UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

JOSÉ LUIZ MATURANA PAGNOSSIM

Uma abordagem híbrida para sistemas de recomendação de notícias

São Paulo
2018

JOSÉ LUIZ MATURANA PAGNOSSIM

Uma abordagem híbrida para sistemas de recomendação de notícias

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 09 de abril de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Profa. Dra. Sarajane Marques Peres

São Paulo

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO
(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB 8 - 4936

Pagnossim, José Luiz Maturana
Uma abordagem híbrida para sistemas de recomendação de notícias / José Luiz Maturana Pagnossim ; orientadora, Sarajane Marques Peres. - 2018.
119 f. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.
Versão corrigida

1. Mineração de dados. 2. Notícias. I. Peres, Sarajane Marques, orient. II. Título

CDD 22.ed.– 006.312

Dissertação de autoria de José Luiz Maturana Pagnossim, sob o título **"Uma abordagem híbrida para sistemas de recomendação de notícias"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 09 de abril de 2018 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Sarajane Marques Peres

Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

Presidente

Prof. Dr. Edmir Parada Vasques Prado

Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

Prof. Dr. Ismar Frango Silveira

Instituição: Universidade Presbiteriana Mackenzie

Profa. Dra. Sahudy Montenegro González

Universidade Federal de São Carlos

À minha mãe Maria de Lourdes pela fé depositada e carinho ao logo de toda minha vida. Ao meu pai José Luiz pelo incentivo à educação e ao trabalho e por estar presente sempre que mais precisava. Aos meus filhos Maria Eduarda e Pedro Henrique por me proporcionarem a felicidade plena e me encherem de orgulho em cada etapa da vida. À minha esposa Luciana pelo amor verdadeiro, cumplicidade e apoio incondicional nos momentos mais importantes da minha vida. A todos vocês, o meu respeito e amor eterno.

Agradecimentos

À minha orientadora, Profa. Dra. Sarajane Marques Peres, por compartilhar seus conhecimentos e me conduzir na direção correta em busca da qualidade acadêmica e científica.

Aos professores da EACH/USP pelos ensinamentos ao longo deste mestrado.

Aos colegas de estudo pelo companheirismo, auxílio mútuo e amizade que carregarei por toda minha vida, em especial ao Sérgio Barbieri, Robinson Cruz, Henrique Passos, Rodrigo Mansho, André Lima, Andrei Martins, Fernando Henrique, Alexandra Katiuska, Arthur Gustavo da Cruz e Helena Rossi.

Aos meus colegas de trabalho pela compreensão e flexibilidade concedida durante o período em que estive imerso nesta pesquisa.

Aos voluntários que participaram do experimento contribuindo para que esta pesquisa fosse viabilizada e também para a evolução da ciência no país.

E finalmente, agradeço a Deus, que me manteve em pé nos momentos mais difíceis desta longa jornada.

"A melhor maneira de prever o futuro é inventá-lo."

(Alan Curtis Kay)

Resumo

PAGNOSSIM, José Luiz Maturana. **Uma abordagem híbrida para sistemas de recomendação de notícias**. 2018. 119 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Sistemas de Recomendação (*SR*) são softwares capazes de sugerir itens aos usuários com base no histórico de interações de usuários ou por meio de métricas de similaridade que podem ser comparadas por item, usuário ou ambos. Existem diferentes tipos de *SR* e dentre os que despertam maior interesse deste trabalho estão: *SR* baseados em conteúdo; *SR* baseados em conhecimento; e *SR* baseado em filtro colaborativo. Alcançar resultados adequados às expectativas dos usuários não é uma meta simples devido à subjetividade inerente ao comportamento humano, para isso, *SR* precisam de soluções eficientes e eficazes para: modelagem dos dados que suportarão a recomendação; recuperação da informação que descrevem os dados; combinação dessas informações dentro de métricas de similaridade, popularidade ou adequabilidade; criação de modelos descritivos dos itens sob recomendação; e evolução da inteligência do sistema de forma que ele seja capaz de aprender a partir da interação com o usuário. A tomada de decisão por um sistema de recomendação é uma tarefa complexa que pode ser implementada a partir da visão de áreas como inteligência artificial e mineração de dados. Dentro da área de inteligência artificial há estudos referentes ao método de raciocínio baseado em casos e da recomendação baseada em casos. No que diz respeito à área de mineração de dados, os *SR* podem ser construídos a partir de modelos descritivos e realizar tratamento de dados textuais, constituindo formas de criar elementos para compor uma recomendação. Uma forma de minimizar os pontos fracos de uma abordagem, é a adoção de aspectos baseados em uma abordagem híbrida, que neste trabalho considera-se: tirar proveito dos diferentes tipos de *SR*; usar técnicas de resolução de problemas; e combinar recursos provenientes das diferentes fontes para compor uma métrica unificada a ser usada para ranquear a recomendação por relevância. Dentre as áreas de aplicação dos *SR*, destaca-se a recomendação de notícias, sendo utilizada por um público heterogêneo, amplo e exigente por relevância. Neste contexto, a presente pesquisa apresenta uma abordagem híbrida para recomendação de notícias construída por meio de uma arquitetura implementada para provar os conceitos de um sistema de recomendação. Esta arquitetura foi validada por meio da utilização de um *corpus* de notícias e pela realização de um experimento *online*. Por meio do experimento foi possível observar a capacidade da arquitetura em relação aos requisitos de um sistema de recomendação de notícias e também confirmar a hipótese no que se refere à privilegiar recomendações com base em similaridade, popularidade, diversidade, novidade e serendipidade. Foi observado também uma evolução nos indicadores de leitura, curtida, aceite e serendipidade conforme o sistema foi acumulando histórico de preferências e soluções. Por meio da análise da métrica unificada para ranqueamento foi possível confirmar sua eficácia ao verificar que as notícias melhores colocadas no ranqueamento foram as mais aceitas pelos usuários.

Palavras-chaves: Sistemas de Recomendação. Mineração de Dados. Mineração de Texto. Raciocínio Baseado em Casos. Recomendação Baseada em Casos. Recomendação Baseada em Conteúdo. Recomendação Baseada em Conhecimento. Filtro Colaborativo. Recomendação de Notícias. Arquitetura de Recomendação Híbrida.

Abstract

PAGNOSSIM, José Luiz Maturana. **A hybrid approach to news recommendation systems**. 2018. 119 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Recommendation Systems (*RS*) are software capable of suggesting items to users based on the history of user interactions or by similarity metrics that can be compared by item, user, or both. There are different types of *RS* and those which most interest in this work are content-based, knowledge-based and collaborative filtering. Achieving adequate results to user's expectations is a hard goal due to the inherent subjectivity of human behavior, thus, the *RS* need efficient and effective solutions to: modeling the data that will support the recommendation; the information retrieval that describes the data; combining this information within similarity, popularity or suitability metrics; creation of descriptive models of the items under recommendation; and evolution of the systems intelligence to learn from the user's interaction. Decision-making by a *RS* is a complex task that can be implemented according to the view of fields such as artificial intelligence and data mining. In the artificial intelligence field there are studies concerning the method of case-based reasoning that works with the principle that if something worked in the past, it may work again in a new similar situation the one in the past. The case-based recommendation works with structured items, represented by a set of attributes and their respective values (within a "case" model), providing known and adapted solutions. Data mining area can build descriptive models to *RS* and also handle, manipulate and analyze textual data, constituting one option to create elements to compose a recommendation. One way to minimize the weaknesses of an approach is to adopt aspects based on a hybrid solution, which in this work considers: taking advantage of the different types of *RS*; using problem-solving techniques; and combining resources from different sources to compose a unified metric to be used to rank the recommendation by relevance. Among the *RS* application areas, news recommendation stands out, being used by a heterogeneous public, ample and demanding by relevance. In this context, the this work shows a hybrid approach to news recommendations built through a architecture implemented to prove the concepts of a recommendation system. This architecture has been validated by using a news *corpus* and by performing an *online* experiment. Through the experiment it was possible to observe the architecture capacity related to the requirements of a news recommendation system and architecture also related to privilege recommendations based on similarity, popularity, diversity, novelty and serendipity. It was also observed an evolution in the indicators of reading, likes, acceptance and serendipity as the system accumulated a history of preferences and solutions. Through the analysis of the unified metric for ranking, it was possible to confirm its efficacy when verifying that the best classified news in the ranking was the most accepted by the users.

Keywords: Recommendation Systems. Data Mining. Text Mining. Case Based Reasoning. Case Based Recommendation. Content Based Recommendation. Knowledge Based Recommendation. Collaborative Filtering. News Recommendation. Hybrid Recommender Architecture.

Lista de figuras

Lista de algoritmos

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

ACM	Association for Computing Machinery
ADO	Activex Data Objects
ASC	American Standard Code for Information Interchange
BNDES	Banco Nacional de Desenvolvimento
CRUD	Create, Read, Update and Delete
CSS	Cascading Style Sheets
EBC	Empresa Brasil de Comunicação
HTML	HyperText Markup Language
K-NN	K-Nearest Neighbor
IAA	Indicador de Avaliação da Arquitetura
ISIL	Índice Silhouette
IR	Índice RAND
IRA	Índice RAND Ajustado
NLTK	Natural Language Toolkit
RBC	Recomendação Baseada em Casos
SQL	Structured Query Language
SR	Sistemas de Recomendação
SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
TM	Text Mining
UML	Unified Modeling Language
UTF-8	8-bit Unicode Transformation Format

Lista de símbolos

k	número de grupos usado no algoritmo *k-means*
D	um conjunto de dados de entrada usado no algoritmo *k-means*
n	objetos contidos no conjunto de dados D
C	um conjunto de dados de saída usado no algoritmo *k-means*
i	um item contido em um grupo - usado no índice *silhouette*
$a(i)$	é a distância média de i a todos demais itens do seu grupo
$b(i)$	distância mínima de i a todos os itens que não pertencem ao seu grupo
Cob	um conjunto de grupos obtidos por meio do algoritmo de agrupamento
$Cref$	um conjunto de grupos de referência, conhecidos a priori
A	quantidade de pares de exemplares Cob iguais e $Cref$ iguais
B	quantidade de pares de exemplares Cob iguais e $Cref$ diferentes
C	quantidade de pares de exemplares Cob diferentes e $Cref$ iguais
D	quantidade de pares de exemplares Cob diferentes e $Cref$ diferentes
Esp	indica o valor esperado do IR ao comparar as partições
m	número de critérios de recuperação de notícias
Pos	posição em que a notícia ficou colocada no ranqueamento
$MURR$	métrica unificada de ranqueamento da recomendação

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Sistemas de Recomendação (*SR*) são softwares nos quais implementa-se uma ou mais técnicas que são capazes de sugerir itens para os usuários (RICCI et al., 2011). Mais especificamente, *SR* são sistemas capazes de prover recomendações: com base no histórico de interações entre os usuários e o sistema; entre os usuários e os itens presentes no sistema; e com base em métricas de similaridade entre os itens ou entre os usuários.

Os *SR* podem ser usados para diferentes propósitos, dependendo da necessidade do negócio, da especificidade do processo de recomendação ou ainda das características do domínio de dados usados para a recomendação. Para atender a tais propósitos, existem diferentes tipos de *SR*, dentre os mais comuns estão os *SR* baseados em filtro colaborativo, os *SR* baseados em conteúdo e os *SR* baseados em conhecimento.

Nos *SR* por filtro colaborativo, o ponto-chave de obtenção de informação para a recomendação está na colaboração dos usuários com o sistema. Neste sentido, Brunialti et al. (2015) define que a filtragem colaborativa permite recomendar ao usuário atual, os itens que outros usuários similares a ele gostaram no passado. Já nos *SR* baseados em conteúdo, as estratégias de recomendação consideram a descrição dos itens para encontrar itens semelhantes, de forma a fazer recomendações com base na similaridade entre itens (BRUNIALTI et al., 2015; LOPS; GEMMIS; SEMERARO, 2011). A recomendação baseada em conteúdo considera dados não estruturados a respeito do item sob recomendação, por exemplo o texto completo (conteúdo) de uma notícia. Finalmente, os *SR* baseados em conhecimento criam as recomendações com base em informações de domínio e de preferências dos usuários (RICCI et al., 2011).

Desde sua criação, a área de *SR* teve uma rápida expansão devido à evolução de aplicações que recomendam filmes, páginas web, notícias, tratamentos médicos, músicas e produtos em geral (RICCI et al., 2011). No que se refere à recomendação de notícias, Kunaver e Požrl (2016) define que as pessoas leem notícias para saber o que aconteceu e o que acontecerá em uma cidade, região, país ou no mundo. Essa demanda por informação apresenta-se como um desafio para os provedores de notícias. Eles precisam ajudar os usuários a encontrar eficientemente as notícias que lhes interessam, já que a seleção manual de notícias interessantes não são viáveis dentro das restrições de tempo comuns para a maioria dos usuários. Neste cenário, os *SR* destacam-se como uma solução possível para

auxiliar os leitores na seleção do conteúdo que melhor se adapta às suas preferências. A recomendação de notícias é utilizada por um público heterogêneo cada vez mais exigente, principalmente devido à evolução dos meios de interação entre usuários e sistemas atualmente disponíveis no estado da prática, sendo o aumento da exigência por qualidade na recomendação, um consenso na área de *SR*.

Cada um dos tipos de *SR* apresenta características que limitam o atendimento à expectativa do usuário, incorrendo em problemas como *cold start* de item e de usuário. Este problema descreve situações em que o recomendador é incapaz de fornecer recomendações significativas devido a uma falta inicial de avaliações, degradando significativamente a performance de recomendadores baseados em filtro colaborativo (BRUSILOVSKY, 2007). Outros problemas estão relacionados à previsibilidade, aleatoriedade e falta de conhecimento das preferências dos usuários. Estes problemas podem ser facilmente detectados por uma simples navegação em recomendadores de notícias disponíveis no estado da prática. Nesse contexto, tem sido proposta a adoção de abordagens híbridas visando minimizar as limitações e resolver os problemas inerentes a cada tipo de *SR*.

Para alcançar resultados adequados às expectativas dos usuários, os *SR* são construídos a partir da aplicação de técnicas capazes de resolver problemas que envolvem a predição da recomendação desejada pelo usuário. No entanto, devido às características subjetivas que descrevem o comportamento humano, essa meta não é simples de ser alcançada. Tais sistemas precisam, portanto, apresentar soluções eficientes e eficazes para: modelagem dos dados que suportarão a predição da recomendação; recuperação da informação inerente a todos os atributos que descrevem os dados; combinação desta informação dentro de métricas de similaridade, relevância ou adequabilidade[1]; criação de modelos preditivos ou descritivos para elaboração da recomendação; e implementação de mecanismos de evolução da inteligência do sistema de forma que ele seja capaz de aprender a partir da interação com o usuário. Ambientes deste tipo apresentam problemas de difícil tratamento e, no presente trabalho, tal complexidade será estudada a partir da visão de áreas como inteligência artificial e mineração de dados.

Dentro da área de inteligência artificial encontram-se estudos referentes ao método de raciocínio baseado em casos. Este método trabalha com o princípio que, se algo funcionou no passado em uma determinada situação, muito provavelmente pode funcionar novamente

[1] O conceito de adequabilidade, neste trabalho, é usado para denotar a qualidade daquilo que está de acordo com as exigências de um contexto, usuário ou ambiente.

em uma nova situação similar àquela passada (KOLODNER, 1993). O raciocínio baseado em casos usa uma associação entre problemas e soluções. As soluções podem ser conhecidas historicamente ou ainda adaptadas para fornecer resoluções a novos problemas (LENZ et al., 1998).

Seguindo os pressupostos discutidos pelo método de raciocínio baseado em casos, Smyth (2007) propõe uma abordagem para uso deste método em problemas de recomendação, apresentando a recomendação baseada em casos (*RBC*). O mesmo autor advoga que a *RBC*, apesar de ser considerada um tipo particular de recomendação baseada em conteúdo, consegue tratar itens de forma estruturada, representados por um conjunto de atributos e seus respectivos valores (modelo denominado "caso"). A recomendação baseada em casos também trata o problema de avaliação de similaridade e objetiva encontrar o produto mais parecido com o que o usuário tem em mente, considerando que o que conta como semelhante muitas vezes envolve conhecimentos específicos do domínio (SAQUIB; SIDDIQUI; ALI, 2017).

No que diz respeito à área de mineração de dados, os *SR* podem ser construídos a partir de modelos preditivos ou descritivos (HAN; KAMBER; PEI, 2011; SILVA; PERES; BOSCARIOLI, 2016). Os *SR* podem também proporcionar o tratamento, manipulação e análise de diferentes formatos de dados, incluindo formatos não estruturados, como imagem, vídeo, áudio e texto. Uma porção substancial das informações disponíveis atualmente são armazenadas em bases de dados textuais, as quais consistem em uma vasta coleção de documentos de várias fontes, como por exemplo notícias, artigos, livros, mensagens de e-mail e páginas na internet (HAN; KAMBER; PEI, 2011). Assim, se o conteúdo dos itens está disponível, as tarefas de mineração podem oferecer também uma modelagem adequada para tratamento de problemas de recomendação.

A seleção de uma abordagem para ser aplicada na construção de um sistema de recomendação deve considerar pontos positivos e negativos, limitações e a adequabilidade desta abordagem para atender a uma determinada finalidade. Essa escolha pode beneficiar um aspecto da recomendação em detrimento de outro. Uma alternativa que visa minimizar os pontos fracos de uma abordagem é a adoção de aspectos híbridos. O caráter híbrido pode ser de diferente natureza e neste trabalho foi considerado: tirar proveito dos diferentes tipos de *SR*, principalmente *SR* por filtro colaborativo, *SR* baseado em conteúdo e conhecimento; usar diferentes técnicas de resolução de problemas (mineração de dados e *RBC*); e combinar

a informação proveniente de diferentes fontes, compondo uma métrica unificada usada para recomendar uma lista ordenada por relevância.

Recomendar atendendo adequadamente às expectativas dos usuários sob diferentes perspectivas é um problema em aberto e atacá-lo utilizando soluções híbridas, como a combinação de técnicas de recomendação e a aplicação de diferentes formas de similaridade, ainda precisam ser mais bem avaliadas. Problemas como *cold start*, previsibilidade e aleatoriedade, assim como desafios relacionados à novidade, diversidade e serendipidade, merecem ser mais investigados e se justifica o estudo de formas para superá-los. Diante desse panorama, a presente pesquisa apresenta uma abordagem híbrida para recomendação de notícias que privilegia similaridade, popularidade, diversidade, novidade e serendipidade, fazendo uso de diferentes estratégias e critérios para gerar a recomendação, podendo ser reusada ou adaptada no futuro.

1.1 Hipótese

Se a utilização adequada de uma abordagem híbrida para recomendação de notícias for bem sucedida, o sistema de recomendação deve equilibrar as limitações em relação ao uso de métodos isolados e fornecer melhores recomendações de acordo com estratégias específicas aplicadas em diferentes contextos.

Se o sistema de recomendação utilizar conhecimento com base em preferências do usuário e for capaz de aprender com a *RBC* por meio de soluções conhecidas e adaptadas, os indicadores de leitura, curtida, aceite de recomendação e serendipidade devem apresentar melhores resultados de acordo com a evolução da aprendizagem do sistema

Se a aplicação de um método de ranqueamento por relevância for eficaz, as notícias mais bem colocadas no ranqueamento devem ser as mais aceitas pelos usuários.

1.2 Objetivos

Esta pesquisa tem como objetivo geral demonstrar que a adoção de uma abordagem híbrida melhora aspectos relacionados à recomendação de notícias. Melhorias estas medidas em termos de: quantidade de notícias lidas; quantidade de notícias curtidas; quantidade de aceites de recomendação; nota de avaliação da recomendação; e serendipidade (quantidade

de recomendações que causaram surpresa positiva nos usuários). Como ambiente de prova de conceito foi construído um protótipo de sistema de recomendação com notícias reais obtidas de um portal público de notícias.

Diante do intuito de demonstrar a eficácia da abordagem híbrida e de aplicá-la ao contexto de recomendação de notícias, um objetivo específico da pesquisa é a formulação e avaliação de uma métrica unificada para ranqueamento da recomendação que seja capaz de combinar critérios de popularidade, similaridade do texto da notícia, similaridade de grupos de notícias, similaridade entre usuários, histórico de navegação do usuário, aceite de recomendação, nota média da recomendação, aleatoriedade e serendipidade.

Como objetivos técnicos, este projeto disponibiliza entregáveis que contribuem adicionalmente aos objetivos já descritos:

- Uma arquitetura para *SR* composta por artefatos elaborados durante a condução deste trabalho, como: diagramas, modelos, especificações, algoritmos e um protótipo de sistema de recomendação.
- Um *corpus* de notícias reais em idioma português do Brasil, disponibilizado em formatos que facilitam seu reuso.
- Um conjunto de dados de navegação de usuários produzidos em um experimento *online* que envolveu a participação de mais de cem pessoas.

1.3 Método de pesquisa

O método de pesquisa utilizado neste trabalho está enquadrado nos seguintes aspectos da metodologia científica: gênero, natureza, abordagem, procedimentos técnicos, fonte de dados, técnicas e instrumentos de coleta de dados, e técnicas de análise de dados.

Esta pesquisa utilizou conhecimentos científicos advindos das áreas da inteligência artificial, mineração de dados e *SR* para fins explícitos de construção de uma ferramenta para uso e intervenção na realidade, e portanto constituiu-se como sendo do gênero prático e natureza aplicada. Como procedimentos para o levantamento do conhecimento das áreas de estudo foram realizados estudos exploratórios e buscas bibliográficas baseadas em métodos de revisão sistemática da literatura.

A avaliação da abordagem é considerada mista, fazendo uso de instrumentos quantitativos e qualitativos. Como instrumentos quantitativos são considerados os índices

de avaliação de tarefas descritivas da mineração de dados (agrupamento) representados pela validação interna e externa do agrupamento.

Diante do exposto, esta pesquisa é caracterizada por procedimentos experimentais, já que usa modelos de mineração de texto em um ambiente sistemático no qual numerosos testes foram feitos para aferição de resultados e também pela realização do experimento envolvendo a participação de voluntários. A pesquisa também é caracterizada por ciência do projeto, uma vez que visa resolver um problema prático levantado a partir da identificação de uma necessidade de negócio, materializando uma ponte entre a ciência e a ação prática.

1.4 Organização deste documento

Este trabalho está organizado em seis capítulos, considerando esta introdução. Os demais capítulos estão divididos da seguinte forma:

- No capítulo 2 é apresentada a fundamentação teórica, com ênfase nos *SR*. São discutidos os aspectos da mineração de dados aplicados em *SR*. Os métodos de raciocínio baseado em casos e *RBC* são fundamentados. O capítulo apresenta as formas de avaliação dos *SR* e o caráter híbrido aplicado neste trabalho. O capítulo traz ainda as pesquisas relacionadas aos temas levantados por este trabalho.
- No capítulo 3 é descrita a abordagem de recomendação aplicada ao domínio de notícias, elaborada para resolver o problema de pesquisa. A abordagem considera a implementação de uma arquitetura para *SR* e apresenta as ferramentas utilizadas ou construídas para apoiar a resolução do problema.
- No capítulo 4 são apresentados os recursos utilizados para validar a abordagem, incluindo: a apresentação do *corpus*; a justificativa e explicação do experimento realizado; a apresentação e discussão dos resultados; e as limitações e ameaças.
- No capítulo 5 são apresentadas: a conclusão da pesquisa; as contribuições; e as indicações de trabalhos futuros.
- Por fim, nos apêndices são apresentadas páginas e formulários utilizados no experimento com objetivo de: solicitar a identificação do participante; informar e solicitar o aceite do participante em relação ao termo de consentimento livre e esclarecido; solicitar a escolha de uma opção para navegação nas telas do experimento; e apresentar um protocolo para disponibilização do *corpus* e do conjunto de indicadores.

2 Referencial teórico

Na área de *SR*, os conceitos fundamentais a serem abordados são: os tipos; as funções; as tarefas; os procedimentos e tarefas de mineração de dados aplicados a sistemas desta natureza; e as formas de avaliação destes sistemas. Além destes conceitos, esta pesquisa tem interesse em abordar os fundamentos sobre os métodos de raciocínio baseado em casos e *RBC*, e o aspecto híbrido utilizado por este trabalho. Ao final deste capítulo são apresentadas as pesquisas correlatas com ênfase nas técnicas, métodos e avaliações que caracterizam o aspecto híbrido. São apresentados também os trabalhos relacionados que tratam os principais desafios e problemas dos *SR* e por fim são observados os trabalhos voltados ao domínio de notícias.

2.1 Sistemas de Recomendação

Em meados da década de 1990, a área de *SR* passou a ser reconhecida como uma área de pesquisa independente e desde então estudos específicos e correlatos cresceram consideravelmente, motivados principalmente por dois fatores: o primeiro está relacionado com os casos de sucessos dos sites de empresas como Amazon, Netflix, Youtube, Yahoo, Tripadvisor, entre outros, que passaram a usar *SR* nos seus respectivos negócios; e o segundo tem a ver com o incentivo por parte de associações de pesquisa, como a ACM[1] que passou a organizar congressos e workshops específicos em *SR* (RICCI et al., 2011).

