que atribui o maior valor na permanência piora muito seu resultado.

Referências Bibliográficas

Adomavicius, G. & Tuzhilin, A. (2005). 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*, 17:734--749. ISSN 1041-4347.

Agichtein, E.; Brill, E.; Dumais, S. & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3--10.

Agrawal, M.; Karimzadehgan, M. & Zhai, C. (2009). An online news recommender system for social networks. In *Proceedings of ACM SIGIR workshop on Search in Social Media*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716--723.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199--213. Springer.

Billsus, D. & Pazzani, M. J. (2007). Adaptive news access. In *The Adaptive Web*, pp. 550--570. Springer.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, pp. 353--355.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Eeview*, 78(1):1--3.

Bühlmann, P. (2000). Model selection for variable length markov chains and tuning the context algorithm. *Annals of the Institute of Statistical Mathematics*, 52(2):287--315. ISSN 1572-9052.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370. ISSN 0924-1868.

Candillier, L.; Jack, K.; Fessant, F. & Meyer, F. (2009). State-of-the-art recommender systems. In *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling*, capítulo 1, pp. 1--22. IGI Global.

Candillier, L.; Meyer, F. & Boullé, M. (2007). Comparing state-of-the-art collaborative filtering systems. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition*, MLDM '07, pp. 548--562, Berlin, Heidelberg. Springer-Verlag.

Chen, X.; Irie, K.; Banks, D.; Haslinger, R.; Thomas, J. & West, M. (2015). Bayesian dynamic modeling and analysis of streaming network data. Relatório técnico, Technical Report, Duke University.

Claypool, M.; Gokhale, A.; Miranda, T.; Murnikov, P.; Netes, D. & Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper. In *Proceedings of ACM SIGIR workshop on Recommender Systems*, volume 60. Citeseer.

Das, A. S.; Datar, M.; Garg, A. & Rajaram, S. (2007). Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, pp. 271--280. ACM.

De Francisci Morales, G.; Gionis, A. & Lucchese, C. (2012). From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pp. 153--162. ACM.

Esiyok, C.; Kille, B.; Jain, B.-J.; Hopfgartner, F. & Albayrak, S. (2014). Users' reading habits in online news portals. In *Proceedings of the 5th Information Interaction in Context Symposium*, pp. 263--266. ACM.

Ferri, C.; Hernández-orallo, J. & Flach, P. A. (2011). Brier curves: a new cost-based visualisation of classifier performance. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 585--592.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359--378.

Gomez-Uribe, C. A. & Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):13.

Guimarães, S.; Ribeiro, M. T.; Assunção, R. & Meira Jr, W. (2013). A holistic hybrid algorithm for user recommendation on twitter. *Journal of Information and Data Management*, 4(3):341.

Hernández-Orallo, J.; Flach, P. & Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13(Oct):2813--2869.

Hsieh, C.-K.; Yang, L.; Wei, H.; Naaman, M. & Estrin, D. (2016). Immersive recommendation: News and event recommendations using personal digital traces. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 51--62. International World Wide Web Conferences Steering Committee.

Jahrer, M.; Toscher, A. & Legenstein, R. A. (2010). Combining predictions for accurate recommender systems. In Rao, B.; Krishnapuram, B.; Tomkins, A. & 0001, Q. Y., editores, *KDD*, pp. 693–702. ACM.

Jannach, D.; Zanker, M.; Felfernig, A. & Friedrich, G. (2011). *Recommender Systems An Introduction*. Cambridge University Press, Cambridge.

Kantor, P. B.; Rokach, L.; Ricci, F. & Shapira, B. (2011). *Recommender systems handbook*. Springer.

Koren, Y.; Bell, R.; Volinsky, C. et al. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30--37.

Kumar, R. & Tomkins, A. (2010). A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 561--570. ACM.

Kwak, H.; Lee, C.; Park, H. & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pp. 591--600. ACM.

Li, L.; Chu, W.; Langford, J. & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661--670. ACM.

Linden, G.; Smith, B. & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76--80. ISSN 1089-7801.

Meuth, R.; Robinette, P. & Wunsch, D. (2008). Computational intelligence meets the netflix prize. In *IEEE International Joint Conference on Neural Networks, IJCNN 2008 (IEEE World Congress on Computational Intelligence)*, pp. 686–691. ISSN 1098-7576.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595--600.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323--351.

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P. & Riedl, J. (1994). Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, pp. 175--186, New York, NY, USA. ACM.

Schafer, J. B.; Frankowski, D.; Herlocker, J. & Sen, S. (2007). Collaborative filtering recommender systems. In Brusilovsky, P.; Kobsa, A. & Nejdl, W., editores, *The adaptive web*, capítulo Collaborative filtering recommender systems, pp. 291--324. Springer-Verlag, Berlin, Heidelberg.

Shardanand, U. & Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pp. 210--217, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Wei, K.; Huang, J. & Fu, S. (2007). A survey of e-commerce recommender systems. In *Proceedings of 2007 International Conference on Service Systems and Service Management*, pp. 1--5. IEEE.

Weigend, A. S. (2003). Analyzing customer behavior at amazon.com. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 5--5, New York, NY, USA. ACM.

Zhao, Z.; Cheng, Z.; Hong, L. & Chi, E. H. (2015). Improving user topic interest profiles by behavior factorization. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1406--1416. International World Wide Web Conferences Steering Committee.