SR são ferramentas capazes de fazer sugestões de "itens" de diversas áreas dentro de um processo de tomada de decisão por parte dos usuários (RICCI et al., 2011). Usuários leigos ou especializados quando aceitam sugestões fornecidas pelos *SR*, se beneficiam das recomendações, principalmente das que foram formuladas especificamente para eles e, ao mesmo tempo, colaboram com o aprendizado do sistema, que registra as interações deles afim de melhorar as recomendações futuras, seja para o mesmo usuário ou para outros.

O pesquisador Jacob Nielsen[2] declarou como sendo a primeira lei do comércio eletrônico: "se o usuário não consegue encontrar o produto ele não poderá comprá-lo". Apesar desta afirmação parecer óbvia, não são poucas vezes em que o usuário tem dificuldade

[1] Association for Computing Machinery.
[2] Nielsen declarou esta lei em seu Alertbox em 2003.

de encontrar um produto de interesse. Este tipo de deficiência sistêmica causa impacto negativo na satisfação do cliente, podendo culminar na troca de um sistema por outro.

Este raciocínio pode ser aplicado a outros domínios de negócio, como o de notícias, em que se a notícia não for recomendada adequadamente ao leitor, ele não a lerá. Para compreender com mais profundidade os *SR* é importante conhecer características a eles associadas, em termos das funções, tarefas, tipos, métodos e as formas de avaliação.

Funções de um sistema de recomendação

O termo funções, refere-se aos possíveis papéis em que um sistema de recomendação pode atuar, em outras palavras, por quais motivos uma empresa, instituição ou um pesquisador utilizaria esse recurso tecnológico. Para que se tenha uma visão introdutória desta vasta área de pesquisa Ricci et al. (2011) destacam as seguintes funções dos *SR*: aumentar o número de itens vendidos; vender itens mais diversificados; aumentar a satisfação do usuário; aumentar a fidelização de clientes; melhorar o conhecimento das preferências do usuário.

Apesar desta definição ser originalmente aplicada ao comércio eletrônico (compra e venda de produtos na internet), ela também pode ser adequadamente aplicada ao domínio de notícias por meio de uma adaptação destas funções: aumentar o número de notícias lidas (consumidas); ter notícias mais diversificadas sendo aceitas; aumentar a satisfação do leitor e do usuário; aumentar a fidelização de clientes; e melhorar o conhecimento das preferências do usuário.

Tarefas de um sistema de recomendação

Apesar das funções descritas anteriormente também influenciarem positivamente a experiência dos usuários com os *SR*, elas foram prioritariamente idealizadas para servir aos interesses dos proprietários dos *SR*, pois fazem referência aos ganhos que podem ser otimizados por meio de melhorias dos *SR* em seus negócios. Para que a relação entre os interesses dos proprietários e dos usuários seja mais equilibrada, Herlocker et al. (2004) apud Ricci et al. (2011) definem as tarefas que podem ajudar a balancear tais interesses:

- Encontrar alguns itens bons: recomendar ao usuário uma lista dos itens que outros usuários mais gostam.
- Anotações no contexto: são anotações realizadas em segundo plano pelos *SR* para registrar dados dentro de um determinado contexto, por exemplo a coleta de cliques do usuário durante a utilização de um sistema.
- Melhorar o perfil: tarefa relacionada com a capacidade do sistema em obter informações do perfil do usuário, incluindo o que ele gosta e não gosta.

Tipos de recomendação

Com a evolução das pesquisas e a ampliação do uso dos *SR* em diversas áreas, a decomposição dos *SR* passou a ser fundamental para permitir o maior aprofundamento dos estudos em áreas mais específicas. São aplicados neste trabalho os conceitos de três diferentes tipos de *SR*: baseado em conteúdo; baseado em conhecimento; e baseado em filtro colaborativo.

- *SR* baseado em conteúdo - sistemas deste tipo consideram as características (conteúdo) de um determinado item consumido no passado para sugerir itens semelhantes (RICCI et al., 2011). A recomendação baseada em conteúdo usa principalmente palavras-chaves como atributo do item. Estas palavras são normalmente extraídas da descrição textual do item. A similaridade entre os itens é calculada com base nos atributos dos itens comparados (RICCI; ROKACH; SHAPIRA, 2015).
- *SR* baseado em conhecimento - estes sistemas recomendam itens com base em conhecimentos específicos a respeito do domínio e das características do item (atributos mais estruturados), sugerindo itens que sejam relevantes e atendam às preferências dos usuários (RICCI et al., 2011). Apesar de haver uma divergência conceitual sobre qual tipo de sistema de recomendação a *RBC* foi derivada, (RICCI; ROKACH; SHAPIRA, 2015) defende que os recomendadores baseados em casos são tipos particulares dos *SR* baseados em conhecimento, já que utilizam soluções conhecidas e registradas no passado para prover recomendações que resolvam novos problemas.
- *SR* baseado em filtro colaborativo - esta abordagem recomenda itens ao usuário corrente, que outros usuários (com preferências similares) gostaram no passado (RICCI et al., 2011). A recomendação baseada em filtro colaborativo produz recomendações

de itens específicos ao usuário com base em avaliações, por meio de interações ou pelo próprio uso do sistema, sem necessidade do sistema possuir muitas informações a respeito dos usuários ou dos itens (KOREN; BELL, 2015). Este tipo de recomendador pode usar a similaridade entre os itens ou entre os usuários do sistema para fazer a recomendação. No contexto deste trabalho, interessa fundamentar uma abordagem conhecida como modelo de vizinhança usuário-usuário. Este modelo, descrito por Koren e Bell (2015), faz predições com base na avaliação e consumo de itens por usuários com ideias semelhantes.

Métodos para recomendação: mineração de dados

Tarefas de mineração de dados vêm sendo usadas para melhorar a qualidade dos *SR*. A escolha destes métodos não é uma tarefa fácil, tendo em vista a variedade de opções disponíveis, principalmente para selecionar, adaptar e aplicar a técnica mais apropriada em razão da necessidade da pesquisa ou do negócio.

Dentre as tarefas para mineração de dados, destacam-se duas que podem ser aplicadas separada ou conjuntamente: classificação e agrupamento. Neste trabalho, foi aplicada a tarefa de agrupamento, que foi precedida por um conjunto de procedimentos de pré-processamento de dados textuais.

Pré-processamento de dados

Em mineração de texto, a resolução para as tarefas de classificação e agrupamento devem ser precedidas por uma etapa de pré-processamento dos dados textuais, afim de garantir maior eficácia do algoritmo de mineração (SILVA; PERES; BOSCARIOLI, 2016).

A necessidade de utilização de grandes volumes de dados reais aplicados em um negócio requer a aplicação de algoritmos para pré-processar estes dados, afim de otimizar a análise dos dados (AMATRIAIN et al., 2011).

Ribeiro-Neto e Baeza-Yates (2011) definem que a realização da etapa de pré-processamento de texto deve aplicar em linhas gerais os seguintes procedimentos :

- Análise Léxica (*tokenizing*): tratamento de pontuação, acentos, letras maiúsculas ou minúsculas e determinação de um separador de termos.

- Eliminação de palavras irrelevantes (*stop words*): remover artigos, preposições, pronomes, numerais, conjunções, advérbios e palavras comuns para o contexto.
- Redução dos termos aos radicais (*stemming*): retirar plural, gerúndio, verbos flexionados, aumentativo e diminutivo.
- Criação de uma representação vetorial para os textos baseada em frequência de termos ou frequência invertida de termos do *corpus*.

Há diversas ferramentas que fornecem suporte para implementação dos procedimentos de pré-processamento de texto, entre as mais importantes estão: PreTexT - desenvolvida pelo Laboratório de Inteligência Computacional[3] do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo; NLTK[4] - uma biblioteca oferecida para ambientes de programação em Python; e TM[5], - uma biblioteca disponível para desenvolvimento em ambiente de programação R (*R-Project*)[6] (DIAZ et al., 2018).

Agrupamento

Ao contrário das tarefas de classificação e regressão, que analisam os conjuntos de dados rotulados (treinamento), o agrupamento analisa os objetos de dados sem consultar os rótulos das classes. Em situações em que os dados rotulados não estão disponíveis, a tarefa de agrupamento pode ser usada para gerar grupos de dados. Os objetos são agrupados com base no princípio de maximizar a similaridade intraclasse e minimizar a similaridade interclasses (HAN; KAMBER; PEI, 2011).

A tarefa de agrupamento é considerada não supervisionada, pois não conhece previamente os rótulos, nesse sentido, usa medidas de distância para agrupar os elementos parecidos (AMATRIAIN et al., 2011).

Um dos algoritmos mais usados em tarefas de agrupamento é o *k-means* que adota métodos eurísticos, como abordagens gulosas, que melhoram progressivamente a qualidade do grupo e se aproximam da solução ótima. Este tipo de algoritmo de agrupamento funciona bem para encontrar grupos em forma esférica em bancos de dados de tamanho pequeno a médio (HAN; KAMBER; PEI, 2011).

[3] http://labic.icmc.usp.br
[4] *Natural Language ToolKit* (LOPER; BIRD, 2002)
[5] Biblioteca para mineração de textos (*package* tm) (FEINERER; HORNIK, 2017)
[6] https://cran.r-project.org/mirrors.html

O algoritmo *k-means* recebe como entrada o número de grupos (k) e um conjunto de dados (D) contendo n objetos. A saída é um conjunto (C) com k grupos. O núcleo do método é especificado no algoritmo 1.

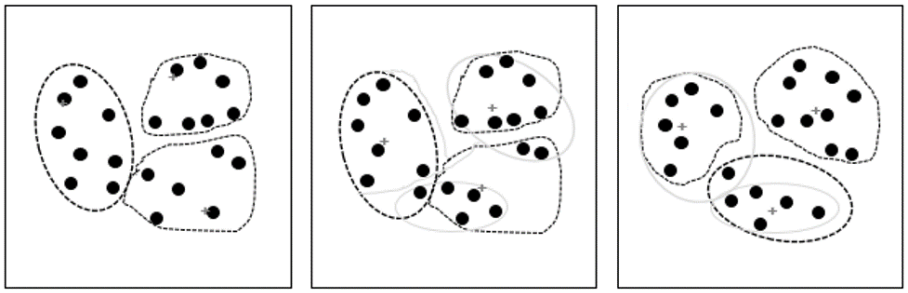
Algoritmo 1 Algoritmo *k-means*

1: **procedure** AGRUPAR()
2: escolha arbitrariamente k objetos de D como centroides iniciais dos grupos
3: **repita**
4: reatribua cada objeto ao grupo ao qual o objeto é o mais similar, com base no valor da média dos objetos no grupo
5: atualize a média do grupo, ou seja, calcule o valor da média dos objetos para cada grupo
6: **até que** não haja mudanças

Fonte: Adaptado de Han, Kamber e Pei (2011).

Algoritmos que implementam a tarefa de agrupamento com base em centroide usam o centroide de um grupo para representar este grupo. Conceitualmente, o centroide de um grupo é o ponto central e pode ser calculado por meio da média dos objetos (ou pontos) atribuídos ao grupo, conforme ilustra a figura 1, que parte da formação inicial dos grupos (a), passa pelo processo de iteração (b) até chegar no agrupamento final (c).

Figura 1 – Agrupamento utilizando o algoritmo *k-means*



Fonte: Adaptado de Han, Kamber e Pei (2011).

A execução de tarefas de agrupamento de dados pode resultar em grupos mal definidos ou ainda atribuir um objeto a um grupo que não é o ideal. Há formas de validar se os grupos estão bem formados e se os objetos neles contidos são de fato parecidos. A avaliação da tarefa de agrupamento é feita por meio de validações interna e externa.

A validação interna analisa se os dados estão apropriados em relação aos grupos gerados, considerando os objetos e suas características dentro de cada grupo, medindo a

coesão interna dos elementos dentro de um grupo e se os grupos estão bem separados um dos outros.

Um índice comummente estudado pela literatura para realizar a validação interna em uma tarefa de agrupamento é o *silhouette*, que segundo Rousseeuw (1987) apud Liu et al. (2010), valida o desempenho do agrupamento com base na diferença entre os pares e entre as distâncias dentro do grupo. Além disso, o número ideal de grupos é determinado pela maximização do valor deste índice. A fórmula do índice *silhouette* (*ISIL*) está ilustrada na equação 1,

$$ISIL = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right))} \tag{1}$$

em que

a(i) é a distância média do dado i a todos os demais dados do seu grupo e

b(i) é a distância mínima do dado i a todos os demais dados que não pertencem ao seu grupo. A análise do resultado indica que quanto maior o valor do índice melhor.

Já a validação externa é utilizada para analisar o agrupamento comparando os grupos gerados pelo algoritmo que implementou o agrupamento, com uma partição conhecida do conjunto de dados de origem. Os dados de origem podem ser obtidos por meio de classes previamente conhecidas deste conjunto ou ainda pelo que se espera em relação aos grupos que serão gerados.

A validação externa sobre um agrupamento utiliza métricas de contagem baseada em pares, como *RAND* (*IR*), proposto por Rand (1971) e *RAND* ajustado. Hubert e Arabie (1985) apud Vinh, Epps e Bailey (2009) definem que este tipo de medida é construída com a contagem de pares de itens em que dois agrupamentos concordam ou discordam. Considere: *Cob* um conjunto de grupos obtidos por meio do algoritmo de agrupamento; e *Cref* um conjunto de grupos de referência, conhecidos a priori. Então (*IR*) é definido por:

$$IR = \frac{(A + D)}{(A + B + C + D)} \tag{2}$$

em que

- A indica a quantidade de pares de exemplares que pertencem a um mesmo grupo *Cob* e uma mesma partição *Cref*;
- B indica a quantidade de pares de exemplares que pertencem a um mesmo grupo *Cob* e partições *Cref* diferentes;
- C indica a quantidade de pares de exemplares que pertencem a grupos *Cob* diferentes e a mesma partição *Cref*;

- D indica a quantidade de pares de exemplares que pertencem a grupos *Cob* diferentes e a partições *Cref* diferentes.

Com base no IR, o cálculo do índice $RAND$ ajustado (IRA) é dado por:

$$IRA = \frac{IR(Cob, Cref) - Esp(IR(Cob, Cref))}{max(IR) - Esp(IR(Cob, Cref))} \tag{3}$$

em que

- *Esp(IR(Cob, Cref))* indica o valor esperado do IR ao comparar as partições *Cob* e *Cref*;
- *max(IR)* indica o valor máximo atingido por essa medida (ou seja, $max(IR) = 1$).

Bobadilla et al. (2013) sugere que a tarefa de agrupamento (mineração de dados) seja acoplada à uma abordagem híbrida para resolver problemas como *cold start*. Já Liao e Lee (2016) propõe uma abordagem com base em tarefas de agrupamento para definição de grupos similares usados na composição das recomendações.

A estratégia definida para uso da tarefa de agrupamento foi estabelecer o valor ideal de k, executando o algoritmo de agrupamento e avaliando os índices interno e externo, e a partir desta análise selecionar os grupos gerados pelo algoritmo, carregar a relação dos grupos e objetos neles contidos (associados à similaridade entre eles). Dessa forma, o algoritmo de recomendação estabelece dois critérios para recomendação baseado em agrupamento: pela maior similaridade intragrupo e maior similaridade intergrupo.

Métodos para recomendação: recomendação e raciocínio baseado em casos

A recomendação baseada em casos teve sua origem a partir dos estudos sobre o raciocínio baseado em casos, principalmente por contribuições associadas à etapa de recuperação da informação e ao processo de avaliação de similaridade, sendo este último, um fator chave para estes tipos de métodos (SMYTH, 2007).

Raciocínio baseado em casos não é apenas uma técnica ou tipo de sistema de recomendação, Lenz et al. (1998) define como sendo uma metodologia capaz de "lembrar de casos úteis", em que algo pesquisado faz lembrar, ou associar alguma informação memorizada. O raciocínio baseado em casos é considerado também um método de resolução de problemas, em que, dado um problema, há que se encontrar uma solução adequada.

Chegar em uma solução final de um problema pode depender de uma série de decisões. Um “caso” considerado novo é dado pela descrição de um objeto (problema) cujo objetivo é descobrir a classe correta (solução) desse objeto. Neste formato, um “caso” é representado como um par ordenado (problema, solução).

Os sistemas baseados em casos dependem de um repositório de dados chamado de base de casos, onde são armazenadas as resoluções de problemas baseadas em experiências passadas. Mas além das soluções armazenadas para problemas já conhecidos, novos problemas podem surgir, sendo resolvidos com a recuperação de um “caso” similar ao problema que se pretende resolver, desta forma, a solução é adaptada para essa situação.

Raciocínio baseado em casos não provê somente soluções para problemas, mas também se preocupa com outras atividades durante seu processo, conforme ciclo ilustrado na figura 2, traduzido de Aamodt e Plaza (1994).

Figura 2 – Ciclo do raciocínio baseado em casos

Fonte: Adaptado de Aamodt e Plaza (1994)

No ciclo do raciocínio baseado em casos é descrito o que ocorre com o “caso” em cada etapa:

1. Recupera - a partir do problema inicial, recupera os “casos” mais similares, que podem ser “casos” já identificados no passado ou um novo “caso” descoberto.

2. Reusa - reutiliza a informação e o conhecimento que os “casos” recuperados no passo anterior podem fornecer para resolver o problema.
3. Revisa - revisa a solução proposta no passo anterior, confrontando-a novamente com a base de casos. Os “casos” são testados e reparados (solução confirmada).
4. Retém - armazena partes da experiência que podem ser úteis para soluções de problemas futuros (lições aprendidas).

As características do método de raciocínio baseado em casos se apresentaram adequadamente aplicáveis ao problema de recomendação de notícias, por este motivo, foram objeto de estudo por este trabalho e serviram de referência para a modelagem da base de casos e para implementação de estratégias de recomendação. Foram adaptadas e implementadas as seguintes etapas do ciclo do raciocínio baseado em casos: Problema (Novo Caso); 1. Recupera; 2. Reusa; 4. Retém; e Casos anteriores (Conhecimento Geral). O processo 3. Revisa não foi aplicado nesta pesquisa, podendo ser objeto de estudo em trabalhos futuros.

Como uma variação do raciocínio baseado em casos surge a recomendação baseada em casos que segundo Smyth (2007) é uma forma particular de recomendação baseada em conteúdo, em que os itens são descritos em termos de características bem definidas, por exemplo na recomendação de produtos caracterizados pelos atributos cor e marca. Este tipo de representação permite ao recomendador fazer julgamentos a respeito da semelhança entre os itens e ainda considera as descrições dos itens, gerando recomendações de itens similares àqueles que outros usuários gostaram no passado, sem considerar as preferências por parte do usuário.

Sistemas que implementam recomendadores baseados em casos, utilizam representações mais estruturadas do item e não somente textos livres. Esta característica sugere que estes recomendadores podem ser particularmente adaptados para domínios onde as descrições dos itens estão presentes em termos de atributo-valores (SMYTH, 2007).

O método de *RBC* também se mostrou adequado para resolver problemas relacionados à recomendação de notícias, já que é capaz de trabalhar com atributos estruturados e não estruturados, podendo utilizar a similaridade de texto para fazer sugestões de notícias (recomendação baseada em conteúdo), mas também é capaz de utilizar características estruturadas da notícia, como o canal de leitura (recomendação baseada em conhecimento).

Por ser um método derivado do raciocínio baseado em casos, a *RBC* utiliza a recomendação a partir de uma solução conhecida (fornecida no passado) e também por meio de uma solução adaptada. Desta forma, o par ordenado (problema, solução), conceituado pelo método de raciocínio baseado em casos, também pode ser mapeado para a dupla (notícia "problema", notícia "solução"), ou seja, dada uma notícia que está sendo lida (problema), o método encontra outra notícia a ser recomendada (solução).

Avaliação de um sistema de recomendação

Determinar qual a recomendação mais relevante em um sistema de recomendação é uma tarefa difícil e encontrar uma lista de recomendações com base em diferentes métricas, torna o trabalho ainda mais complexo. Apesar destes desafios, os *SR* precisam considerar as formas de avaliar as recomendações como instrumentos para melhor tomada de decisão.

Shani e Gunawardana (2011) listam as propriedades essenciais que servem de base para avaliação da recomendação: estudo da preferência do usuário; acurácia de colocação de item; cobertura; novidade; diversidade; utilidade e serendipidade.

- Estudo da preferência do usuário - uma forma de selecionar uma propriedade ou um algoritmo a ser implementado em um sistema de recomendação é utilizar um mecanismo que estuda o comportamento do usuário durante o uso do sistema, variando técnicas, propriedades e algoritmos, selecionando a opção com maior número de votos ao final do estudo.
- Ranqueamento do item - é responsável por predizer uma lista ordenada de itens que sejam de interesse do usuário. Esse tipo de propriedade pode ser obtida por meio de um ranqueamento de referência, que é quando o sistema possui registradas notas de avaliações do item. Uma forma de apresentar estes itens é listando o resultado em ordem decrescente de nota (por exemplo 5 estrelas, 4 estrelas, e assim sucessivamente).
- Cobertura - o termo cobertura é comummente usado para definir a proporção de itens que pode ser recomendada por um sistema de recomendação. Três tipos de cobertura são de interesse deste trabalho:

  - Cobertura do item no espaço - é um tipo de cobertura cuja medida é a porcentagem de todos os itens que são recomendados aos usuários durante um experimento.
  - Cobertura do usuário no espaço - a cobertura também pode ser medida em relação à proporção de usuários para o qual o sistema pode recomendar os itens.
  - *Cold start* - este termo refere-se ao desempenho do sistema em relação aos itens novos ou usuários novos e serve para medir a cobertura do sistema nas seguintes situações: por um determinado período de tempo; em subconjunto de itens; ou em um grupo específico de usuários. Por exemplo, um sistema pode sempre incluir um item "frio" dentro da lista de recomendação, mesmo que este não tenha indicadores de popularidade (obviamente ocorre em itens novos).

- Novidade - é a propriedade utilizada pelo sistema para recomendar itens que o usuário desconhece. Esta propriedade pode servir de forma indireta para que o sistema mensure o grau de exploração de sua base de itens. Os *SR* podem usar esta propriedade para gerar um controle itens já visitados pelo usuário. Este último requisito deve ser usado com cautela, pois em alguns casos a recomendação de itens que o usuário já visitou pode causar uma insatisfação dele em relação ao sistema, por outro lado, nunca mais recomendar um item, pode resultar em uma limitação sistêmica não desejada pelo usuário.
- Diversidade - é uma propriedade normalmente entendida como o inverso da similaridade, já que envolve a sugestão de itens diversificados afim de apresentar ao usuário a variedade de itens do conjunto de dados. Se um sistema recomenda somente itens similares está incorrendo no problema de previsibilidade.
- Utilidade - a utilidade de um sistema de recomendação é normalmente medida em função dos resultados alcançados por influência da recomendação, por exemplo, em um sistema de *e-commerce* quando um novo algoritmo de recomendação é implantado, a utilidade pode ser medida pelo aumento das vendas dos produtos após a operação do novo algoritmo, em comparação com os resultados anteriores.
- Serendipidade - serve para medir o quão surpreendente são as recomendações de sucesso. Uma outra forma de identificar se uma recomendação é serendípida é questionar ao usuário se o item que está sendo recomendado causou nele uma surpresa positiva. Há estratégias que visam alançar a serendipidade, como: diversidade,

similaridade de usuário, novidade e aleatoriedade, embora neste último caso uma boa alternativa é equilibrar o acaso com precisão.

- Privacidade - em *SR* que utilizam dados obtidos a partir de interações dos usuários, a questão da privacidade deve ser tratada de forma bem transparente entre as partes, já que os usuários colaboram voluntariamente com o sistema na expectativa de obterem melhores recomendações para si, mas não esperam que suas preferências sejam divulgadas publicamente. Dessa forma, um sistema de recomendação deve ter a preocupação de manter sigilo sobre os dados cadastrais de preferências dos usuários e deve também questionar ao usuário se as interações que ele proporciona podem ser usadas para gerar predições (juntamente com as de outros usuários).

As propriedades supracitadas podem ser avaliadas com base nos aceites de recomendação registrados pelos usuários do sistema de recomendação, comparando as medições em momentos distintos de uso do sistema ou ainda com outros critérios de recomendação.

Adicionalmente à propriedades, este trabalho considera avaliar as recomendações com base em: similaridade de conteúdo; similaridade de usuário; similaridade intragrupo e intergrupo; e por meio de uma métrica unificada para ranqueamento;

Avaliação com base em similaridade de conteúdo

Os *SR* podem também ser avaliados considerando sua capacidade de recomendar itens similares. A similaridade de conteúdo é usada para recomendação de maneira direta, ou seja, se um leitor acessou uma determinada notícia, o sistema recomenda uma outra que tenha um conteúdo similar.

Há diversas métricas que podem ser utilizadas para calcular similaridade entre itens. Algumas são mais adequadas para resolver problemas de recomendação baseada em conteúdo, já que para se obter a similaridade entre documentos texto é necessário o uso de estruturas como vetores de dados numéricos, que normalmente são esparsos (como vetores de termo e frequência). Para problemas desta natureza, a medida cosseno é fortemente recomendada para cálculo de similaridade (HAN; KAMBER; PEI, 2011).

Sobre a medida cosseno, Han, Kamber e Pei (2011) explicam que um documento pode ser representado por milhares de atributos, em que cada um destes atributos pode ser uma palavra específica (como uma palavra-chave) ou ainda uma frase no documento. Assim,

cada documento é um objeto de um vetor de termo e frequência (que são tipicamente longos e esparsos, ou seja, têm muitos valores de zero). Há medidas de distância que não funcionam bem com dados numéricos esparsos. Por exemplo, dois vetores de termo e frequência podem ter muitos valores de zero em comum, o que significa que os documentos correspondentes não compartilham muitas palavras e que isso não os torna semelhantes. Mas nem sempre isso procede, para isso é preciso de uma medida que se concentre nas palavras que os dois documentos têm em comum e na frequência de ocorrência de tais palavras, ou seja, é necessária uma medida para dados numéricos que ignore os valores iguais a zero. Neste contexto, a medida de similaridade cosseno atende adequadamente a problemas desta natureza, sendo útil na comparação de documentos do tipo texto (ou vetores de palavras).

Rezende, Marcacini e Moura (2011) apud Tan, Steinbach e Kumar (2005) descrevem que a medida de similaridade cosseno é definida de acordo com ângulo cosseno formado entre os vetores de dois documentos. Assim, se o valor da medida de similaridade cosseno é zero, o ângulo entre os vetores é noventa graus, ou seja, os documentos não compartilham nenhum termo. Por outro lado, se o valor da similaridade for próximo do número um, o ângulo entre os vetores é próximo de zero, indicando que os documentos compartilham termos e portanto são similares.

Avaliação com base em similaridade entre usuários

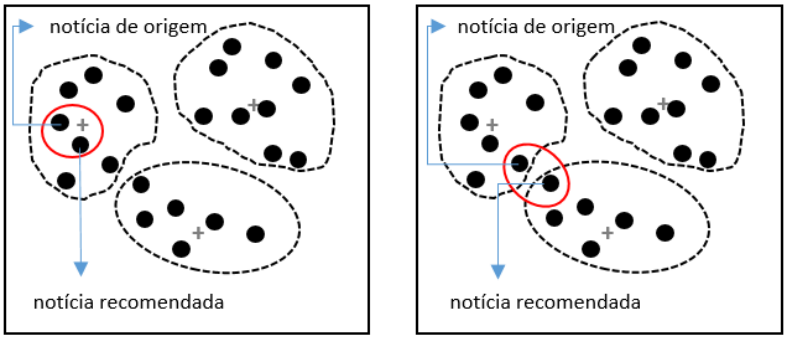
O modelo de recomendação usuário-usuário usa a similaridade entre usuários considerando o que eles têm de interesse em comum. Koren e Bell (2015) descrevem que alguns recomendadores lidam com itens que mudam rapidamente, tornando as relações item-item muito voláteis. Por outro lado, uma base de usuários estável estabelece um relacionamento de longo prazo entre os usuários. Recomendadores de artigos da *web* ou notícias mudam rapidamente pela própria natureza dos itens, mas podem possuir uma relação estável com os usuários. Nesses casos, as recomendações baseadas em modelo usuário-usuário são mais atraentes.

Este modelo faz parte da estratégia de recomendação baseada em filtro colaborativo e foi implementado no protótipo de sistema de recomendação utilizado no experimento conduzido por esta pesquisa.

Avaliação com base em similaridade intragrupo e intergrupos

Os itens que compõem um grupo gerado a partir de uma tarefa de agrupamento possuem maior similaridade com outros itens que fazem parte do mesmo grupo (na maioria dos casos). Com base nesta proximidade entre itens é possível fazer recomendações com base em similaridade intragrupo. Por outro lado, considere um cenário em que o item 1 pode estar distante do centroide do seu grupo, por este motivo, pode estar mais próximo do item 2 que também pode estar distante do centroide do grupo dele. Nestes casos, a recomendação do item 2 se dá por meio de uma similaridade intragrupo em relação ao item 1. Na figura 3 estão ilustradas estes tipos de similaridade e como o recomendador pode utilizá-las para formulação da estratégia de recomendação.

Figura 3 – Similaridade intragrupo e intergrupos



Fonte: José Luiz Maturana Pagnossim, 2018

Considerações sobre avaliações com base em similaridade

Uma forma de avaliar recomendações baseadas em similaridade de conteúdo, entre usuários e por agrupamento (intragrupo e extra grupos) é medir os aceites de recomendação registrados pelos usuários do sistema de recomendação, comparando as medições em momentos distintos de uso do sistema. Outra forma é confrontar o resultados dessas avaliações de acordo com os demais critérios para recuperação da notícia.

Avaliação com base em popularidade

Indicadores de popularidade são medidos por meio de interações entre os usuários e os *SR* e são medidos de acordo com o interesse do usuário pelo item.

Tatar et al. (2014) descreve que métricas de popularidade são baseadas em atividades de participação do usuário, como: comentários; votos; e compartilhamentos por meio de redes sociais ou serviços de e-mail. O autor defende que, no contexto de notícias, um bom indicador para mediar a popularidade é por meio da quantidade de comentários que os usuários publicam para cada notícia.

A figura 4 ilustra outros exemplos das formas possíveis de avaliação da popularidade de uma notícia (curtida, serendipidade e nota de avaliação em escala 1-5 estrelas). O conteúdo da notícia apresentada nesta figura foi obtida do portal EBC[7].

Figura 4 – Indicadores de popularidade: um exemplo

Fonte: José Luiz Maturana Pagnossim, 2018

Avaliação com base no ranqueamento da recomendação

O ranqueamento de uma recomendação pode ser feito de várias formas, por exemplo por meio de indicadores de popularidade como aceites, curtidas e compartilhamentos.

[7] http://radioagencianacional.ebc.com.br/pesquisa-e-inovacao/audio/2017-11/universo-descoberta-de-planetas-parecidos-com-terra-desafia

Cada método de ranqueamento pode estimar a relevância de um item usando um determinado critério, e um método é considerado adequado se a lista ordenada (estimada) estiver próxima da lista ideal (ou de uma lista de referência) (TATAR et al., 2014).

Tatar et al. (2014) explica que se um jornal *online* tem de um lado uma notícia que ele pretende publicar e do outro um conteúdo que se pretende promover, este tem que ter mecanismos para ordenar automaticamente qual destas publicações é mais importante. A partir desta tomada de decisão, o sistema deve exibir ao leitor: primeiro a notícia melhor classificada e depois a outra. O autor defende como forma de avaliação a comparação entre diferentes métodos de ranqueamento por meio de um experimento com usuários, de forma que se possa medir a eficácia destes métodos.

A figura 5 ilustra um exemplo com três métodos de ranqueamento utilizando ordenações diferentes para recomendar os mesmos itens.

Figura 5 – Avaliação de métodos de ranqueamento: um exemplo

Ranqueamento	Método de ranqueamento 1		Método de ranqueamento 2		Método de ranqueamento 3	
	Item	# Curtidas	Item	# Curtidas	Item	# Curtidas
1	Item 1	300	Item 2	200	Item 3	100
2	Item 2	200	Item 1	300	Item 2	200
3	Item 3	100	Item 3	100	Item 1	300

Fonte: José Luiz Maturana Pagnossim, 2018

Neste exemplo é possível observar que o método de ranqueamento 1 obteve melhor eficácia em relação aos outros métodos, pois conseguiu estimar de forma exata a ordem de relevância dos itens dentro da lista (de acordo com o resultado do indicador de popularidade "#Curtidas". Por outro lado, nota-se a ineficácia do método de ranqueamento 3, que estimou o oposto do ideal.

Avaliação das estratégias de recomendação

SR construídos a partir de abordagens híbridas demandam a aplicação de diferentes estratégias para gerar a recomendação. Estas estratégias envolvem adequar a recomendação dependendo do contexto em que o sistema, os itens e os usuários estão inseridos. Por exemplo, um sistema pode recomendar itens com base em similaridade de conteúdo durante um período de *cold start* de item. Estratégias de recomendação com base na recuperação ou adaptação do "caso", dentro do método proposto pela *RBC*, também são opções que podem ser usadas na recomendação e posteriormente avaliadas. Algumas avaliações possíveis são:

verificar os aceites de recomendação em relação à estratégia de recomendação; analisar a distribuição das recomendações emitidas pelo sistema de acordo com as estratégias utilizadas. Estas avaliações permitem a evolução do sistema de recomendação por meio da calibragem adequada de uso da estratégias e refinamento os algoritmos de recomendação.

2.2 Sistemas híbridos

Sistemas híbridos baseiam-se na combinação de dois ou mais tipos de sistemas de recomendação (RICCI et al., 2011). Os sistemas híbridos tendem a gerar melhores resultados em relação ao uso de técnicas isoladas, pois as técnicas são combinadas de forma que uma possa compensar a limitação da outra.

Um problema conhecido da recomendação baseada em filtro colaborativo está relacionado aos novos itens (*cold start* de item), ou seja, há uma dificuldade em recomendar itens que não têm avaliações registradas. Este aspecto não é limitante em abordagens baseadas em conteúdo, uma vez que a previsão para novos itens é baseada em uma descrição que normalmente está disponível. O caráter híbrido é aplicado em diferentes frentes:

- Uso de diferentes tipos de recomendação: baseada em conteúdo; baseada em conhecimento; e baseada em filtro colaborativo.
- Uso de diferentes estratégias de recomendação: baseada em critérios de recuperação da notícia; baseada em casos (conhecidos e adaptados).
- Uso de procedimentos de pré-processamento de texto e da tarefa de agrupamento.
- Cálculo de uma métrica unificada para ranqueamento da recomendação, considerando: popularidade; similaridade do conteúdo da notícia; similaridade dos grupos de notícias; similaridade entre usuários; histórico de navegação do usuário; aceite de recomendação; nota média da recomendação; aleatoriedade; e serendipidade.

2.3 Pesquisas correlatas

A área de *SR* é vasta e possibilita a aplicação de diferentes técnicas e métodos para geração da recomendação. Também são utilizadas diferentes estratégias de implementação da recomendação com distintas formas de avaliação. As áreas de aplicação também são as mais diversas, sendo o tratamento dos dados uma tarefa particular para cada domínio.

Esta seção apresenta uma busca exploratória da literatura com objetivo de criar um mapa dos trabalhos correlatos às ideias discutidas nesta pesquisa.

A busca exploratória realizada não pretende alcançar o mesmo nível de detalhe de uma revisão sistemática da literatura, embora tenha contado com características deste tipo de revisão, descritas por Kitchenham (2004) em seu relatório técnico que descreve os procedimentos para execução de uma revisão sistemática. Este relatório técnico descreve também os estágios durante a condução de uma revisão sistemática: identificação da pesquisa; seleção dos estudos primários; avaliação da qualidade do estudo; extração e monitoramento dos dados; e síntese dos dados.

O método utilizado para realização desta etapa da pesquisa foi organizado da seguinte forma:

- A identificação da pesquisa - definição da área, tema e palavras chaves.
- Seleção de estudos primários - por meio da elaboração de *strings* de busca (6 *strings*). Foi utilizado um critério de pelo menos uma citação.
- Escolha da fonte de dados - foi utilizado o Scopus[8] que é um dos maiores indexadores de bases de dados e comumente usada para realização de buscas para revisões sistemáticas da literatura.
- Escolha da janela de tempo a ser explorada - busca por publicações a partir do ano de 2010. Como resultado das tomadas de decisão supracitadas, foram obtidos 266 publicações científicas.
- Avaliação da qualidade do estudo - leitura da publicação para análise da aderência ao tema e relevância. Após esta etapa foram selecionadas 14 publicações, que são comentadas nesta seção.
- Síntese dos dados - resumo das publicações selecionadas e agrupamento por assunto.

2.3.1 Descrição dos trabalhos selecionados

Após os estágios supracitados foi possível estabelecer um agrupamento das publicações por assunto, que estão organizadas de acordo com: as abordagens para recomendação e soluções híbridas; problemas e desafios dos *SR*; e domínio dos dados. Na tabela 1, é apresentado um resumo das publicações selecionadas.

[8] https://www.scopus.com/

Tabela 1 – Tabela resumo dos trabalhos correlatos

Id	Grupo	Título da publicação	Referência
1	Abordagens híbridas	Recommender systems survey	Bobadilla et al. (2013)
2	Abordagens híbridas	Classifications of Recommender Systems...	Saquib et al. (2017)
3	Abordagens híbridas	A hybrid multi-criteria recommender system...	Kermany et al. (2017)
4	Abordagens híbridas	A web-based personalized business partner...	Lu et al. (2012)
5	Abordagens híbridas	Review of Web Personalization	Malik et al. (2012)
6	Abordagens híbridas	Further Experiments in Opinionated Product...	Dong et al. (2014)
7	Abordagens híbridas	Automatically Recommending Multimedia...	Bermingham et al. (2015)
8	Problemas e desafios	A clustering based approach to improving...	Liao et al. (2016)
9	Problemas e desafios	Dealing with the new user cold-start problem...	Son et al. (2016)
10	Problemas e desafios	Incorporating popularity in a personalized...	Jonnalagedda et al. (2016)
11	Problemas e desafios	Diversity in recommender systems – A survey	Kunaver et al. (2017)
12	Domínio de dados	RCV1: A new benchmark collection for text...	Lewis et al. (2004)
13	Domínio de dados	A hybrid recommendation system for news...	Kunaver et al. (2016)
14	Domínio de dados	From popularity prediction to ranking online...	Tatar et al. (2014)

Fonte: José Luiz Maturana Pagnossim, 2018

Abordagens para recomendação e soluções híbridas

Este grupo de trabalhos relacionados levanta aspectos voltados às abordagens adotadas para recomendação. A maior parte dos trabalhos utiliza uma solução híbrida que mescla mais de uma técnica, método ou estratégia para a recomendação ou avaliação da mesma.

Bobadilla et al. (2013) em um trabalho de revisão sobre os sistemas de recomendação levanta os principais tipos de abordagens usadas para recomendação: baseada em conteúdo; baseada em informações demográficas; baseada em filtro colaborativo; baseado em memória; e baseado em modelos de classificação. O autor sugere que soluções híbridas que combinem estas abordagens podem melhorar a qualidade na recomendação. Sugere ainda que a tarefa de agrupamento (mineração de dados) pode ser usada acoplada à abordagem híbrida para resolver problemas como *cold start*.

Saquib, Siddiqui e Ali (2017) em um trabalho mais recente de revisão sobre os tipos de sistemas de recomendação, descreveu suas características e aplicações, com destaque aos *SR* baseados em: filtro colaborativo; métodos recursivos; demográfico; conhecimento; contexto; redes sociais; técnicas *fuzzy* e genética; casos; limitação (ou restrição); e híbridos (que apresentavam diferentes combinações de tipos de *SR*). Por meio dos resultados foi possível concluir que uma técnica isolada não pode superar uma abordagem híbrida

Uma outra abordagem híbrida proposta por Kermany e Alizadeh (2017) apresentou melhores resultados em termos de acurácia de predição e eficiência na redução da esparsidade. Esta abordagem foi aplicada para a recomendação de filmes e mesclou conceitos de *fuzzy* multi-critério; *SR* baseado em filtro colaborativo; *SR* baseado em informações demográficas; e uma ontologia baseada em item.

Com o crescimento do o *e-commerce*, as empresas que atuam neste setor buscam cada vez mais soluções que forneçam recomendações personalizadas para seus usuários e clientes em potencial. Neste cenário, Lu et al. (2012) definem que os sistemas de recomendação são ferramentas efetivas para implementação de serviços personalizados em ambiente *web* e propõem uma abordagem híbrida que envolve semântica *fuzzy* e recomendação baseada em filtro colaborativo. Os resultados mostraram que a abordagem proposta supera as limitações do filtro colaborativo quando este é usado isoladamente.

Melhorar a satisfação de usuários em um ambiente *web* é uma meta difícil de ser alcançada. Para alcançar resultados eficazes neste ambiente Malik e Fyfe (2012) sugere a utilização de recomendações baseadas em conteúdo, conhecimento e filtro colaborativo. Os resultados demonstram que por meio desta abordagem híbrida os usuários são beneficiados com recomendações que provêm exatamente o que eles necessitam.

Outras estratégias mesclam algumas das abordagens supracitadas com métodos de raciocínio baseado em casos, como é a pesquisa de Dong, O'Mahony e Smyth (2014) que obtém dados a partir de descrições geradas por comentários de usuários para gerar descrições estruturadas dos produtos, o que proporciona condições de utilizar uma abordagem de recomendação baseada em casos e ainda tirar conclusões a respeito do sentimento do usuário. O autor demonstra que os resultados da recomendação baseada em casos ficaram dentro do esperado e com potencial de melhoria, já a análise de sentimento não apresentou ganhos.

A combinação de recomendação baseada em conteúdo e raciocínio baseado em casos está presente na pesquisa de Bermingham et al. (2015) que apresenta uma solução híbrida (incluindo as duas abordagens citadas) para recomendação de tarefas terapêuticas para pessoas com Alzheimer. O autor mostra que a solução foi efetiva na identificação exata de tarefas para sessões de terapia para pacientes com Alzheimer.

Problemas e desafios dos *SR*

Um problema conhecido na área de *SR* que impacta principalmente recomendadores baseados em filtro colaborativo é conhecido como *cold start*. Para resolver este tipo de problema, estudiosos da área de *SR* propõem variadas soluções. Liao e Lee (2016) apresentam uma abordagem com base em tarefas de agrupamento para definição de grupos similares usados na composição das recomendações. Analisando os resultados foi possível concluir que esta abordagem melhorou a eficiência do recomendador sem perda de qualidade na recomendação e compensou o problema de *cold start* (que limitava o uso de recomendação baseada em filtro colaborativo).

Um outro trabalho que discorre sobre problemas causados pelo efeito do *cold start* é apresentado por Son (2016) que compensa a limitação da recomendação baseada por filtro colaborativo por meio da recomendação baseada em conteúdo.

Paralelamente aos problemas enfrentados pelos *SR* há trabalhos que visam estudar desafios típicos destes tipos de sistema, como a popularidade e a diversidade. Jonnalagedda et al. (2016) apresentam uma abordagem de recomendação de notícias que utiliza dois principais critérios pra gerar a recomendação: a popularidade e o perfil do usuário. O autor compara os resultados deste recomendador com um recomendador pessoal que dispunha de dados previamente gerados (*baseline*). Os resultados apontaram piores resultados da abordagem proposta em relação ao *baseline* quando usada somente a popularidade como propriedade para compor a recomendação, já quando a recomendação foi combinada (popularidade com o perfil do usuário), a abordagem proposta superou o recomendador *baseline*.

Outro tema que vêm sendo cada vez mais discutido na área de *SR* é a recomendação com base em diversidade. Kunaver e Požrl (2017) apresenta uma revisão da literatura com ênfase na diversidade. Em seu trabalho é relatado o aspecto subjetivo do conceito de diversidade para o usuário, por exemplo, algumas pessoas podem considerar que a recomendação do filme *Star Trek* para quem assistiu o filme *Star Wars* é baseada em similaridade, enquanto alguns usuários mais especializados nestes gêneros de filmes podem considerar que são filmes muito diferentes e que neste sentido, a recomendação estaria usando a diversidade ao invés da similaridade. O autor sugere que a diversidade seja tratada nas atividades que antecedem a composição da recomendação e não avaliada somente

na fase que sucede a recomendação (como faz a maioria dos algoritmos pesquisados pelo autor). Dessa forma, a lista de recomendação deve ser montada considerando diferentes critérios, evitando recomendações com itens muito parecidos.

Domínio dos dados

Os procedimentos e algoritmos utilizados para mineração de texto dependem do domínio de dados, principalmente no que se refere ao idioma do texto e do quão são estruturáveis os dados, a fim de se utilizar meta-dados como como fonte para organização de informações. Na área de *SR* não é diferente, em que as dependências iniciam pela definição do formato do dado, por exemplo: texto; vídeo; música; e imagens, mas também podem estar associadas aos meta-dados que serão selecionados para compor as recomendações. Por fim, dependem também da área de aplicação, pois as formas de interação com o usuário e as estratégias de recomendação mudam dependendo do objetivo de negócio em que o recomendador é utilizado.

Neste contexto foram feitas buscas sobre o domínio de notícias com objetivo de identificar se seria possível reutilizar um *corpus* público de notícias e também conhecer as técnicas, estratégias e formas de avaliação em *SR* de notícias.

Lewis et al. (2004) apresenta um *benchmark* construído sobre um *corpus* com mais de oitocentas mil notícias da agência *Reuters*. As principais contribuições deste trabalho são: as métricas de eficácia; as formas de avaliação; a extração de características; os algoritmos (*K-NN*[9] e *SVM*[10]); e a classificação de texto.

Buscou-se também por abordagens que utilizassem mecanismos de interação de usuários com um sistema de recomendação de notícias. Neste sentido, Kunaver e Požrl (2016) apresentam um sistema de recomendação de notícias em um cenário de mobilidade dos usuários, para isso, o autor utiliza recomendação baseada em conteúdo e baseada em informações demográficas. A abordagem proposta utiliza o perfil do usuário e considera também a geolocalização do usuário em relação ao seu dispositivo móvel. Para medir os resultados foi realizado um experimento (que contou com a participação de 35 voluntários) baseado no método de estudo do usuário. Por meio deste experimento os usuários reconheceram a qualidade do recomendador. O aplicativo móvel utilizado no experimento foi

[9] K-Nearest Neighbor.
[10] Support Vector Machines

integrado a um jornal (público) português. A tarefa de agrupamento foi implementada na plataforma *Carrot2* [11] que organiza de forma automática pequenas coleções de documentos. Foi utilizada medida cosseno para cálculo de similaridade entre as notícias. A avaliação foi realizada em duas sessões, a primeira comparou três estratégias: a abordagem de recomendação proposta; as notícias mais populares; e notícias selecionadas aleatoriamente. A segunda sessão avaliou abordagens baseadas no perfil do usuário e na geolocalização. A pesquisa demonstra melhores resultados a favor do algoritmo implementado pelo trabalho.

Um desafio encontrado em recomendadores de notícias está relacionado ao ranqueamento dos itens que serão apresentados na lista de recomendação. É um consenso entre estudiosos da área e entre os usuários de *SR* que se um item está posicionado no topo da lista, este tem mais chances de ser consumido. Neste contexto, Tatar et al. (2014) propõem um método de ranqueamento da recomendação baseado em predição de popularidade e compara os resultados com alguns *baseline* e também com algoritmos de ranqueamento. Os resultados indicaram que o modelo de predição proposto foi eficaz para resolver o problema de ranqueamento de notícias. O autor usou como método de validação o experimento *offline* e considerou que esta forma de avaliação foi uma das limitações do trabalho, sugerindo que em trabalhos futuros o experimento considere o uso de uma plataforma *web* que seja capaz de capturar a reação do usuário em relação ao resultado do ranqueamento.

2.3.2 Considerações sobre os trabalhos correlatos

Nota-se que alguns trabalhos encontram-se na intersecção entre os assuntos definidos para organizar a análise dos trabalhos correlatos, o que possibilitaria enquadrar alguns destes trabalhos em outro grupo e ainda assim o assunto estaria adequado. Por exemplo, o trabalho de Jonnalagedda et al. (2016) poderia ser enquadrado nos três grupos, já que propõe uma abordagem para recomendação de notícias que utiliza uma estratégia híbrida para melhorar recomendações com base em popularidade.

Analisando os trabalhos correlatos, observa-se a forte tendência de utilização de abordagens híbridas, mesclando as mais diversas técnicas. Os resultados têm demonstrado que independente da combinação utilizada, a solução híbrida apresenta indicadores superiores em relação à comparação de técnicas isoladas e valores de *baseline*.

[11] http://project.carrot2.org/index.html

No que diz respeito aos problemas enfrentados pelos *SR*, o *cold start* é apresentado como um problema típico, que normalmente é compensado por abordagens baseadas em conteúdo, conhecimento, personalização do usuário, ou ainda com o apoio de tarefas de mineração de dados como o agrupamento.

Outras propriedades amplamente discutidas em *SR* são a popularidade e a diversidade. Os estudos indicam que a popularidade continua sendo um fator chave e relevante para se utilizar como critério de recomendação, mas também sugerem que quando usado combinado com outros critérios são obtidos melhores resultados. O efeito desta combinação de critérios reflete na diversidade, que também serve de contra-ponto da similaridade. A diversidade também tem se mostrado eficaz quando aplicada em soluções híbridas, compensando limitações de recomendadores que usam estratégias isoladas.

A tendência de adoção de abordagem híbrida também se mostrou presente no que diz respeito ao domínio de notícias. Além desta constatação, outros aspectos foram verificados nos trabalhos relacionados, como: o uso de tarefas de classificação e agrupamento como mecanismos de apoio para montagem da recomendação; a aplicação de medidas de distância para análise de similaridade entre documentos texto (incluindo a media cosseno); a adoção de um método para ranqueamento dos itens dentro da lista de recomendação; e as formas de validação (como uso de experimento *offline* e estudo do usuário).

A busca por trabalhos correlatos corroborou com os objetivos e os problemas de pesquisa levantados por este trabalho, indicando que há problemas em aberto e desafios a serem superados, e que portanto merecem ser mais explorados.

Por meio da análise dos trabalhos relacionados, apesar de encontrar referências isoladas ou ainda que combinadas com relação ao uso de técnicas, métodos, formas de avaliação, tratamento do domínio de dados e estudo das propriedades, problemas e desafios em *SR*, não foi possível encontrar um trabalho específico que englobasse todos os aspectos discutidos por esta pesquisa.

3 Apresentação da abordagem

Apesar dos avanços ocorridos nos últimos anos na área de *SR*, alguns desafios ainda não estão totalmente superados e merecem ser mais explorados, como: o tratamento de *cold start* de item; a previsibilidade; e a baixa relevância nas recomendações. Para que problemas dessa natureza sejam adequadamente solucionados, soluções híbridas têm sido propostas como forma de maximizar a eficácia dos resultados e minimizar as limitações relacionadas ao uso isolado de técnicas de recomendação.

Para alcançar os objetivos traçados para esta pesquisa e validar a hipótese delineada, este capítulo apresenta uma abordagem híbrida para recomendação de notícias que combina as técnicas, métodos e formas de avaliação fundamentadas no capítulo 2. A abordagem foi materializada por meio da implementação de uma arquitetura para *SR* de notícias. Foi desenvolvido um ambiente para prova de conceito desta arquitetura, denominado protótipo do sistema de recomendação de notícias. Este protótipo serviu de ambiente para realização de um experimento *online* que proporcionou a extração de dados para análise dos resultados.

Este trabalho priorizou tratar em abrangência vários e diferentes conceitos ao invés de tratar alguns poucos em profundidade. Assim, foram aplicadas as formas básicas dos métodos, técnicas, algoritmos e propriedades de um sistema de recomendação. Neste sentido, estudos intensivos de variações de parâmetros e diferentes combinações dos conceitos não fizeram parte do escopo do estudo.

Com o objetivo de apresentar a abordagem híbrida para sistemas de recomendação de notícias, este capítulo está organizado da seguinte forma:

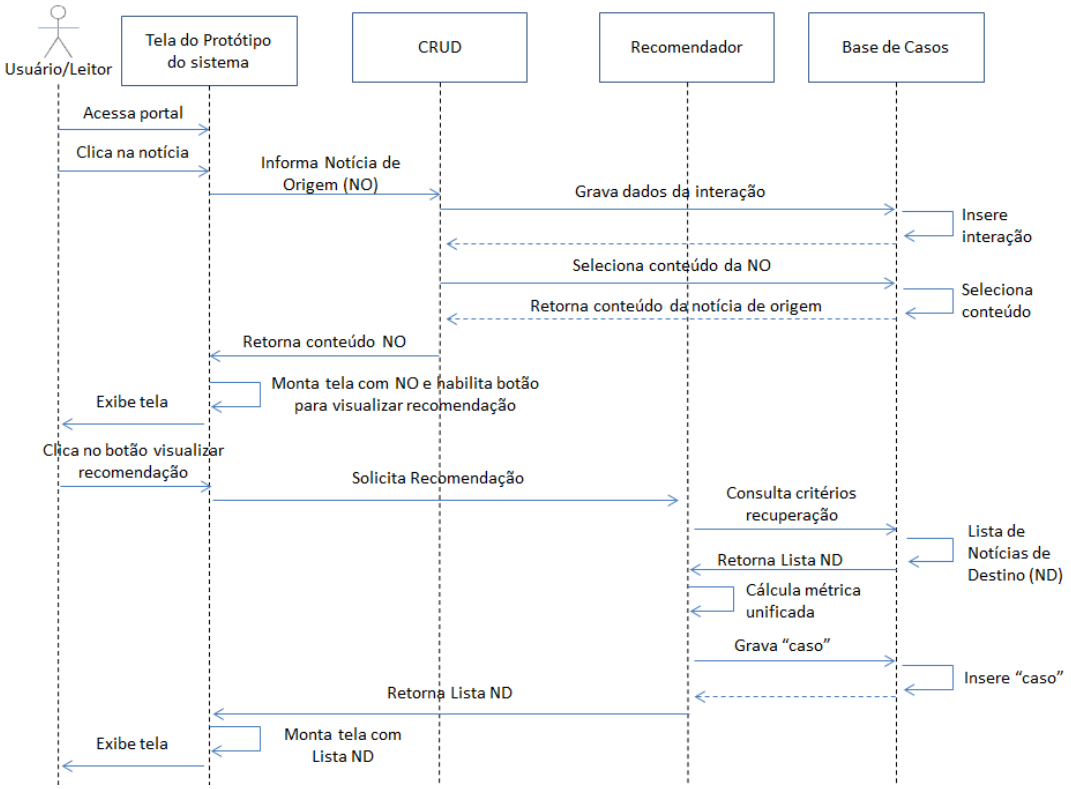
- Processo de recomendação - apresentação do processo de recomendação de notícias com foco na interação entre o usuário e o protótipo do sistema de recomendação.
- Arquitetura para *SR* de notícias - apresentação do desenho conceitual da arquitetura, descrição dos módulos, componentes e funcionalidades implementadas.

3.1 Processo de recomendação

O processo de recomendação envolve as ações do leitor ou usuário com o protótipo do sistema de recomendação de notícias. A interface gráfica deste protótipo foi acessada

pelo leitor ou usuário em um ambiente *web*. As consultas e o armazenamento dos dados foi intermediada por uma camada denominada *CRUD*[1]. A lógica de recomendação foi implementada na camada denominada Recomendador. Por fim, a modelagem da base de casos foi implementada em um modelo de banco de dados relacional denominado Base de Casos. Este processo de recomendação é apresentado em um diagrama de sequência UML[2], conforme ilustrado na figura 6.

Figura 6 – Processo de recomendação de notícias por meio de critérios de recuperação



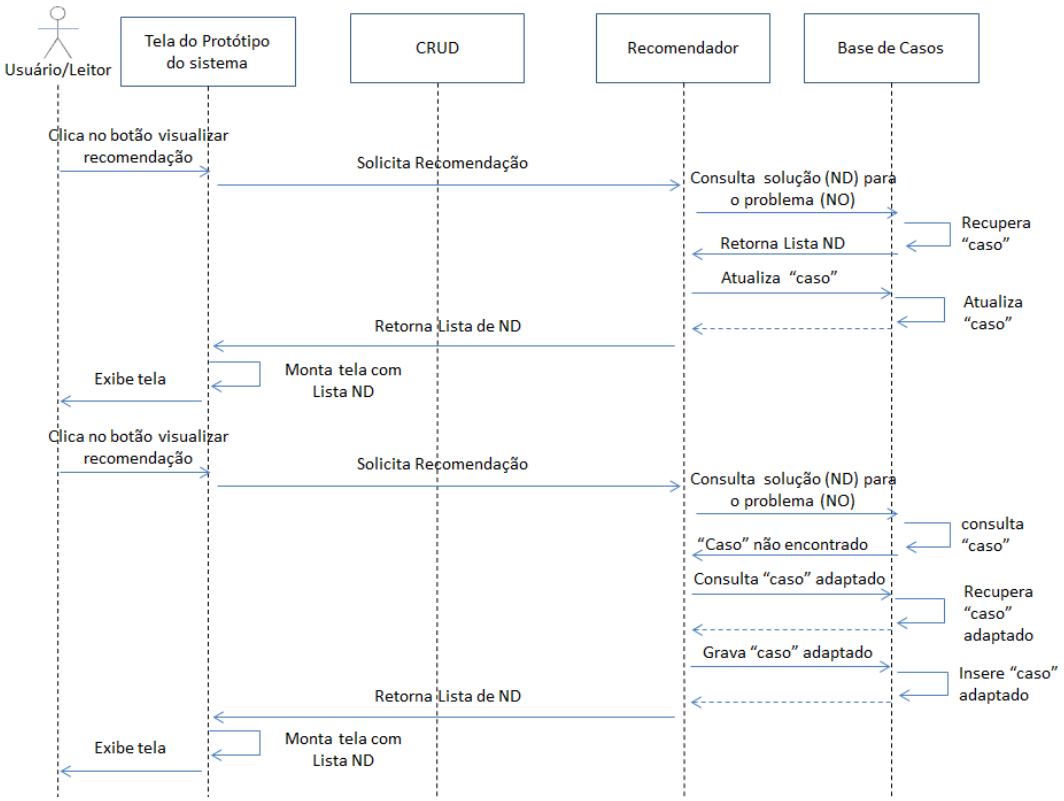
Fonte: José Luiz Maturana Pagnossim, 2018

Neste diagrama, é demonstrado o o processo de recomendação partindo de uma interação inicial de um leitor. Neste cenário em que o sistema não tem indicadores sobre os itens, os usuários, tão pouco os "casos" armazenados, o recomendador faz recomendações com base nos critérios de recuperação advindos das tarefas de mineração de dados e de um fator aleatório.

Ao registrar as recomendações e consequentemente associá-las ao modelo de raciocínio baseado em casos, o sistema reune condições de recuperar um "caso" conhecido (se encontrá-lo de forma direta) ou ainda adaptar um "caso" a ser recomendado. Estes cenários estão representados na figura 7.

[1] Acrônimo para *Create, Read, Update and Delete*
[2] Unified Modeling Language

Figura 7 – Processo de recomendação de notícias por meio da *RBC*



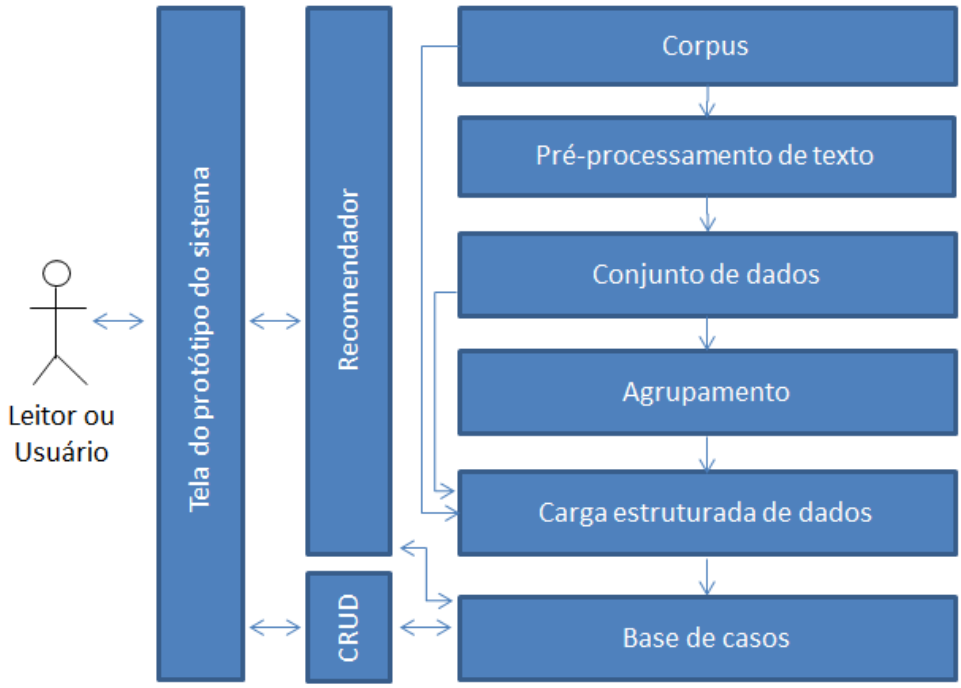
Fonte: José Luiz Maturana Pagnossim, 2018

3.2 Arquitetura para sistemas de recomendação de notícias

Na área de *SR*, um sistema pode ser considerado um produto já finalizado, que tenha passado por todo ciclo de vida da engenharia de software e que esteja homologado e implantado para utilização dos usuários finais. Já uma arquitetura entende-se ser uma organização que integra: camadas; módulos; processos; interfaces; princípios; diretrizes e padrões, servindo de modelo para a construção de um sistema. Assim, o trabalho em questão especifica e constrói as funcionalidades necessárias de uma arquitetura e as aplica em um protótipo de sistema de recomendação de notícias, de forma que seja possível avaliar a hipótese por meio de uma prova de conceito. A figura 8 apresenta o desenho conceitual da arquitetura.

Os desenhos, modelos, especificações e códigos-fontes que compõem esta arquitetura constituem como uma contribuição técnica deste trabalho.

Figura 8 – Arquitetura para sistemas de recomendação de notícias



Fonte: José Luiz Maturana Pagnossim, 2018

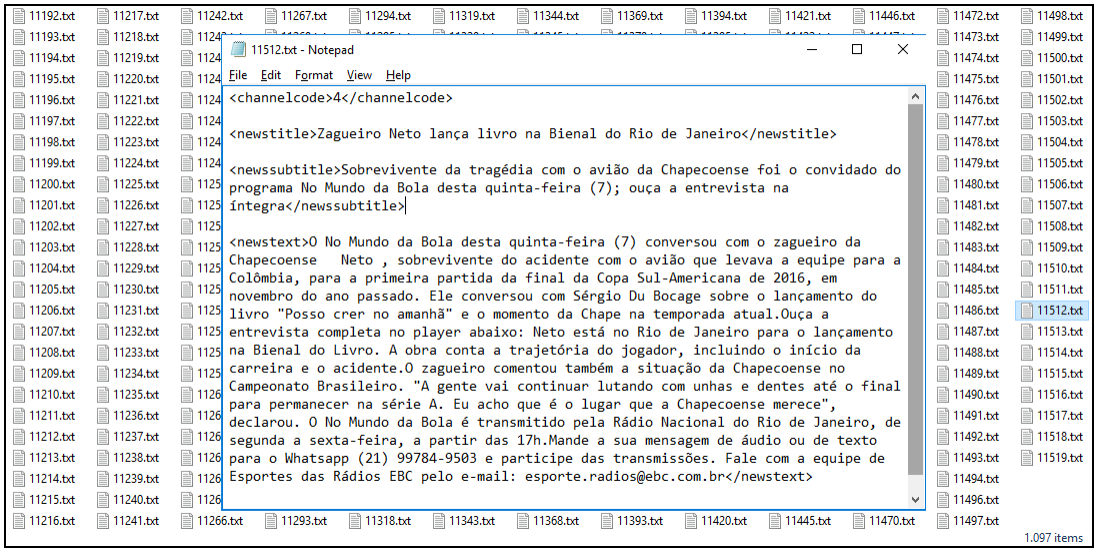
3.2.1 *Corpus*

Um *corpus* pode ser definido simplesmente como uma "coleção de textos". Se parecer muito amplo, uma qualificação possível de ser feita é: um *corpus* é uma coleção de textos quando considerado como um objeto de estudo linguístico ou literário (KILGARRIFF, 2001). Esta definição é feita pelo autor no contexto da *web* como sendo um grande *corpus*.

No contexto deste trabalho, a *web* foi utilizada para obtenção de notícias a partir do portal de notícias da Empresa Brasil de Comunicação (EBC). Foi feito um recorte com cerca de mil notícias obtidas no período de abril de 2017 até setembro de 2017. O conteúdo das notícias foi organizado em arquivos texto, possibilitando sua utilização pelos procedimentos de pré-processamento de texto implementado neste trabalho.

A figura 9 traz uma exemplo de notícia armazenada em formato texto como parte do *corpus* de notícias elaborado por este projeto. A descrição completa do *corpus* é feita na seção 4.1.

Figura 9 – *Corpus* de notícias: um exemplo



Fonte: José Luiz Maturana Pagnossim, 2018

3.2.2 Pré-processamento de texto

O módulo de pré-processamento de texto utilizado neste projeto seguiu os procedimentos especificados no relatório técnico elaborado por Diaz et al. (2018). A implementação foi feita no ambiente *R* utilizando comandos nativos da linguagem e as funções de mineração de texto disponíveis na biblioteca *TM*. O relatório em questão, apresenta os detalhes técnicos a respeito da implementação dos procedimentos de pré-processamento em *R*. A seguir são listados os principais comandos com os respectivos parâmetros utilizados pelo algoritmo de pré-processamento de texto implementado neste trabalho:

- Carga dos documentos para a memória - o carregamento dos documentos para a memória foi feito por meio de um comando que percorreu os arquivos texto a partir de um diretório de entrada e recebeu como parâmetros: a codificação "UTF-8"[3]; o indicador para ignorar letras minúsculas e maiúsculas (foi definido como falso); o modo de leitura do arquivo (foi definido como texto); e o idioma (definido como português).
- Preparação dos dados - para otimizar os procedimentos de pré-processamento, as seguintes ações de preparação do texto foram utilizadas: conversão do texto para letras minúsculas; remoção de pontuações; remoção de espaços desnecessários; e remoção de números.

[3] 8-bit Unicode Transformation Format

- Análise Léxica - foi utilizado “espaço” como separador (*token*) de termos, que é o padrão do *R* para mineração de texto.
- Eliminação de palavras irrelevantes - definiu-se uma lista de palavras para serem removidas do texto, principalmente composta por artigos, pronomes, preposições e verbos, além de palavras que apareciam de forma recorrente devido à fonte das notícias, como: “ebc”; “agênciabrasil” e “radionacional”.
- Redução para o radical - foi utilizada a biblioteca *SnowballC* que implementa o algoritmo *word steeming* de Porter[4].
- Criação de uma representação vetorial para os textos - a representação foi baseada na métrica *TF-IDF*[5] normalizada. Esta etapa transformou o texto em uma matriz de termos (colunas) e documentos (linhas). Foram utilizados também parâmetros para considerar palavras de pelo menos duas letras. Não foi limitada a quantidade de vezes que um termo deveria aparecer para ser incorporado na representação vetorial, desta forma, se uma palavra aparecia pelo menos uma vez no texto, ela era incluída na representação vetorial.

Na figura 10 é ilustrado um exemplo da matriz *TF-IDF* gerada pelos procedimentos de pré-processamento. Os documentos destacados na matriz possuem a palavra “angola” em comum. No documento “2058” é possível observar que o índice *TF-IDF* normalizado é maior que no documento “2563”, o que significa que no primeiro documento há uma maior incidência da palavra em questão. A demonstração da incidência da palavra “angola” nestes documentos está ilustrada na figura 11.

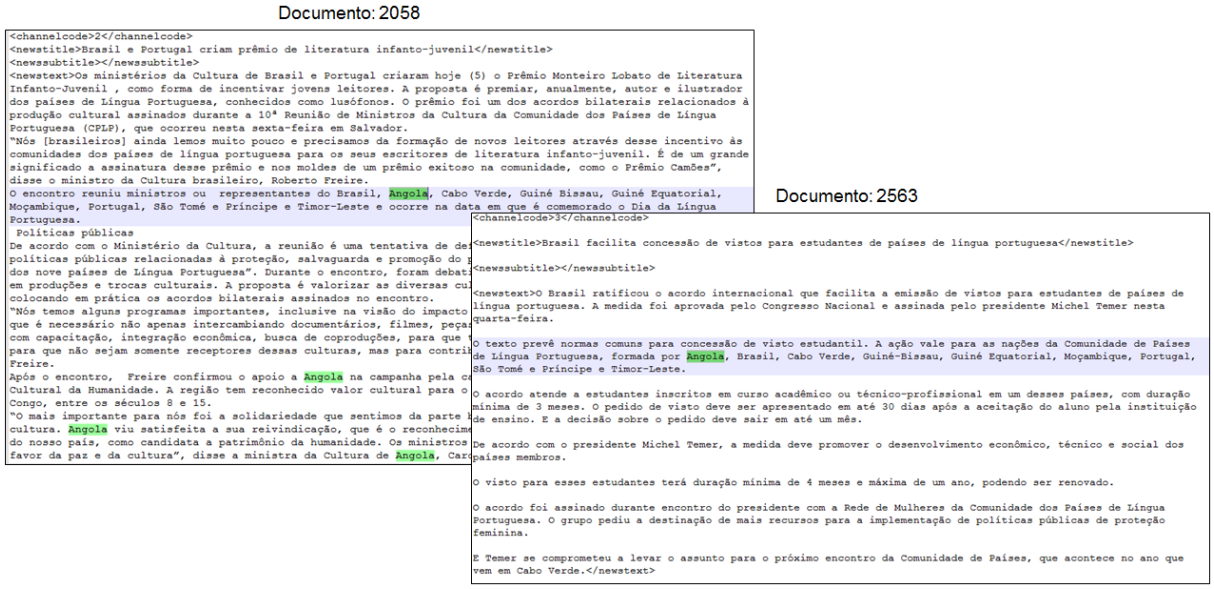
Figura 10 – Matriz *TF-IDF*: um exemplo

	acordo	africanas	ainda	algo	algun	ambique	amer	amizad	ancestralidades	angola
2058	0,014792233	0,024100773	0,004653552	0,0150209	0,009521799	0,021476101	0,024100773	0,024100773	0,024100773	0,085904403
2061	0,011116057	0	0,020982289	0	0	0	0	0	0	0
2065	0	0	0	0	0	0	0	0	0	0
2563	0,054850032	0	0	0	0	0,059725506	0	0	0	0,059725506
2564	0	0	0	0	0	0	0	0	0	0

Fonte: José Luiz Maturana Pagnossim, 2018

[4] https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf
[5] Term Frequency-Inverse Document Frequency

Figura 11 – Análise da incidência da palavra “angola” nos documentos



Fonte: José Luiz Maturana Pagnossim, 2018

3.2.3 Conjunto de dados

O termo conjunto de dados refere-se ao resultado produzido pela tarefa de pré-processamento sobre o *corpus*, que normalmente é uma estrutura de dados como: uma matriz de termo e frequência; uma matriz de similaridade ou de dissimilaridade; ou uma matriz transposta de *TF-IDF*. Estas estruturas de dados podem ser exportadas em arquivos ou planilhas a partir dos ambientes de programação em que foram geradas, ou ainda podem ser usadas como parâmetro de entrada para outras tarefas de mineração de dados, como classificação e agrupamento.

No algoritmo de pré-processamento utilizado neste projeto, o procedimento utilizado foi gerar uma matriz transposta da matriz *TF-IDF*. Esta matriz transposta foi passada como parâmetro para uma uma função de cálculo da distância (usando a medida cosseno). A partir da matriz de distância, foi gerado o conjunto de dados[6].

Um recorte do conjunto de dados gerado pelo algoritmo de pré-processamento implementado neste projeto está ilustrado na figura 12, que destaca a distância (em medida cosseno) entre um documento (destacado no eixo vertical) com outros três documentos (apresentados no eixo horizontal). Ao fazer o cruzamento entre os eixos nota-se que a distância de um documento em relação a ele mesmo é zero. Isto ocorre pelo fato da faixa de

[6] Em *R* esta estrutura é denominada *data frame*

valores da medida cosseno estar entre $[0-1]$, e quanto mais próximo de zero, mais similares são os documentos, no outro extremo, quanto mais próximo de um, mais diferentes são. A figura também ilustra os títulos das notícias referentes aos documentos utilizados por este recorte.

Figura 12 – Matriz de distância: medida cosseno

	10935	10936	10937	10938
10935	0.0	0.995153530232612	0.992254261916939	0.987302927311235
10970	0.826895138128555	0.991210878296086	0.993488856549057	0.985158137249534
11161	0.319507764518661	0.988779635960306	0.988584315354435	0.979790371458938
2554	0.999783068026568	0.985249099100905	0.994882720540757	0.996106229400787

Marque no relógio: às 20h34 começa o o eclipse lunar que avermelha a Lua
Petrobras reduz preço da gasolina em 1,4% e sobe o diesel em 0,7%
Entenda o que é a deflação e os efeitos da queda de preços na economia
Petrobras reduz em 0,5% preços do diesel e da gasolina nas refinarias

Fonte: José Luiz Maturana Pagnossim, 2018

3.2.4 Agrupamento

Em mineração de dados, a tarefa de agrupamento é considerada não supervisionada, já que não possui conhecimento a priori sobre os rótulos. O *corpus* de notícias utilizado neste trabalho, foi obtido do portal de notícias EBC com rótulos pré-definidos caracterizados pelos canais de leitura, apesar disto, esta pesquisa buscou novos grupos de notícias que não fossem conhecidos a priori. Neste contexto, a tarefa de agrupamento foi considerada adequada para que esse objetivo fosse cumprido.

Tarefas de agrupamento usam medidas de distância para encontrar grupos de itens similares. Alguns algoritmos que implementam tarefas de agrupamento precisam das medidas de distância como parâmetro de entrada para sua execução, como é o caso do *k-means*, cuja escolha para aplicação neste trabalho adveio adveio do conhecimento da literatura relacionada à mineração de textos, que considera o *k-means* como sendo simples e rápido de ser implementado, além de se mostrar eficiente quando usado para agrupamento de conjuntos de dados textuais.

O algoritmo *k-means* utilizado em linguagem *R* requer dois parâmetros de entrada para sua execução: o primeiro refere-se ao conjunto de dados gerado a partir da matriz de distância; e o segundo trata-se do valor de k que representa a quantidade de grupos pretendida na saída do algoritmo.

A matriz de distância armazena o valor da distância entre todos os documentos do conjunto de dados, ou seja, para o *corpus* de notícias utilizado por este trabalho, a estrutura de dados resultante ficou dimensionada em 1097 linhas por 1097 colunas.

Na literatura é possível encontrar métodos para determinação da quantidade de grupos que se pretende gerar a partir da tarefa de agrupamento. Esta decisão não é fácil de ser tomada, já que se pode acarretar em grupos mal formados ou em itens deslocados de sua real proximidade.

Neste trabalho, foi utilizado um procedimento que visa determinar o valor ideal para o número de grupos com base em validações internas e externas. Neste sentido foram feitas variações na execução do algoritmo *k-means* e os resultados gerados foram submetidos às validações externa (índice *RAND* ajustado) e interna (índice *silhouette*). Foram simulados valores para k na faixa de $[2-10]$. Os parâmetros utilizados para determinação do número de grupos estão ilustrados na tabela 2.

Tabela 2 – Parâmetros para determinação do valor ideal de k

Parâmetro	**Quantidade**	**Descrição**
Fonte de conteúdo	1	Portal EBC
Documentos no corpus	1097	Arquivos texto
Rótulos na origem	7	Canais de notícias
Algoritmo	1	*k-means*
Variação de grupos	9	k de 2 até 10
Medidas	1	Cosseno
Validações do agrupamento	2	Externa e Interna
Variação de execuções	30	Para cada k, 30 execuções

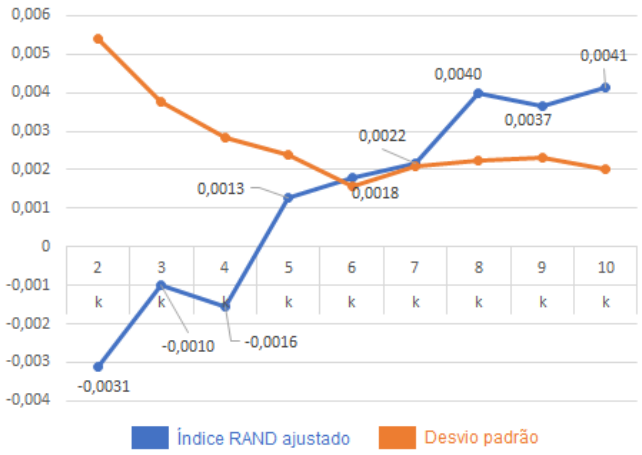
Fonte: José Luiz Maturana Pagnossim, 2018

As variações na execução do algoritmo *k-means* resultaram em trinta valores diferentes para cada índice utilizado. A partir desta extração foi possível calcular a média e o desvio padrão dos índices, para cada tipo de validação e para cada variação de k. Vale lembrar que para efeito de análise de ambos indicadores, quanto maior seus valores, melhor é o índice. A média e o desvio padrão consolidado para o índice *RAND* ajustado está ilustrado na figura 13. Vale destacar que a validação externa recebeu como parâmetro de entrada os rótulos obtidos das notícias a partir da fonte de conteúdo (portal EBC).

Observa-se no gráfico que o valor do índice *RAND* ajustado tende a aumentar conforme o valor de k é incrementado. O maior valor calculado por este índice foi de $0,0041$ para $k = 10$.

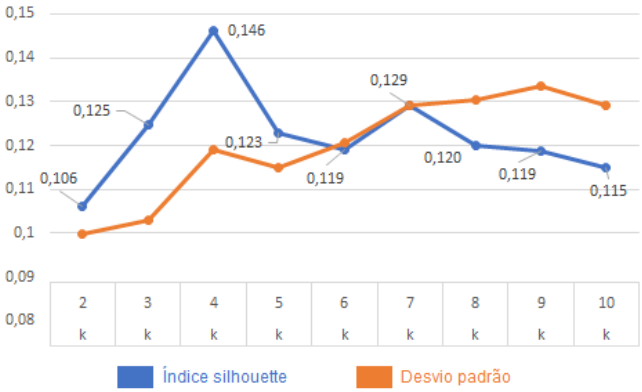
A figura 14 ilustra a média e o desvio padrão consolidado para o índice *Silhouette*.

Figura 13 – Resultado da validação externa usando o índice *RAND* ajustado



Fonte: José Luiz Maturana Pagnossim, 2018

Figura 14 – Resultado da validação interna usando o índice *Silhouette*



Fonte: José Luiz Maturana Pagnossim, 2018

Observa-se no gráfico que o valor do índice *silhouette* cresce a cada iteração de k até que $k = 4$, na sequência o índice cai para $k = 5$ e $k = 6$, volta a crescer para $k = 7$ e a partir de $k = 8$, o valor do índice começa a cair. Neste caso, o maior valor do índice foi para $k = 4$.

Com base na análise comparativa dos tipos de validação, ilustrada nos gráficos, a definição do valor de k, ficaria de 8 a 10 (no caso do índice *RAND* ajustado), já para o índice *silhouette*, o valor ideal apontado foi para $k = 4$. Uma forma de priorizar um dos dois índices é considerando que a validação *silhouette*, por ser interna, utiliza dados objetivos (como a similaridade entre os documentos), enquanto que a validação externa, por meio do índice *RAND* ajustado, utiliza uma fonte de dados externa para comparação, que neste caso foram os canais de notícias obtidos do portal EBC. Porém, para $k = 4$ (melhor índice *silhouette*), o *RAND* ajustado apresentou um valor negativo. Desta forma,

buscou-se o maior valor do índice *silhouette* a partir de $k > 4$, e quanto mais próximo do maior valor de k, mais haveria uma convergência em relação ao índice *RAND* ajustado.

Neste contexto, a decisão sobre o valor de k para utilização no algoritmo de agrupamento deste projeto foi para $k = 7$, por ser o maior valor de k para o índice *silhouette*, considerando $k > 4$ (para que o *RAND* ajustado não fosse negativo) e $k <= 7$ (para que o valor do índice *silhouette* não entrasse na curva descendente) .

Este procedimento gerou arquivos que foram usados para calcular a média e desvio padrão dos índices para cada valor de k. Após este processo foi verificado em qual execução (dentre as 30) o valor do índice *silhouette* havia sido maior. Como resultado destas verificações, obteve-se a formação dos grupos pelo algoritmo de agrupamento.

A título de exemplo, a tabela 3, ilustra um recorte do resultado do agrupamento produzido pelo algoritmo *k-means* em que foram selecionadas as 6 primeiras notícias para que se pudesse ter uma ideia de como a saída do algoritmo ficou disposta.

Tabela 3 – Resultado do agrupamento: recorte de seis notícias

Documento	**Grupo**	**Título da notícia**
2058	4	Brasil e Portugal criam prêmio de literatura infanto-juvenil
2061	4	Alunos das redes federal e municipais do Rio vão perder gratuidade ...
2065	7	Abertas inscrições para 1.400 vagas em cursos técnicos do Instituto...
2071	2	Bate Bola Nacional repercute demissão de Eduardo Baptista
2073	2	Atlético-MG vence o Sport Boys e encaminha classificação ...
2076	2	Bate Bola Nacional destaca a participação dos clubes brasileiros...

Fonte: José Luiz Maturana Pagnossim, 2018

Nota-se na tabela que apesar da rotina de agrupamento não conhecer os rótulos a priori, os exemplos demonstram que os grupos foram organizados próximos do que seriam os canais de notícias definidos em um jornal.

3.2.5 Carga estruturada de dados

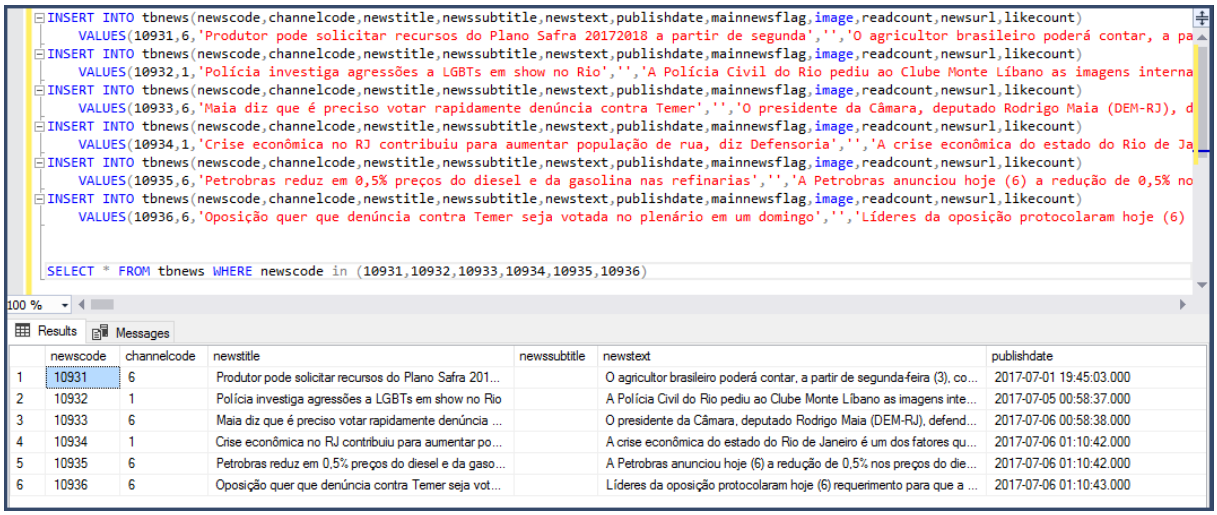
Nesta camada da arquitetura (figura 8) foram elaborados comandos para integração com a base de casos a partir de três fontes de dados de origem: o *corpus*; o conjunto de dados; e os grupos identificados pela tarefa de agrupamento.

A partir do *corpus* foram carregados os dados da notícia e os dados de domínios, como o canal de leitura. Por meio do conjunto de dados foi carregada na base de casos a

matriz de distância cosseno. Por fim, foi carregada na base de casos, a relação entre os documentos (notícias) e os grupos definidos pelo algoritmo de agrupamento.

Os comandos foram elaborados em linguagem *SQL*[7], compatível com a ferramenta utilizada como repositório da base de casos. A figura 15 ilustra um exemplo de carga massiva do *corpus* de notícias para a base de casos.

Figura 15 – Carga estruturada de dados: um exemplo de carga a partir do *corpus*



```
INSERT INTO tbnews(newscode,channelcode,newstitle,newssubtitle,newstext,publishdate,mainnewsflag,image,readcount,newsurl,likecount)
    VALUES(10931,6,'Produtor pode solicitar recursos do Plano Safra 20172018 a partir de segunda','','O agricultor brasileiro poderá contar, a pa
INSERT INTO tbnews(newscode,channelcode,newstitle,newssubtitle,newstext,publishdate,mainnewsflag,image,readcount,newsurl,likecount)
    VALUES(10932,1,'Polícia investiga agressões a LGBTs em show no Rio','','A Polícia Civil do Rio pediu ao Clube Monte Líbano as imagens interna
INSERT INTO tbnews(newscode,channelcode,newstitle,newssubtitle,newstext,publishdate,mainnewsflag,image,readcount,newsurl,likecount)
    VALUES(10933,6,'Maia diz que é preciso votar rapidamente denúncia contra Temer','','O presidente da Câmara, deputado Rodrigo Maia (DEM-RJ), d
INSERT INTO tbnews(newscode,channelcode,newstitle,newssubtitle,newstext,publishdate,mainnewsflag,image,readcount,newsurl,likecount)
    VALUES(10934,1,'Crise econômica no RJ contribuiu para aumentar população de rua, diz Defensoria','','A crise econômica do estado do Rio de Ja
INSERT INTO tbnews(newscode,channelcode,newstitle,newssubtitle,newstext,publishdate,mainnewsflag,image,readcount,newsurl,likecount)
    VALUES(10935,6,'Petrobras reduz em 0,5% preços do diesel e da gasolina nas refinarias','','A Petrobras anunciou hoje (6) a redução de 0,5% no
INSERT INTO tbnews(newscode,channelcode,newstitle,newssubtitle,newstext,publishdate,mainnewsflag,image,readcount,newsurl,likecount)
    VALUES(10936,6,'Oposição quer que denúncia contra Temer seja votada no plenário em um domingo','','Líderes da oposição protocolaram hoje (6)

SELECT * FROM tbnews WHERE newscode in (10931,10932,10933,10934,10935,10936)
```

100 % — Results — Messages

	newscode	channelcode	newstitle	newssubtitle	newstext	publishdate
1	10931	6	Produtor pode solicitar recursos do Plano Safra 201...		O agricultor brasileiro poderá contar, a partir de segunda-feira (3), co...	2017-07-01 19:45:03.000
2	10932	1	Polícia investiga agressões a LGBTs em show no Rio		A Polícia Civil do Rio pediu ao Clube Monte Líbano as imagens inte...	2017-07-05 00:58:37.000
3	10933	6	Maia diz que é preciso votar rapidamente denúncia ...		O presidente da Câmara, deputado Rodrigo Maia (DEM-RJ), defend...	2017-07-06 00:58:38.000
4	10934	1	Crise econômica no RJ contribuiu para aumentar po...		A crise econômica do estado do Rio de Janeiro é um dos fatores qu...	2017-07-06 01:10:42.000
5	10935	6	Petrobras reduz em 0,5% preços do diesel e da gaso...		A Petrobras anunciou hoje (6) a redução de 0,5% nos preços do die...	2017-07-06 01:10:42.000
6	10936	6	Oposição quer que denúncia contra Temer seja vot...		Líderes da oposição protocolaram hoje (6) requerimento para que a ...	2017-07-06 01:10:43.000

Fonte: José Luiz Maturana Pagnossim, 2018

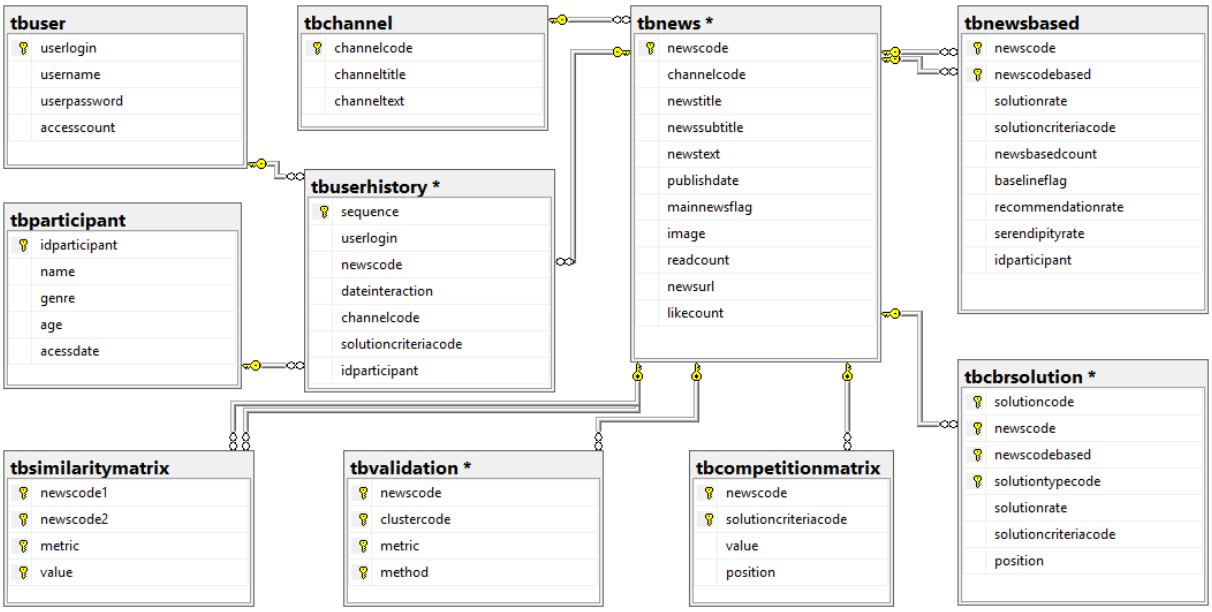
3.2.6 Base de casos

Para que fosse possível implementar o método de *RBC* e também as funcionalidades para validação de um sistema de recomendação, se fez necessária a adoção de uma modelagem de dados que utilizasse a abordagem relacional[8]. Este modelo lógico foi implementado na ferramenta *Microsoft SQL Server*[9]. A base de casos também foi usada como repositório de dados durante e após a execução do experimento *online*, que é detalhado na seção 4.2. A apresentação da base de casos é feita por meio de um modelo relacional, ilustrado na figura 16 e também pela descrição de todas as relações e os respectivos relacionamentos.

---

7 Structured Query Language
8 Notação de um modelo de dados relacional (NAVATHE; ELSMARI, 2013)
9 *Microsoft SQL Server 2017 Express.*

Figura 16 – Modelo da base de casos: notação de modelo de dados relacional



Fonte: José Luiz Maturana Pagnossim, 2018

- Relação *tbnews* - principal relação do modelo, responsável pelo armazenamento das notícias e seus atributos descritivos. Possui relacionamentos com outras sete relações do modelo.
- Relação *tbchannel* - relação de domínio para armazenar informações sobre os canais de notícias obtidos do *corpus*. Possui relacionamento com as relações *tbuser* e *tbnews*.
- Relação *tbuser* - relação responsável para armazenar o cadastro de usuário. Possui relacionamento com a relação *tbuserhistory*.
- Relação *tbparticipant* - relação responsável pelo armazenamento do registro do participante do experimento, independente se ele optou por navegar em modo anônimo ou cadastrado. Possui relacionamento com a relação *tbuserhistory*.
- Relação *tbuserhistory* - relação responsável pelo registro do histórico de navegação de usuários e leitores anônimos. Possui relacionamento com as relações *tbparticipant*, *tbuser* e *tbnews*
- Relação *tbnewsbased* - relação que tem o papel de representar o relacionamento entre uma notícia e as demais notícias relacionadas. Serve para estabelecer um vínculo entre uma notícia de origem e uma notícia de destino. Responsável por armazenar os contadores de aceite de recomendação e serendipidade, além de registrar a nota média da recomendação.

- Relação *tbsimilaritymatrix* - relação que recebe a carga da matriz de distância a partir do conjunto de dados. Possui relação com a relação *tbnews*.
- Relação *tbvalidation* - relação que recebe a carga da tarefa de agrupamento, incluindo a identificação do documento e o grupo que este documento está associado. Possui relacionamento com a relação *tbnews*.
- Relação *tbcompetitionmatrix* - relação criada para armazenar de forma temporária a matriz de competição usada para cálculo da métrica unificada para ranqueamento da recomendação. Possui relacionamento com a relação *tbnews*.
- Relação *tbcbrsolution* - relação criada para armazenar as soluções recomendadas pelo método *RBC*. Trata-se da representação do "caso". Nesta relação são guardados os atributos que caracterizam: o tipo de solução (conhecida ou adaptada); o critério de recuperação da notícia; o valor calculado para a métrica unificada de ranqueamento da recomendação; e a posição em que a notícia ficou colocada no ranqueamento.

3.2.7 Tela do protótipo do sistema

A tela do protótipo do sistema de recomendação foi desenvolvida em plataforma *web* e foi disponibilizada na internet para ser usada pelos participantes do experimento.

Foi utilizado como plataforma o ambiente de desenvolvimento integrado *Visual Studio.Net*[10], linguagem de programação *C#* e tipo de projeto *ASP.Net Web Application*. A camada de apresentação utilizou conteúdos simples e padronizados em *HTML*[11] e *CSS*[12] de forma que a interface gráfica ficasse objetiva e intuitiva ao usuário.

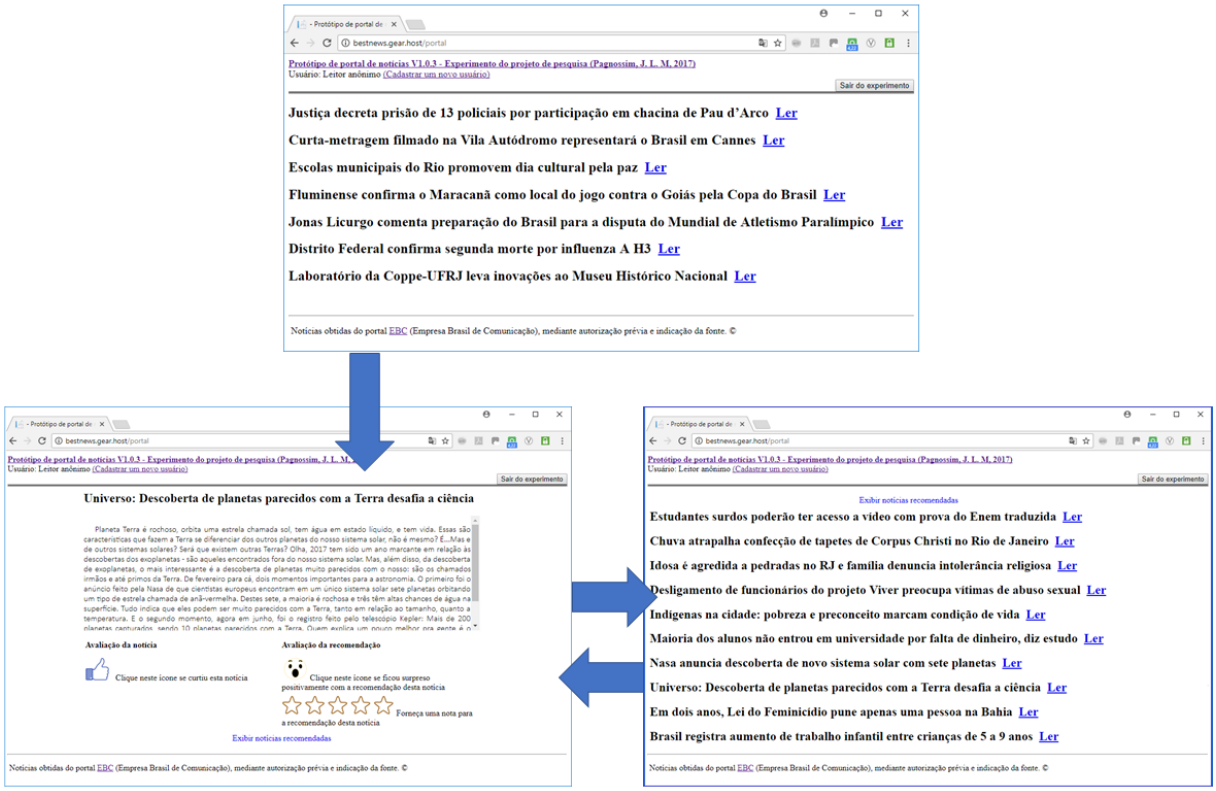
Os procedimentos de navegação entre as telas são descrito detalhadamente na seção 4.2.5. A figura 17 ilustra uma visão resumida do fluxo entre as telas.

3.2.8 *CRUD*

Este módulo é responsável pela integração entre a tela do protótipo do sistema e a base de casos. Esta integração provê os métodos para recuperação e manipulação de dados como: inserir, alterar, excluir e consultar.

[10] Microsoft Visual Studio Community 2015
[11] HyperText Markup Language
[12] Cascading Style Sheets

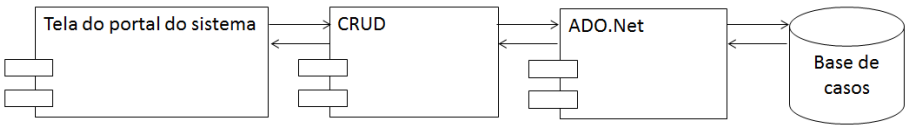
Figura 17 – Fluxo resumido da navegação entre as telas do protótipo do sistemal



Fonte: José Luiz Maturana Pagnossim, 2018

A programação desta camada foi feita na mesma solução em que foi desenvolvida a tela do protótipo do sistema, porém utilizando classes como estrutura de dados principais. Nestas classes foram implementados os métodos de conexão e manipulação da base de casos. O componente de acesso a dados utilizado foi o *ADO.Net*[13]. A figura 18 ilustra o diagrama de componentes que envolve o módulo *CRUD*.

Figura 18 – Diagrama de componentes: módulo *CRUD*



Fonte: José Luiz Maturana Pagnossim, 2018

[13] Activex Data Objects da plataforma Microsoft.Net

3.2.9 Recomendador

O recomendador é o núcleo de inteligência do sistema de recomendação, já que implementa os métodos de recomendação fundamentados neste trabalho.

Pode parecer pertinente que os *SR* façam recomendações utilizando critérios de similaridade, mas a similaridade pode não atender à prerrogativas de qualidade de recomendação como diversidade, serendipidade e novidade. Dentro desse contexto, para maximizar a qualidade da recomendação em relação às prerrogativas citadas, a alternativa adotada foi combinar diferentes critérios para recuperação da notícia e associá-los à estratégias de recomendação. Desta forma, o recomendador reune melhores condições para tomada de decisão e consequentemente mais chances de alcançar os resultados esperados.

Construir um algoritmo de recomendação com objetivo de sugerir notícias similares e relevantes ao leitor deve levar em consideração aspectos relacionados à: similaridade (entre as notícias e entre usuários); interação dos leitores com o portal de notícias; preferência do leitor; e critérios que permitam algum grau de imprevisibilidade. De posse dessas informações, uma estratégia de tomada de decisão do recomendador é ponderar diferentes critérios de recuperação de notícias, de forma que o resultado seja uma lista diversificada de notícias adequadas ao contexto e suficientemente boas às exigências dos usuários.

Considerando as particularidades da abordagem em discussão neste trabalho, a especificação do núcleo do sistema de recomendação segue formalizada nesta seção. Para a construção da formalização foram estabelecidas convenções conforme especificadas no quadro 1.

Critérios para recuperação da notícia

Para que a abordagem apresentada por este trabalho pudesse atingir os objetivos no que se refere à lista de recomendação de notícias, principalmente em aspectos relacionados à similaridade, popularidade, diversidade, novidade e serendipidade, uma lista com treze critérios para recuperação de notícias foi especificada, conforme ilustrada no quadro 2.

A escolha dos critérios foi fundamentada em tarefas, funções e propriedades de um sistema de recomendação, assim como as características dos diferentes tipos de *SR*. Desta forma, buscou-se atingir critérios que privilegiassem: recomendação baseada em

Quadro 1 – Convenções

Descrição	**Sigla**
Sistema de Recomendação	$\mathcal{SR}$
Base de Casos	$\mathcal{BC}$
Similaridade	$\boldsymbol{sim}$
Uma notícia	$\mathcal{N}$, representa um item
Uma lista	$\mathcal{L}$, um conjunto de elementos quaisquer
Um leitor	$\mathcal{E}$, que representa um leitor anônimo ao $\mathcal{SR}$
Uma lista de leitores	$\mathcal{LE} = \{E_1, \ldots E_n, \ldots, E_N\}$
Um usuário	$\mathcal{U}$, representa um usuário cadastrado no $\mathcal{SR}$
Uma lista de usuários	$\mathcal{LU} = \{U_1, \ldots U_n, \ldots, U_N\}$
Canal favorito do $\mathcal{U}$	$\mathcal{CFU}$
Uma interação de $\mathcal{E}$	$\mathcal{I} = \{E_1, N_1\}$ um $\mathcal{E}$ interage com uma $\mathcal{N}$
Uma interação de $\mathcal{U}$	$\mathcal{I} = \{U_1, N_1\}$ um $\mathcal{U}$ interage com uma $\mathcal{N}$
Uma $\mathcal{L}$ de $\mathcal{I}$ de um leitor	$\mathcal{LIE} = \{I_1, \ldots, I_m, \ldots, I_M\}$
Uma $\mathcal{L}$ de $\mathcal{I}$ de um usuário	$\mathcal{LIU} = \{I_1, \ldots, I_m, \ldots, I_M\}$
Uma notícia de origem	$\mathcal{NO}$, é a $\mathcal{N}$ de entrada para a recomendação
Uma notícia de destino	$\mathcal{ND}$, é a $\mathcal{N}$ recomendada pelo $\mathcal{SR}$
Uma $\mathcal{L}$ de $\mathcal{ND}$ ao $\mathcal{E}$	$\mathcal{LNDE} = \{ND_1, \ldots ND_n, \ldots, ND_N\}$
Uma $\mathcal{L}$ de $\mathcal{ND}$ ao $\mathcal{U}$	$\mathcal{LNDU} = \{ND_1, \ldots ND_n, \ldots, ND_N\}$
Uma solução conhecida	$\mathcal{SC}$
Uma solução adaptada	$\mathcal{SA}$
Um "caso" de uma $\mathcal{SC}$ a um $\mathcal{E}$	$\mathcal{CASOESC} = \{E_1, I_1, NO_1, , SC, LNDE\}$
Um "caso" de uma $\mathcal{SC}$ a um $\mathcal{U}$	$\mathcal{CASOUSC} = \{U_1, I_1, NO_1, , SC, LNDU\}$
Um "caso" de uma $\mathcal{SA}$ a um $\mathcal{E}$	$\mathcal{CASOESA} = \{E_1, I_1, NO_1, , SA, LNDE\}$
Um "caso" de uma $\mathcal{SA}$ a um $\mathcal{U}$	$\mathcal{CASOUSA} = \{U_1, I_1, NO_1, , SA, LNDU\}$

Fonte: José Luiz Maturana Pagnossim, 2018

Quadro 2 – Critérios para recuperação da notícia

Critério	**Descrição**
$\{CR_1\}$	Recupera $\mathcal{ND}$ com maior $\boldsymbol{sim}$ de conteúdo em relação à $\mathcal{NO}$
$\{CR_2\}$	Recupera a $\mathcal{N}$ com maior $\boldsymbol{sim}$ intergrupo
$\{CR_3\}$	Recupera a $\mathcal{N}$ com maior $\boldsymbol{sim}$ intragrupo
$\{CR_4\}$	Recupera uma $\mathcal{N}$ aleatória que não tenha sido recomendada anteriormente
$\{CR_5\}$	Recupera a $\mathcal{N}$ mais lida dentro do canal
$\{CR_6\}$	Recupera a $\mathcal{N}$ mais lida do portal
$\{CR_7\}$	Recupera a $\mathcal{N}$ mais curtida do canal
$\{CR_8\}$	Recupera a $\mathcal{N}$ mais curtida do portal
$\{CR_9\}$	Recupera $\mathcal{ND}$ com mais recomendações aceitas em relação à $\mathcal{NO}$
$\{CR_{10}\}$	Recupera $\mathcal{ND}$ com maior nota de avaliação em relação à $\mathcal{NO}$
$\{CR_{11}\}$	Recupera $\mathcal{ND}$ com maior indicador de serendipidade em relação à $\mathcal{NO}$
$\{CR_{12}\}$	Recupera a última $\mathcal{N}$ lida pelo $\mathcal{U}$ mais similar
$\{CR_{13}\}$	Recupera a $\mathcal{N}$ mais lida do $\mathcal{CFU}$ corrente

Fonte: José Luiz Maturana Pagnossim, 2018

conteúdo ($\{CR_1\}$); similaridade com base em agrupamentos ($\{CR_2\}$ e $\{CR_3\}$); diversidade e novidade ($\{CR_4\}$); popularidade ($\{CR_5, \ldots, CR_{10}\}$); serendipidade ($\{CR_{11}\}$); recomendação baseada em filtro colaborativo usuário-usuário ($\{CR_{12}\}$); e personalização de usuário ($\{CR_{13}\}$).

Para que o critério $\{CR_{12}\}$ seja recomendado, o algoritmo de recomendação recupera o canal favorito por meio do histórico de navegação do usuário corrente. Recupera o usuário com mais leituras do mesmo canal favorito e estabelece a similaridade entre os usuários. Já o critério $\{CR_{13}\}$ obtém o canal favorito por meio do histórico de navegação do usuário. Estes critérios não são recomendados a leitores anônimos, mas somente aos usuários cadastrados no sistema.

Todos os critérios apresentados foram implementados no módulo de recomendação, integrado ao protótipo de sistema de recomendação usado como ambiente de prova de conceito desta abordagem. Estes critérios são acionados pelo recomendador dependendo da disponibilidade de um ou mais indicadores e também de acordo com as estratégias utilizadas para fornecer as recomendações.

Estratégias para recomendação baseada em critérios de recuperação da notícia

Na lógica de construção da recomendação utilizada na abordagem em discussão, os critérios de recuperação de notícias são combinados com as estratégias de recomendação. O objetivo dessa combinação é a produção de recomendações adequadas aos leitores e usuários. Tais estratégias podem ser construídas de diferentes formas, a depender do problema ou desafio a ser superado e merecem um trabalho de calibração contínua a ser analisada pelos responsáveis pelo sistema. Para o contexto de apresentação deste trabalho é pertinente discutir as estratégias adotadas em termos do problema de recomendação que elas visam resolver, conforme demonstradas no quadro 3.

Quadro 3 – Estratégias para recomendação com base em critérios

Estratégia	**Problema associado**	**Critérios combinados**
$\{ES_1\}$	*Cold start* de item	$\{CR_1\}$, $\{CR_2\}$, $\{CR_3\}$ e $\{CR_4\}$
$\{ES_2\}$	Previsibilidade e aleatoriedade	$\{CR_5, \ldots, CR_{11}\}$
$\{ES_3\}$	Falta de personalização de usuário	$\{CR_{12}\}$ e $\{CR_{13}\}$

Fonte: José Luiz Maturana Pagnossim, 2018

Para atacar tais problemas, as estratégias devem combinar técnicas e propriedades fundamentadas pelos *SR*, mas para cada estratégia adotada, um ou mais efeitos colaterais podem ser gerados e assim o sistema implementa uma outra estratégia para combater os efeitos colaterais, de tal forma que a solução híbrida disponha de recursos capazes de resolver os problemas e minimizar os efeitos colaterais.

No contexto deste trabalho, a $\{ES_1\}$ ataca o problema de *cold start* de item fornecendo recomendações com base em: similaridade de conteúdo, similaridade por agrupamento e aleatoriedade. Os efeitos colaterais destes tipos de recomendações estão associados à previsibilidade e à própria aleatoriedade. Desta forma, para atacar estes efeitos colaterais, a $\{ES_2\}$ proporciona recomendações que privilegiam: novidade, diversidade, relevância, popularidade e serendipidade. O efeito colateral desta estratégia é a falta de personalização de usuário, que é atacada pela $\{ES_3\}$ por meio do uso da recomendação baseada em filtro colaborativo (modelo de vizinhança usuário-usuário) e pelo estudo das preferências do usuário (com base no histórico de navegação).

A $\{ES_3\}$ apesar de complementar as demais estratégias, ainda assim, pode gerar efeitos colaterais, por exemplo de privacidade, já que a recomendação usuário-usuário pode ocasionar uma recomendação por um lado surpreendente, mas às vezes indesejada. Na abordagem proposta, o efeito colateral da privacidade não foi motivo de avaliação, sendo necessários estudos que envolvam o estabelecimento de protocolos de privacidade entre o sistema e os usuários. Por se tratar de um experimento científico aprovado pelo comitê de ética e com os voluntários devidamente esclarecidos, não houve risco deste efeito ocorrer.

Métrica unificada para ranqueamento da recomendação ($\mathcal{MURR}$)

Na etapa final da recomendação o recomendador tem que ser capaz de listar e ordenar os itens que estão sendo recomendados. Do ponto de vista prático, a sugestão fornecida por cada critério de recuperação de notícias apresentada no quadro 2, já agrega qualidade à recomendação, mas o desafio é encontrar dentro da lista de notícias uma ordem de relevância para apresentá-la aos leitores e usuários do sistema. Para isso foi necessário definir e normalizar uma unidade de medida comum entre os critérios de recuperação, já que estes possuíam diferentes escalas de valores.

É importante considerar que se o sistema conseguir reunir indicadores suficientes para fazer a recomendação com base nos treze critérios especificados no quadro 2, isto significa que cada notícia recuperada foi a melhor colocada dentro do critério pelo qual foi selecionada.

A métrica unificada para ranqueamento da recomendação ($\mathcal{MURR}$) foi formulada a partir de um cenário de competição entre as notícias selecionadas pelos critérios de recuperação. A competição é baseada na soma de colocações que a notícia assume dentro de cada critério de recuperação sob consideração. Desta forma, quanto menor for esta somatória, melhor é a colocação da notícia dentro da lista ordenada.

Para demonstrar a fórmula desta métrica, considere $\boldsymbol{m}$ o número de critérios de recuperação de notícias e $\boldsymbol{Pos}$ a posição em que a notícia ficou colocada no ranqueamento de um determinado critério (CR). A ($\mathcal{MURR}$) está formalizada na equação 4,

$$MURR = \sum_{m=1}^{m=13} Pos(CR_m) \tag{4}$$

Como ilustração da lógica de construção desta métrica, considere o melhor cenário possível para colocação da notícia dentro do ranqueamento, ou seja ela ficou em primeiro lugar no critério de recuperação pelo qual foi selecionada e em segundo lugar em todos os outros critérios. Somando, desta forma, 1 ponto do seu critério e 24 pontos referentes aos 2 pontos de cada um dos outros 12 critérios, totalizando 25 pontos (menor valor que esta métrica pode assumir). Um outro exemplo, considera a pior colocação possível, ou seja, 157 pontos no total, obtidos a partir do seguinte raciocínio: 1 ponto do critério pelo qual a notícia foi vencedora e 13 (última colocação) pontos multiplicados pelos 12 critérios restantes.

Nesta primeira etapa da simulação é populada uma estrutura denominada de matriz de competição, composta pelas notícias selecionadas (N_m) em relação aos critérios de recuperação (CR_m). Para cada noticia a matriz recebe a posição em que esta notícia ficou colocada em relação a cada critério. Vale ressaltar que as células que mostram a notícia vencedora para cada critério estão em destaque e que uma mesma notícia não pode ser vencedora em mais de um critério, dessa forma não há notícias repetidas na lista de recomendação. Um exemplo do uso desta métrica está ilustrado na figura 19.

Figura 19 – Matriz de competição: simulação do cálculo da *MURR*

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8	CR9	CR10	CR11	CR12	CR13
N1	1	2	2	2	12	2	6	12	5	5	12	8	13
N2	2	1	12	3	11	3	2	11	6	6	11	7	13
N3	3	12	1	5	10	5	3	10	7	7	10	6	13
N4	5	9	11	1	2	7	5	2	8	8	9	5	13
N5	4	10	9	4	1	4	4	7	9	9	8	4	13
N6	7	11	10	7	9	1	7	9	10	10	7	3	13
N7	6	8	8	6	8	6	1	8	11	11	6	2	13
N8	8	7	7	8	7	8	8	1	12	12	5	12	13
N9	9	5	5	9	5	9	9	5	1	2	4	11	13
N10	10	6	6	10	6	10	10	6	2	1	3	10	13
N11	12	3	3	12	3	12	12	3	3	4	1	9	13
N12	11	4	4	11	4	11	11	4	4	3	2	1	13
N13	13	13	13	13	13	13	13	13	13	13	13	13	1

Fonte: José Luiz Maturana Pagnossim, 2018

A segunda etapa é a ordenação pelo menor valor da $\mathcal{MURR}$ e caso haja empate no valor da $\mathcal{MURR}$, o desempate é feito pelo atributo de data/hora de publicação, privilegiando o aspecto da novidade. Esta etapa está ilustrada na figura 20.

Figura 20 – Matriz de competição: ranqueamento pela *MURR*

	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8	CR9	CR10	CR11	CR12	CR13	MURR	Posição
N1	1	2	2	2	12	2	6	12	5	5	12	8	13	82	1
N12	11	4	4	11	4	11	11	4	4	3	2	1	13	83	2
N4	5	9	11	1	2	7	5	2	8	8	9	5	13	85	3
N5	4	10	9	4	1	4	4	7	9	9	8	4	13	86	4
N9	9	5	5	9	5	9	9	5	1	2	4	11	13	87	5
N2	2	1	12	3	11	3	2	11	6	6	11	7	13	88	6
N11	12	3	3	12	3	12	12	3	3	4	1	9	13	90	7
N3	3	12	1	5	10	5	3	10	7	7	10	6	13	92	8
N10	10	6	6	10	6	10	10	6	2	1	3	10	13	93	9
N7	6	8	8	6	8	6	1	8	11	11	6	2	13	94	10
N6	7	11	10	7	9	1	7	9	10	10	7	3	13	104	11
N8	8	7	7	8	7	8	8	1	12	12	5	12	13	108	12
N13	13	13	13	13	13	13	13	13	13	13	13	13	1	157	13

Fonte: José Luiz Maturana Pagnossim, 2018

Estratégias para *RBC*

Além de estratégias que combinam critérios de recuperação de notícias apresentadas anteriormente, a abordagem faz uso da *RBC* e a partir dos recursos deste tipo de recomendação são definidas duas outras estratégias, conforme ilustradas no quadro 4.

A estratégia $\{ES_4\}$ é implementada para as situações em que a $\mathcal{NO}$ já tenha sido anteriormente consumida. Na ocasião do aceite desta recomendação, a lógica da *RBC*

Quadro 4 – Estratégias para RBC

Estratégia	**Contexto da utilização**
$\{ES_4\}$	Existe uma solução conhecida para o problema
$\{ES_5\}$	Trata-se de um novo problema, busca-se uma solução adaptada

Fonte: José Luiz Maturana Pagnossim, 2018

armazena informações sobre tal consumo na forma de "casos" (conforme a quíntupla que descreve $\mathcal{CASOESC}$ e $\mathcal{CASOUSC}$). Na reincidência da leitura da mesma $\mathcal{NO}$ por outro leitor ou usuário, a lista de recomendação associada ao "caso" pode ser reutilizada.

A estratégia anterior descreve uma situação em que a reutilização da solução pode ser feita de forma direta, fazendo uso da mesma recomendação realizada no passado. Mas quando o sistema se depara com um novo problema, o "caso" é desconhecido, o que proporciona a aplicação da estratégia de recomendação $\{ES_5\}$. A implementação desta estratégia faz uso matriz de distância, recuperando a $\mathcal{ND}$ mais similar em relação à $\mathcal{NO}$. De posse desta $\mathcal{ND}$ similar, o recomendador verifica se ela tem uma solução conhecida registrada. Se esta solução for localizada, o sistema recupera uma das quíntuplas especificadas em $\mathcal{CASOESA}$ e $\mathcal{CASOUSA}$ e a define como uma solução adaptada.

A principal vantagem destas estratégias é o reaproveitamento de soluções conhecidas e adaptadas, considerando que se elas foram úteis no passado, podem ser úteis também em recomendações futuras. O reaproveitamento também ocorre em relação à ordenação da lista de recomendação, já que um "caso" para ser conhecido foi gerado a partir de outras estratégias, que obrigatoriamente passaram pela aplicação da $\mathcal{MURR}$. Uma outra vantagem está associada a melhor eficiência na recuperação da recomendação em relação à execução do algoritmo baseado em critérios de recuperação de notícias. Apesar desta vantagem descrita, ela não foi alvo de estudo neste trabalho.

A partir da especificação das estratégias de recomendação, algumas premissas podem ser derivadas, a fim de formalizar a relação entre as estratégias e o sistema de recomendação:

- A $\{ES_1\}$ é obrigatoriamente a primeira recomendação do sistema a partir do momento da sua implantação.
- A $\{ES_2\}$ é usada se o sistema reunir os indicadores de popularidade.
- A $\{ES_3\}$ só é possível de ser aplicada se houver usuários cadastrados no sistema.
- A $\{ES_4\}$ só é possível de ser aplicada se houver uma solução conhecida para o problema apresentado.

- A $\{ES_5\}$ só é possível de ser aplicada se não houver uma solução conhecida para $\mathcal{NO}$ e se houver pelo menos uma solução conhecida para a notícia mais similar em relação à $\mathcal{NO}$.

A partir destas premissas foi implementada uma regra para equilibrar o uso das estratégias pelo sistema, conforme ilustrada no algoritmo 2.

Algoritmo 2 Algoritmo Estratégias de Recomendação

```
1: procedure UTILIZARESTRATEGIASRECOMENDACAO()
2:     Se existe uma solução RBC conhecida
3:         Seja vAleatorio um valor aleatório entre verdadeiro e falso
4:         Se vAleatorio é igual a verdadeiro
5:             Usa a solução conhecida RBC para recomendar
6:         Senao
7:             Usa o algoritmo de recomendação baseado nos critérios de recuperação
8:         FimSe
9:     Senao
10:        Se existe solução RBC adaptada
11:            Seja vAleatorio um valor aleatório entre verdadeiro e falso
12:            Se vAleatorio é igual a verdadeiro
13:                Usa a solução adaptada RBC para recomendar
14:            Senao
15:                Usa o algoritmo de recomendação baseado nos critérios de recuperação
16:            FimSe
17:        Senao
18:            Usa o algoritmo de recomendação baseado nos critérios de recuperação
19:        FimSe
20:    FimSe
```

Fonte: José Luiz Maturana Pagnossim, 2018

4 Validação da abordagem

A validação da abordagem foi conduzida utilizando o método de avaliação *online*. Segundo Gunawardana (2015), "a avaliação *online* está interessada em medir o comportamento do usuário quando este interage com diferentes sistemas de recomendação". Neste trabalho, o método de avaliação *online*, foi aplicado em um experimento que utilizou um protótipo de portal de notícias, construído para provar o conceito da abordagem apresentada. Tal experimento foi responsável por capturar dados de navegação dos usuários durante a participação destes no experimento. Este capítulo apresenta o *corpus* utilizado, detalha o formato do experimento, apresenta e discute os resultados. Ao final são apresentadas as limitações e ameaças ao trabalho.

4.1 O corpus

Obter um *corpus* público de notícias que esteja pronto para utilização em um projeto é uma tarefa que demanda pesquisa e esforço para analisar as características do *corpus* em relação às necessidades do projeto. Há alguns *corpora* de notícias em idioma inglês já explorados por outras pesquisas, como os *datasets*: *Reuters*[1] e *20 NewsGroups*[2]. No caso da base *Reuters*, trata-se de um *corpus* com mais de vinte e uma mil notícias publicadas pela agência *Reuters* no ano de 1987. Já a base 20 *NewsGroups* é um *corpus* com aproximadamente vinte mil documentos, cujo conteúdo não contém textos de notícias, mas sim, mensagens extraídas de fóruns de discussão a respeito de notícias. Para o idioma português (Brasil) encontrar um *corpus* de notícias pronto para ser usado é uma tarefa ainda mais difícil.

Durante a fase de pesquisa por um *corpus* para utilização neste trabalho, foi localizada uma base em idioma português (Brasil) com notícias do jornal folha de São Paulo. Após contato com o grupo Linguateca[3], que detém os direitos de uso deste *corpus*, foi concedida a autorização para utilização deste conteúdo. Após a exploração deste conteúdo, foi possível concluir que, apesar do *corpus* ter uma quantidade considerável de documentos

[1] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection
[2] https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups
[3] http://www.linguateca.pt/

(cerca de cem mil), o conteúdo das notícias era resumido (semelhante a manchetes de jornal), adicionalmente, tratavam-se de notícias dos anos de 1994 e 1995.

Ao analisar as características destes *corpora* já disponíveis e confrontar com os requisitos desta pesquisa, principalmente no que diz respeito aos critérios de recomendação de notícias, os fatores preponderantes para tomada de decisão sobre a obtenção de um destes *corpus* ou pela elaboração de um *corpus* próprio foram: o idioma e a data de publicação das notícias.

Com relação ao idioma, optou-se pelo idioma português (Brasil), já que o experimento *online* seria realizado com grupos de usuários do Brasil. Sobre a data de publicação, constatou-se que, para atender fatores como novidade e serendipidade, a utilização de notícias mais recentes era essencial.

Diante deste cenário, a alternativa foi buscar um portal público com notícias brasileiras para que pudesse servir como fonte de conteúdo de notícias e a partir deste portal elaborar um *corpus* específico para utilização neste trabalho.

Dentre as opções de portais públicos de notícias que pudessem servir como base para este projeto foram avaliados: o portal do Banco Nacional de Desenvolvimento (BNDES)[4]; o portal de notícias da prefeitura da cidade de Curitiba, no estado do Paraná[5]; e o portal da Empresa Brasil de Comunicação (EBC)[6]. No caso dos portais do BNDES e da prefeitura de Curitiba, foram atendidos os requisitos relacionados ao caráter público das notícias, ao idioma e à data atual de publicação das notícias, por outro lado, o conteúdo de ambos era voltado integralmente aos interesses específicos de cada entidade, não atendendo ao critério da diversidade e também poderia prejudicar a avaliação da serendipidade.

A opção que melhor atendeu aos interesses desta pesquisa, considerando os critérios do idioma, do caráter público da base, da diversidade e da novidade foi o portal da EBC[7], tal obtenção foi iniciada após autorização concedida pela gerência de conteúdo web do portal. O portal da EBC é organizado por assuntos em sete diferentes canais de notícias: cidadania, cultura, educação, esportes, infantil, notícias (que engloba política, economia, saúde, internacional e meio ambiente) e tecnologia.

---

4 http://www.bndes.gov.br/wps/portal/site/home/imprensa/noticias/

5 http://www.curitiba.pr.gov.br/servicos/cidadao/0/0/0/1

6 http://www.ebc.com.br

7 Segundo o site da EBC "a Empresa Brasil de Comunicação é uma instituição da democracia brasileira: pública, inclusiva, plural e cidadã. Criada em 2007 para fortalecer o sistema público de comunicação (...). Os veículos da EBC têm autonomia para definir produção, programação e distribuição de conteúdos. Atualmente, são veiculados conteúdos jornalísticos, educativos, culturais, esportivos e de entretenimento."

Foram obtidas mil e noventa e sete notícias dos sete diferentes canais durante o período de abril de 2017 a setembro de 2017. Na tabela 4 é ilustrada a distribuição das notícias pelos canais. O conteúdo das notícias obtidas foi organizado em três formatos. O primeiro em arquivo texto, codificados em formato *ASC II plain text*[8]. Esse formato foi gerado para facilitar a manipulação do conteúdo pelos procedimentos de pré-processamento de texto. O segundo formato, apesar também ser um formato texto, segue o padrão *comma separated value (csv)*, o que facilita a utilização em planilhas eletrônicas. O terceiro formato foi construído em uma ferramenta de banco de dados relacional para ser integrado ao protótipo de portal de notícias.

Tabela 4 – Organização do *corpus* EBC de notícias

Canal da notícia	**Total de notícias**
Cidadania	96
Cultura	124
Educação	257
Esporte	137
Infantil	14
Notícias	344
Tecnologia	125

Fonte: José Luiz Maturana Pagnossim, 2018

Na figura 21 é ilustrado um exemplo da exibição de uma notícia pelo portal da EBC[9]. Neste exemplo é possível observar que a notícia pertence ao canal de cultura, além de outros atributos também utilizados por este trabalho, como o título da notícia, a data de publicação e o conteúdo textual da notícia.

Na figura 22 é apresentada a mesma notícia citada anteriormente, mas neste caso organizada em formato de arquivo texto.

O formato *csv*, que facilita a visualização em formato tabular, é apresentado aberto em planilha eletrônica, conforme ilustrado na figura 23.

4.2 O experimento

Dentre os formatos de experimentos comummente aplicados em *SR*, destacam-se: o experimento *offline*, que é realizado usando um conjunto de dados pré-coletados de usuários

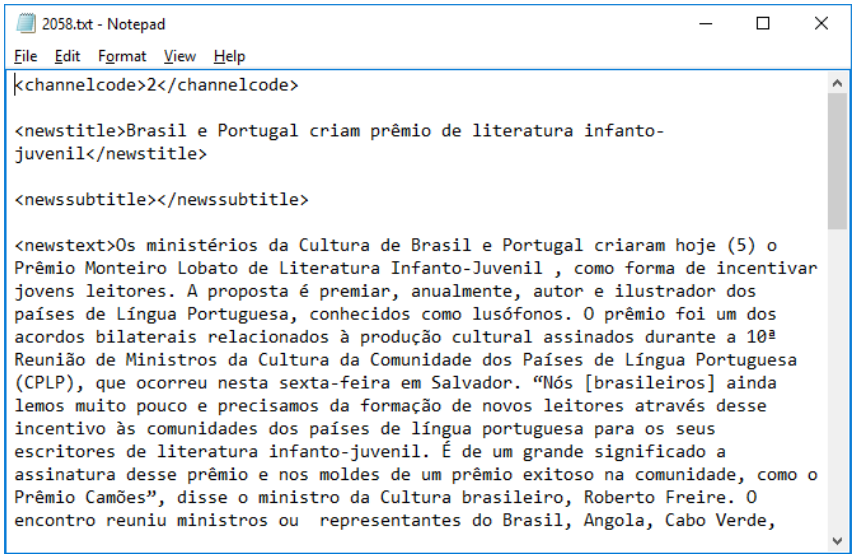
[8] ASC (American Standard Code for Information Interchange).

[9] http://agenciabrasil.ebc.com.br/cultura/noticia/2017-05/brasil-e-portugal-criam-premio-de-literatura-infanto-juvenil

Figura 21 – Exemplo de uma notícia no portal da EBC

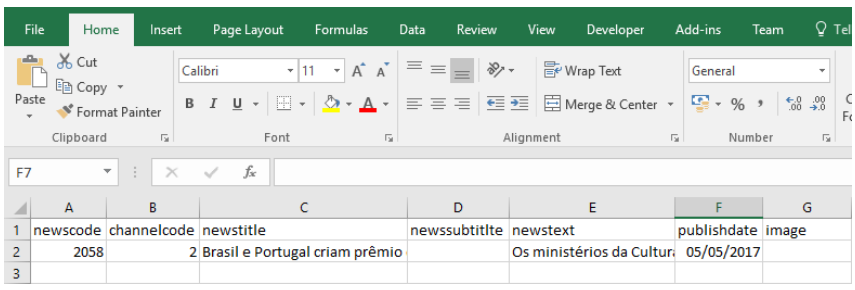
Fonte: site da EBC na internet

Figura 22 – Conteúdo da notícia no formato texto



Fonte: José Luiz Maturana Pagnossim, 2018

Figura 23 – Conteúdo da notícia em formato *c*sv exibido em uma planilha



Fonte: José Luiz Maturana Pagnossim, 2018

selecionando ou classificando itens (GUNAWARDANA, 2015); o estudo do usuário, que é conduzido por meio de um conjunto de cenários de teste e estimulando os usuários a executarem tarefas de interação com o sistema, enquanto isso, seus comportamentos são observados e coletados (GUNAWARDANA, 2015); e o experimento *online* que segundo Gunawardana (2015), "a experiência que fornece a evidência mais forte quanto ao verdadeiro valor do sistema é uma avaliação *online*, em que o sistema é usado por usuários que executam tarefas reais".

O experimento *online* foi considerado o método mais adequado às necessidades deste trabalho, pois possibilitou de forma mais intuitiva a interação do usuário com o sistema de recomendação, proporcionando a obtenção de indicadores gerados durante a navegação do usuário pelo sistema. O conceito de experimento *online* está respaldado no formato como o experimento foi disponibilizado aos usuários e utilizado pelos mesmos, possibilitando coletar indicadores de navegação dos participantes pelo sistema por meio de uma experiência real. Vale ressaltar que o *corpus* elaborado foi um recorte de notícias obtido em um período determinado de tempo, sendo portanto o conteúdo das notícias considerado estático e não online.

O objetivo principal do experimento foi coletar dados de navegação provenientes da interação de usuários com um protótipo de portal de notícias, usar esses dados para gerar uma avaliação sobre o sistema e apresentar um protocolo para reutilização dos dados em trabalhos futuros e pesquisas correlatas. Visto que o experimento envolveu a participação de seres humanos, um projeto foi encaminhado ao comitê de ética sob o título "Avaliação de sistema de recomendação de notícias por meio de interações de usuários" e aprovado por este comitê com a identificação "CAAE 68557417.4.0000.5390".

A navegação dos usuários pelo sistema permitiu a valoração dos seguintes indicadores para avaliação desta pesquisa: quantidade de notícias lidas; quantidade de "curtidas" das notícias; quantidade de aceites das recomendações; nota das recomendações; e as avaliações dos usuários sobre a surpresa positiva causada pela recomendação (serendipidade). Os indicadores calculados a partir do uso do sistema pelos usuários serviram de base para comparação de diferentes critérios e estratégias de recomendação usadas no sistema.

Além das informações já apresentadas até o momento sobre a organização do experimento, outras definições foram estabelecidas como: os perfis dos participantes; a organização das sessões do experimento; o tempo de participação das pessoas no experi-

mento; o protótipo do portal de notícias; a definição dos procedimentos de navegação; e o protocolo para disponibilização dos dados. Todos estes aspectos são apresentados a seguir.

4.2.1 Perfis das pessoas convidadas para participarem do experimento

Foram convidados para participar do experimento quatro grupos diferentes de pessoas, com objetivo de diversificar as interações durante as sessões do experimento.

- Grupo A - estudantes de graduação do curso de bacharelado de sistemas de informação da EACH/USP.
- Grupo B - estudantes de pós-graduação do curso de mestrado em sistemas de informação da EACH/USP.
- Grupo C - estudantes de graduação dos cursos de bacharelado em ciência da computação e tecnologia em análise e desenvolvimento de sistemas da Faculdade Carlos Drummond de Andrade (campus Tatuapé, São Paulo-SP).
- Grupo D - pessoas das áreas de tecnologia da informação e educação que fazem parte do convívio profissional do pesquisador.

As pessoas foram convidadas pelo pesquisador e pela orientadora desta pesquisa, por meio de uma mensagem eletrônica formal contendo os principais objetivos do experimento e endereço na internet para acesso ao sistema.

4.2.2 Organização das sessões do experimento

A partir da definição dos quatro perfis de participantes, descritos na seção anterior, formou-se uma base de pessoas a serem convidadas para o experimento. A estratégia utilizada para melhor distribuição dos convidados de acordo com os perfis, foi fazer convites separados por sessão para diferentes sub-grupos[10] pertencentes aos quatro grupos definidos nos perfis, com exceção da sessão 1 que tinha um objetivo particular, explicado a seguir, assim como os demais objetivos específicos das outras sessões. Vale ressaltar que um participante pode participar mais de uma vez em uma ou mais sessões do experimento, por este motivo são contabilizadas quantidades de participações e não de participantes.

[10] Não foi feito um controle sistemático para controlar se a participação de uma pessoa ocorreu na mesma sessão a qual ela foi convidada ou em uma sessão posterior.

- Sessão 1 - testes do sistema de recomendação. Para testar o sistema foi selecionado um sub-grupo do grupo B especializado em sistemas de recomendação e mineração de dados. Este sub-grupo contribuiu com a identificação de defeitos e sugestões de melhorias. Ao final desta sessão, uma nova versão do protótipo foi disponibilizada com as correções dos defeitos encontrados. Os dados de navegação, gerados pelos participantes, foram descartados para efeito de apuração dos resultados e este sub-grupo de participantes não foi convidado para participar das sessões seguintes.
- Sessão 2 - estabelecimento de uma base inicial de navegação para comparação com as próximas sessões do experimento e análise dos efeitos de *cold start* de itens e *cold start* de usuários. Para esta sessão foi convidado um sub-grupo de pessoas selecionado a partir dos quatro grupos definidos na seção 4.2.1. Ao final do período estabelecido para essa sessão (uma semana), foram registradas trinta participações. Os dados gerados nesta sessão foram mantidos para uso na sessão seguinte.
- Sessão 3 - comparação entre sessões em um cenário com dados já registrados pelo sistema de recomendação. Para esta sessão foi convidado um sub-grupo de pessoas selecionado a partir dos quatro grupos definidos na seção 4.2.1 (exceto pessoas convidadas anteriormente). Foi definido o mesmo prazo da sessão anterior (uma semana), mas após três dias da abertura desta sessão, o número de participações já havia superado a sessão anterior. Neste momento foi guardada uma cópia da base de dados para comparação com a sessão anterior e as pessoas que acessaram o sistema a partir deste momento foram contabilizadas na sessão seguinte. Foram registradas trinta e duas participações e apesar do número de participações ser maior em relação à sessão anterior, o número de leitura foi ligeiramente inferior nesta sessão, o que proporcionou uma análise comparativa igualitária entre as duas sessões.
- Sessão 4 - nesta sessão participaram pessoas que acessaram o experimento após o encerramento da sessão 3, ou seja, eram pessoas que foram convidadas para participarem da sessão 2 ou da sessão 3, mas não tinham conseguido acessar o experimento dentro dos prazos estabelecidos nas respectivas sessões. Os dados desta sessão foram separados para serem comparados com as sessões 2 e 3. Apesar do número de participações ter sido maior nesta última sessão (quarenta e nove), foram gerados indicadores, tabelas e gráficos, a partir dos dados coletados, de forma a relativizar a análise dos resultados.

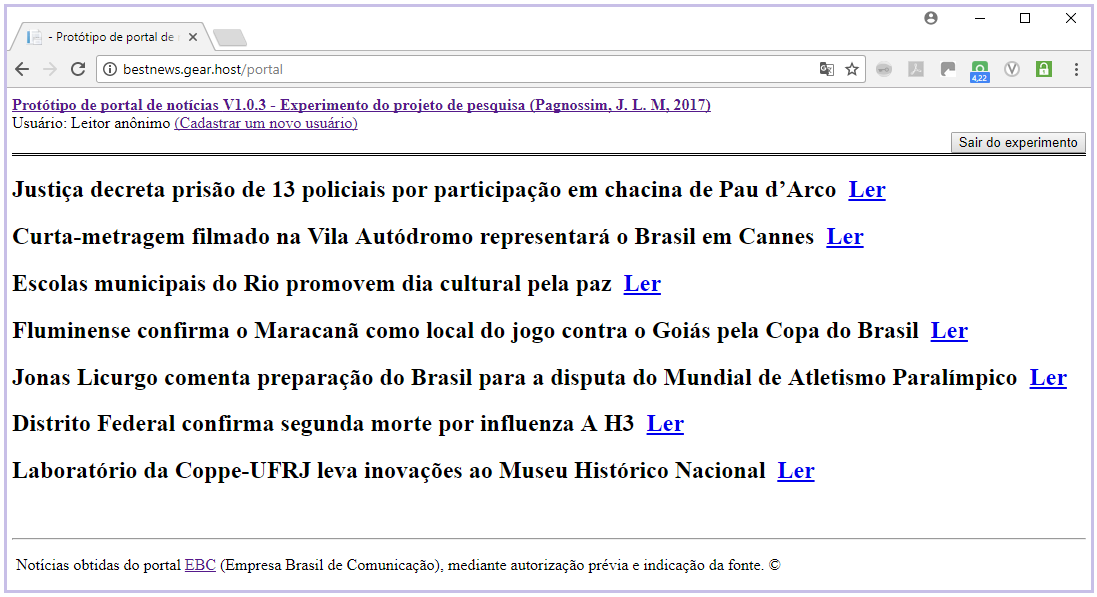
4.2.3 Tempo de participação no experimento

Embora não houvesse obrigatoriedade de tempos mínimo e máximo para permanência do participante no ambiente do experimento, na mensagem formal enviada aos convidados foi sugerido que o sistema fosse usado por pelo menos dez minutos. Ao final do experimento, foi registrado um tempo médio (por participação) de 8 minutos e 14 segundos com um desvio desvio padrão de 9 minutos e 27 segundos.

4.2.4 O protótipo do portal de notícias

Foi construído um protótipo de portal de notícias para atender especificamente aos objetivos deste projeto. Este protótipo constituiu uma plataforma de apoio para provar o conceito da abordagem apresentada nesta pesquisa. Na figura 24 é ilustrada uma tela exibida pelo sistema para interação inicial do participante com o experimento.

Figura 24 – Protótipo do portal de notícias: tela de interação inicial do participante



Fonte: José Luiz Maturana Pagnossim, 2018

4.2.5 Definição dos procedimentos de navegação

Ao acessar o endereço do experimento na internet, o sistema direcionou o convidado à tela inicial com as instruções do experimento, conforme apêndice A – Página de Instruções Iniciais. Ao avançar nesta página, foi apresentada uma outra tela contendo o termo de

consentimento livre e esclarecido (TCLE), conforme apêndice B. O participante tinha que concordar com os termos para prosseguir e na sequência o sistema exibia uma tela para o participante escolher o modo de navegação durante o uso do experimento, podendo optar: pela navegação como um leitor anônimo; cadastrar um novo usuário; ou acessar usando um usuário já cadastrado.

No apêndice C é apresentada a tela de modo de navegação, em que é possível observar as opções que o participante tem para navegar pelo experimento. Ao escolher um dos modos de navegação, o sistema direcionava o participante à tela inicial do portal de notícias do experimento, já descrita na seção 4.2.4. Esta tela, apresentava sete notícias, sendo uma de cada canal. Dentro de cada canal a escolha da notícia a ser exibida foi aleatória, de forma que não houvesse influência dessa recomendação inicial nos indicadores do sistema.

Ao escolher uma notícia, a partir da tela inicial do portal, o sistema exibia uma tela com o conteúdo da notícia e os ícones para curtir a notícia ou ainda avaliar a recomendação em termos de serendipidade e nota da recomendação. Esta tela está ilustrada na figura 25.

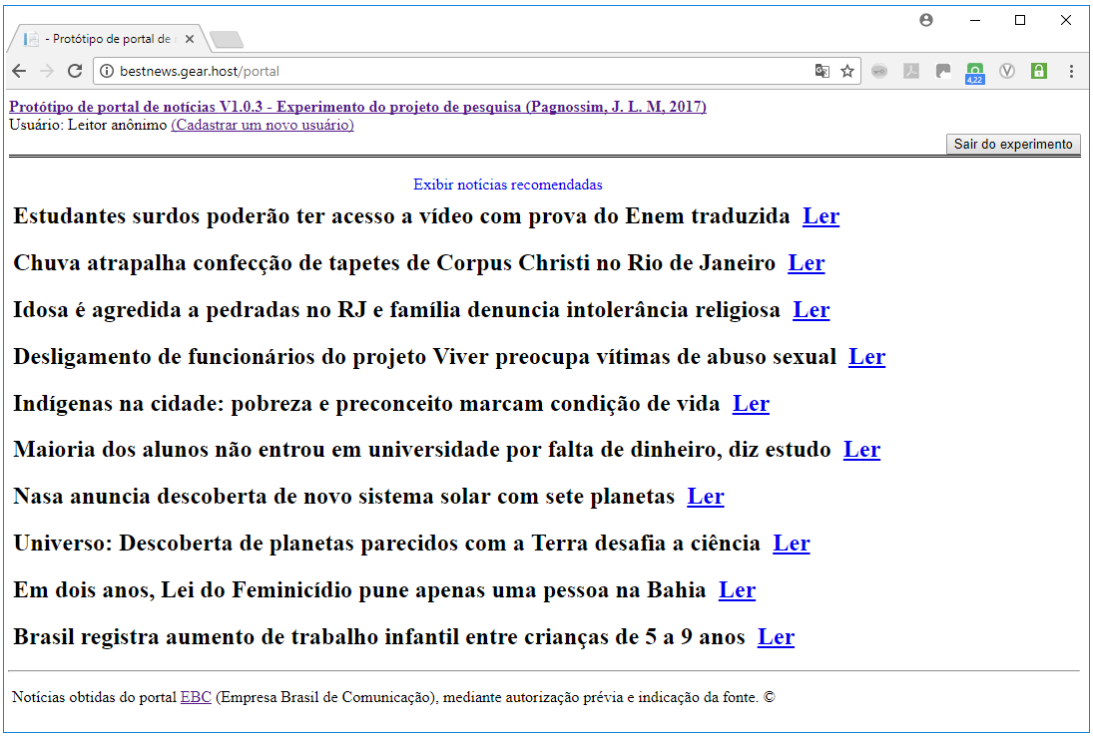
Figura 25 – Protótipo do portal de notícias: tela para leitura e interação

Fonte: José Luiz Maturana Pagnossim, 2018

A partir desta tela, além dos ícones para interagir com a notícia e com a recomendação, o participante podia solicitar para que o sistema exibisse notícias recomendadas,

ao escolher essa opção, o sistema listava na parte inferior da mesma tela, as notícias recomendadas pelo sistema. Esta parte da tela está ilustrada na figura 26.

Figura 26 – Protótipo do portal de notícias: lista de notícias recomendadas



Fonte: José Luiz Maturana Pagnossim, 2018

4.2.6 Protocolo para disponibilização dos dados

Os dados utilizados e gerados por este trabalho foram mantidos pelos pesquisadores em ambiente privado, já que envolve a disponibilização de notícias reais publicadas na internet pelo portal EBC. Para que os dados sejam adequados para acesso público e possam ser reutilizados por outros estudos, alguns procedimentos devem ser seguidos. Estes procedimentos estão detalhados no apêndice D.

4.3 Resultados e análises

Esta seção apresenta os resultados extraídos a partir da execução do experimento, bem como as análises dos resultados em relação ao experimento como um todo e também comparações entre as sessões que o compuseram.

Para demonstrar tais resultados e análises, está seção está organizada de forma a apresentar: o perfil dos participantes e o tempo de utilização durante do experimento; os resultados referentes às interações sobre a notícia, definidos como indicadores de leitura e curtida; os resultados relacionados à avaliação da recomendação que envolvem o indicador de aceite de recomendação, a nota da recomendação e o indicador de serendipidade; as estratégias de recomendação com base nos treze critérios de recuperação de notícias e utilizando a RBC; e a eficácia da ($\mathcal{MURR}$). Após a validação dos resultados do experimento é apresentada uma avaliação da arquitetura de recomendação de notícias sob o ponto de vista do atendimento aos requisitos funcionais exigidos por um sistema de recomendação. Vale ressaltar que os dados da sessão 1 não foram contabilizados nos resultados, conforme motivos descritos na seção 4.2.2.

4.3.1 Análise dos resultados do experimento

Foram estabelecidas convenções para que fosse possível relacionar os resultados gerados em decorrência do experimento com as propriedades e formas de avaliação fundamentadas neste trabalho, com base nas definições de (RICCI et al., 2011). Estas convenções são especificadas no quadro 5.

Quadro 5 – Formas de avaliação do sistema de recomendação

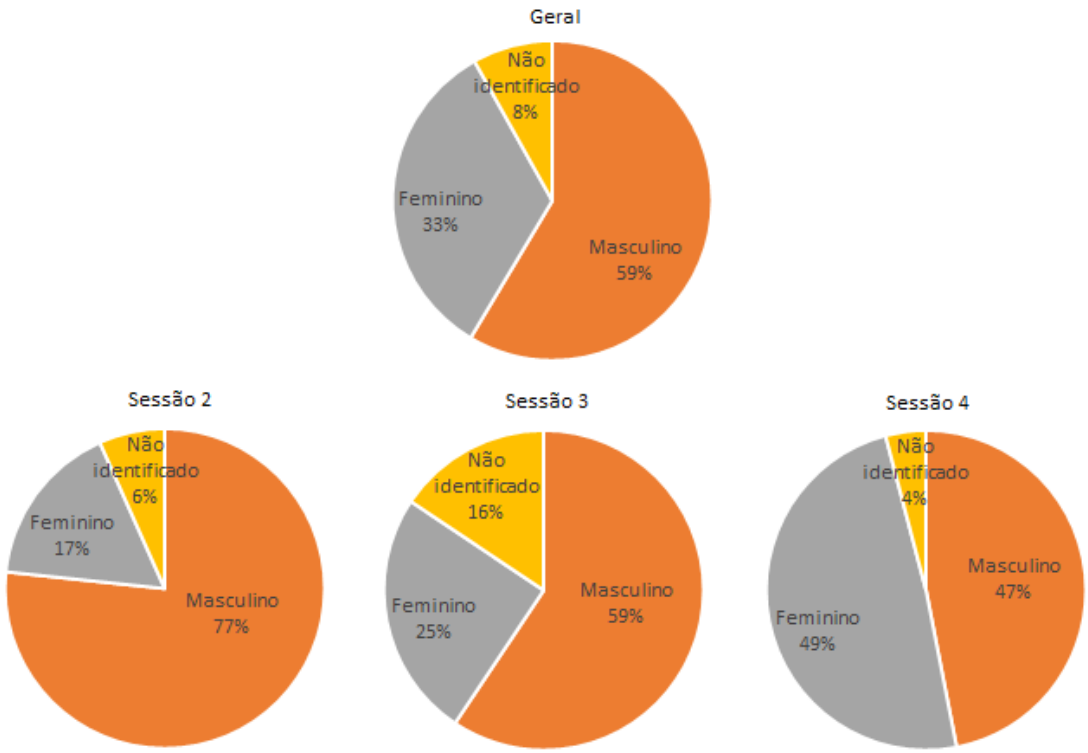
Formas de avaliação
$\{AV_1\}$ Estudo da preferência do usuário
$\{AV_2\}$ Ranqueamento de item (implementado por meio da $MURR$)
$\{AV_3\}$ Cobertura do item no espaço
$\{AV_4\}$ Cobertura do usuário no espaço
$\{AV_5\}$ *Cold start* de item
$\{AV_6\}$ Novidade
$\{AV_7\}$ Diversidade
$\{AV_8\}$ Utilidade
$\{AV_9\}$ Serendipidade
$\{AV_{10}\}$ Similaridade de conteúdo
$\{AV_{11}\}$ Similaridade entre usuários
$\{AV_{12}\}$ Similaridade intragrupos
$\{AV_{13}\}$ Similaridade intergrupos
$\{AV_{14}\}$ Popularidade
$\{AV_{15}\}$ Estratégias de recomendação

Fonte: José Luiz Maturana Pagnossim, 2018

Perfil dos participantes e tempo de utilização

Foram contabilizadas 111 participações no experimento, sendo 30 na sessão 2, 32 na sessão 3 e 49 na sessão 4. A título de conhecimento, são apresentados gráficos que demonstram o perfil do público participante do experimento. Estes dados foram obtidos durante a identificação do participante na tela inicial do experimento. As informações fornecidas não foram usadas pelo sistema para realizar recomendações, mas somente para conhecer o perfil do público. Na figura 27 estão ilustrados os gráficos com as participações de acordo com o gênero. Já na figura 28 estão ilustrados os gráficos de participações de acordo com a faixa etária.

Figura 27 – Participações por gênero

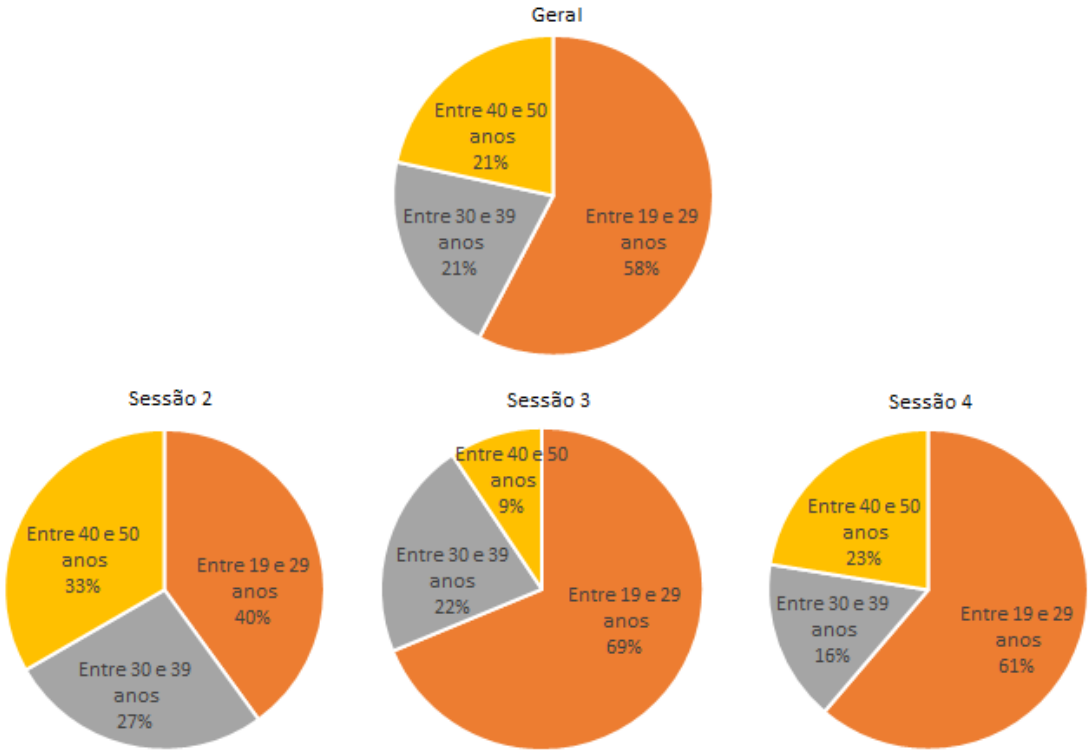


Fonte: José Luiz Maturana Pagnossim, 2018

A informação de gênero foi definida pelo sistema como opcional durante a fase de preenchimento dos dados do participante, podendo o participante deixar em branco ou escrever o gênero que melhor lhe definisse. Neste contexto, os resultados relacionados ao gênero, apresentaram os seguintes registros preenchidos pelos participantes: “Feminino”, “Masculino” e quando não preenchido, foi rotulado pelo sistema como “Não identificado”.

Ao analisar os gráficos de perfil do público participante foi possível observar uma participação majoritária de jovens do sexo masculino, o que se justifica pelo contexto em

Figura 28 – Participações por faixa etária



Fonte: José Luiz Maturana Pagnossim, 2018

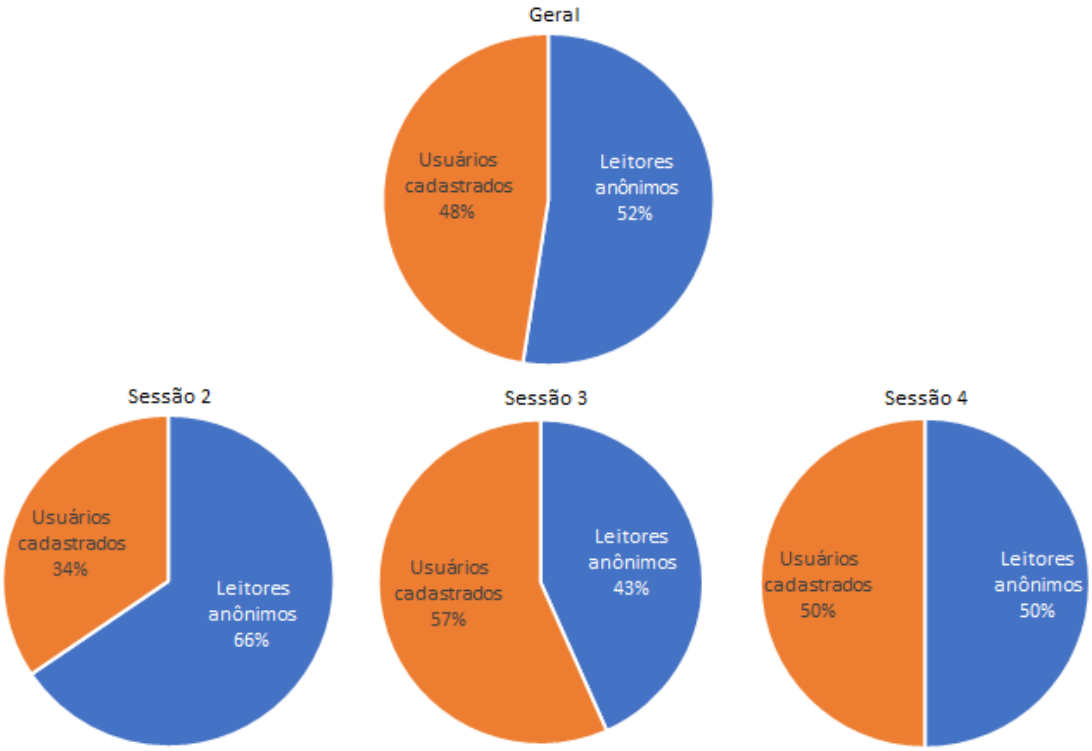
que está situada a população convidada, que na sua maioria era composta por estudantes de graduação e pós-graduação de cursos de sistemas de informação e afins.

Após a identificação do participante, era questionado se o mesmo desejava navegar pelo sistema de forma anônima ou se preferia se cadastrar ($\{AV_4\}$). Na figura 29 é ilustrada a proporção desta opção entre os participantes.

Neste gráfico é possível observar que o modo de navegação escolhido pela maioria foi o anônimo, ou seja, 52% dos participantes não receberam recomendações com base nos critérios relacionados ao usuário. Já os que optaram por se cadastrar (48%) receberam as recomendações relacionadas a filtro colaborativo e pelo histórico de navegação usuário.

Com relação ao tempo de utilização do sistema pelos participantes durante as sessões do experimento, não foi definida uma obrigatoriedade de tempos mínimo e máximo, mas foi recomendada uma permanência de 10 minutos por participante. Ao final do experimento, o tempo médio de permanência por participação foi de 8 minutos e 14 segundos, valor próximo do recomendado, embora tenha sido notado que alguns participantes utilizaram o sistema por pouco tempo, por exemplo o menor tempo de participação foi de 12 segundos, enquanto o maior tempo foi de 1 hora e 1 minuto. Esta discrepância refletiu no desvio

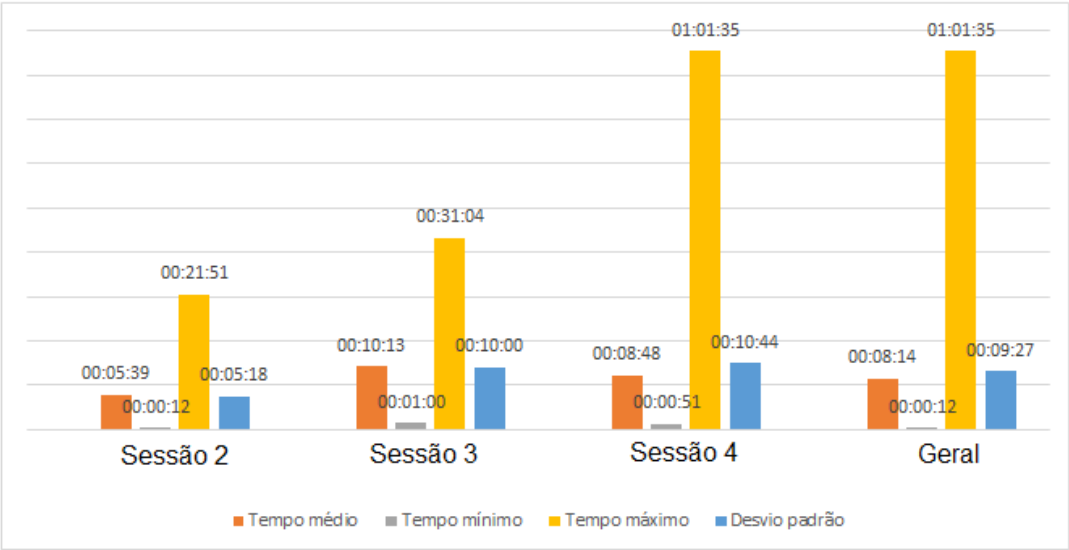
Figura 29 – Participação por opção de modo de navegação



Fonte: José Luiz Maturana Pagnossim, 2018

padrão, que foi de 9 minutos e 27 segundos. Na figura 30 estão ilustrados os tempos de participação (formato *hh:mm:ss*[11]) ao longo do experimento.

Figura 30 – Tempo de participação por sessão



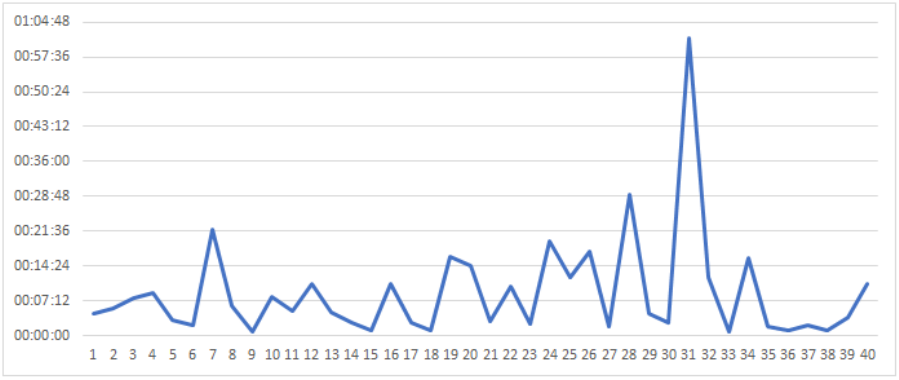
Fonte: José Luiz Maturana Pagnossim, 2018

[11] hh(hora); mm(minuto); ss(segundo)

Os gráficos de tempo evidenciaram uma discrepância entre os tempos mínimo e máximo de utilização ao longo das três sessões. É possível observar que o tempo médio registrado sessão 3 foi 80% maior que na sessão anterior. Já na quarta sessão, o tempo ficou 7% maior do que a média geral.

Um dado que chamou atenção foi o valor do tempo máximo da sessão 4 (causado pela permanência de um único participante), o que tenderia a aumentar o valor da média, mas isto não ocorreu devido aos demais participantes terem registrado tempos próximos da média. O gráfico ilustrado na figura 31 auxilia nesta análise.

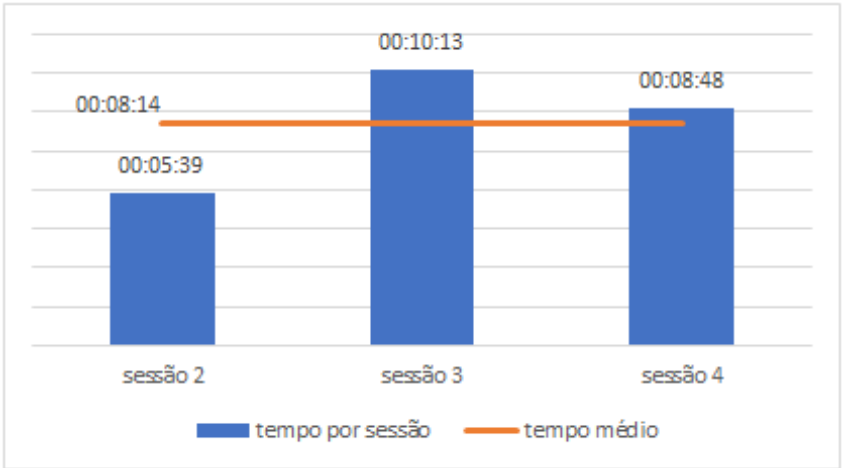
Figura 31 – Tempo por participante na sessão 4



Fonte: José Luiz Maturana Pagnossim, 2018

Na figura 32 está ilustrada a comparação entre os tempos médios de cada sessão em relação ao tempo médio geral, em que é possível observar que as sessões 3 e 4 ficaram com os tempos acima da média geral.

Figura 32 – Tempo de participação



Fonte: José Luiz Maturana Pagnossim, 2018

Avaliações sobre as notícias

Estas avaliações registraram indicadores exclusivamente associados às notícias, como a quantidade de leitura que uma notícia obteve, ou ainda a quantidade de curtidas de uma notícia. Estes indicadores foram registrados no experimento por meio da navegação dos participantes.

O indicador de leitura

O indicador de leitura foi registrado no experimento por meio de um atributo associado à notícia, calculado no momento em que o participante clicava em uma notícia. Esta ação por parte do participante, disparava um evento no sistema que incrementava um contador de leituras para a notícia selecionada ($\{AV_{14}\}$). O sistema registrou em cada sessão do experimento, além do número total de leituras, outros indicadores derivados deste, como a quantidade de notícias distintas lidas e o percentual de exploração do *corpus*, calculado a partir da divisão entre a quantidade de notícias distintas lidas pela quantidade total de notícias do *corpus*. Um outro indicador foi gerado para analisar a média de leitura por participação. O objetivo da criação deste indicador foi relativizar a análise, considerando que a quantidade de participações foi diferente entre as sessões do experimento. Este indicador está formalizado na equação 5.

$$MediaLeituraParticipacao = \frac{quantidadeTotalLeituras}{quantidadeParticipacoes} \tag{5}$$

Os indicadores calculados a partir das leituras sobre as notícias, estão ilustrados na tabela 5.

Tabela 5 – Indicadores de leitura

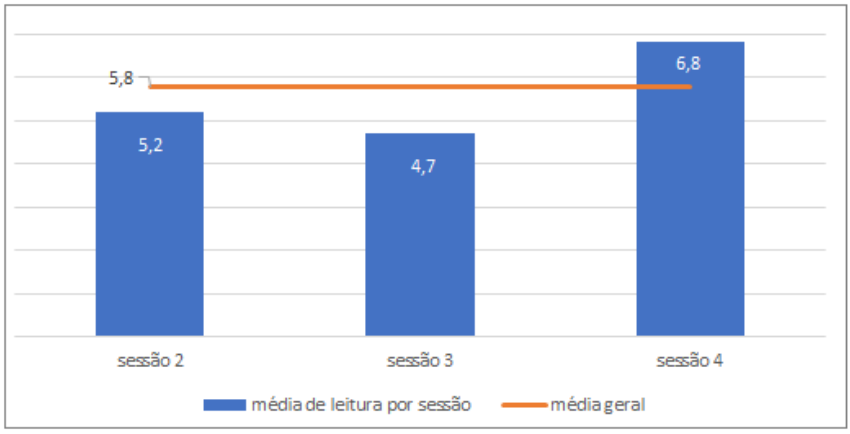
Indicador	**Sessão 2**	**Sessão 3**	**Sessão 4**	**Geral**
Quantidade total de leituras ($\{AV_{14}\}$)	156	150	334	640
Quantidade de notícias distintas lidas	79	95	190	264
Exploração do corpus ($\{AV_3\}$)	7%	9%	17%	24%
Quantidade de participações	30	32	49	111
Média de leituras por participação	5,2	4,7	6,8	5,8

Fonte: José Luiz Maturana Pagnossim, 2018

Estes resultados demonstram que a sessão 2 apresentou uma quantidade total de leituras maior que da sessão 3, consequentemente o indicador que armazenou a média de leituras por participação também foi maior na sessão 2 em relação à sessão 3. Por outro lado, os participantes da sessão 3, por terem lido uma quantidade de notícias distintas maior que da sessão 2, tiveram um percentual maior de exploração do *corpus* ($\{AV_3\}$). Essas informações permitem concluir que os participantes da sessão 2 leram mais vezes as mesmas notícias se comparados com os participantes da sessão 3, que por sua vez exploraram de forma mais distinta as notícias do *corpus*.

Para que fosse possível comparar as três sessões do experimento, foi gerado o indicador de média de leituras por participação. Comparando todas as sessões, notou-se que a sessão 4 apresentou a maior média, esta análise comparativa está ilustrada no gráfico da figura 33.

Figura 33 – Média de leitura por sessão



Fonte: José Luiz Maturana Pagnossim, 2018

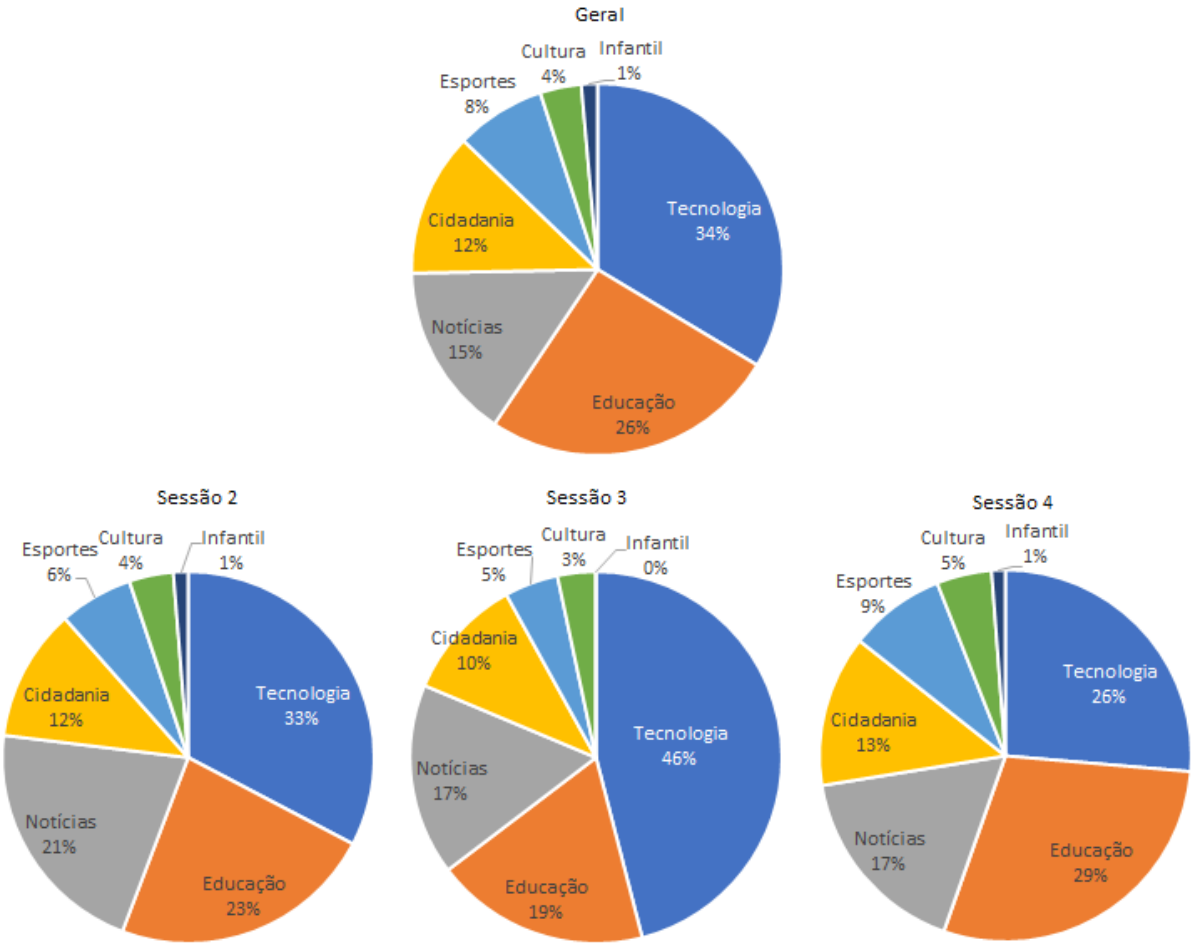
Por meio deste indicador foi possível observar que a sessão 4 alcançou uma média de leitura maior que as sessões anteriores. Este fenômeno pode estar ligado ao fato do algoritmo de recomendação ter reunido mais elementos históricos para sugerir itens aos participantes, resultando em um maior interesse dos participantes pelo sistema. Sugere-se também que as duas primeiras sessões ainda estavam sob o efeito inicial dos *SR* conhecido como *cold start* de item ($\{AV_5\}$).

O percentual de exploração do *corpus* ($\{AV_3\}$) ao final do experimento alcançou 24% de notícias distintas lidas, ou seja, foram lidas 264 notícias distintas de um total de 1097 notícias existentes no *corpus*. Sobre os 76% das notícias não exploradas, duas possibilidades podem ser levantadas. A primeira é que o tempo de utilização e o número

de participantes foram pequenos se comparados com tamanho do *corpus*. A segunda é que os participantes leram um grupo de notícias em comum, aceitando, na maior parte dos casos, recomendações por popularidade e similaridade no lugar das recomendações que estimulavam diversidade e novidade.

Uma outra visão derivada do indicador de leitura apresenta os canais mais lidos ($\{AV_{14}\}$), conforme demonstrada na figura 34.

Figura 34 – Canais mais lidos



Fonte: José Luiz Maturana Pagnossim, 2018

Os gráficos demonstram que o canal mais lido foi o de tecnologia, resultado esperado devido ao perfil do público convidado para participar do experimento. Apesar de esperado esse resultado no âmbito geral, houve uma preferência pelo canal educação na sessão 4. Uma possibilidade deste efeito ter ocorrido é pelo fato de parte dos convidados serem professores. Apesar da distribuição e dos convites terem sido feitos por igual, pode ter ocorrido uma maior incidência deste perfil nesta sessão. Essa possibilidade não tem como ser confirmada, já que não o experimento não solicitava e registrava a profissão do participante.

O indicador de curtida

O indicador de curtida foi registrado no experimento por meio de um atributo associado à notícia, calculado no momento em que o participante clicava no botão curtir. Esta ação, disparava um evento no sistema que incrementava um contador de curtidas para a notícia selecionada ($\{AV_{14}\}$). O sistema registrou em cada sessão do experimento, além do número total de curtidas, a quantidade de notícias distintas curtidas e considerando que a quantidade de participações foi diferente entre as sessões do experimento, um outro indicador foi gerado para registrar a média de curtida por sessão, conforme formalizado na equação 6.

$$MediaCurtidaParticipacao = \frac{quantidadeTotalCurtidas}{quantidadeParticipacoes} \tag{6}$$

O objetivo da criação deste último indicador foi relativizar a análise das informações sobre curtida entre as diferentes sessões do experimento. Os indicadores calculados a partir das curtidas, estão ilustrados na tabela 6.

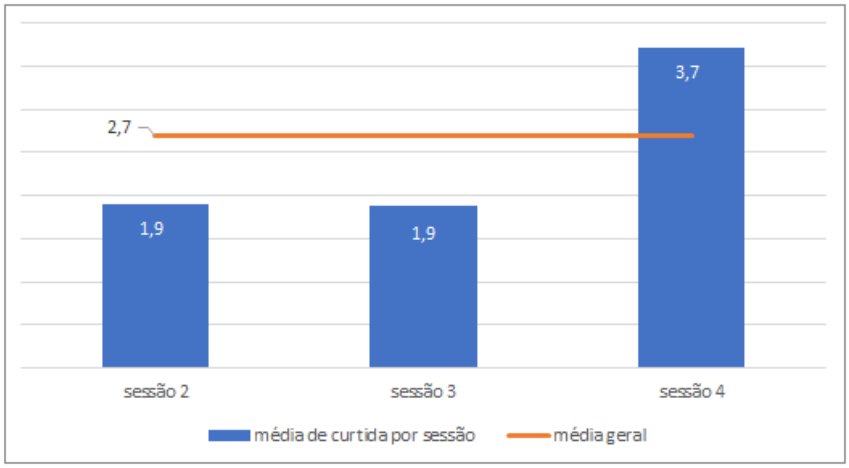
Tabela 6 – Indicador de curtida

Indicador	**Sessão 2**	**Sessão 3**	**Sessão 4**	**Geral**
Quantidade total de curtidas	57	60	182	299
Quantidade de notícias distintas curtidas	37	42	118	149
Quantidade participações	32	30	49	111
Média de curtidas por participação	1,9	1,9	3,7	2,7

Fonte: José Luiz Maturana Pagnossim, 2018

A quantidade total de curtidas apresentou uma diferença mínima a favor da sessão 3 em relação à sessão 2, assim como já havia sido observado no indicador de leitura, porém a favor da sessão 2 em relação à sessão 3. Estas diferenças podem ser consideradas irrelevantes, o que reforça a suposição feita anteriormente de que as duas primeiras sessões estariam inseridas em um contexto de *cold start* de item ($\{AV_5\}$). Esta possibilidade é reforçada ao analisar o indicador de média de curtidas, que foram idênticos nas sessões 2 e 3, enquanto que na sessão 4, este indicador foi expressivamente maior que os valores das sessões anteriores. No gráfico apresentado na figura 35 é demonstrada de forma mais clara a diferença das curtidas.

Figura 35 – Média de curtida por sessão



Fonte: José Luiz Maturana Pagnossim, 2018

O fato das médias de leitura e curtida da sessão 4 terem sido maiores em relação às as sessões anteriores, reforça a possibilidade de melhoria do sistema após o período de uso inicial.

Uma outra visão derivada do indicador de curtida apresentou os canais mais curtidos entre as sessões do experimento, conforme demonstrada na figura 36.

Os resultados contidos nestes gráficos demonstram que o canal mais curtido foi o tecnologia. Em linhas gerais, estes resultados acompanharam a preferência dos participantes já demonstrada nos indicadores de leitura.

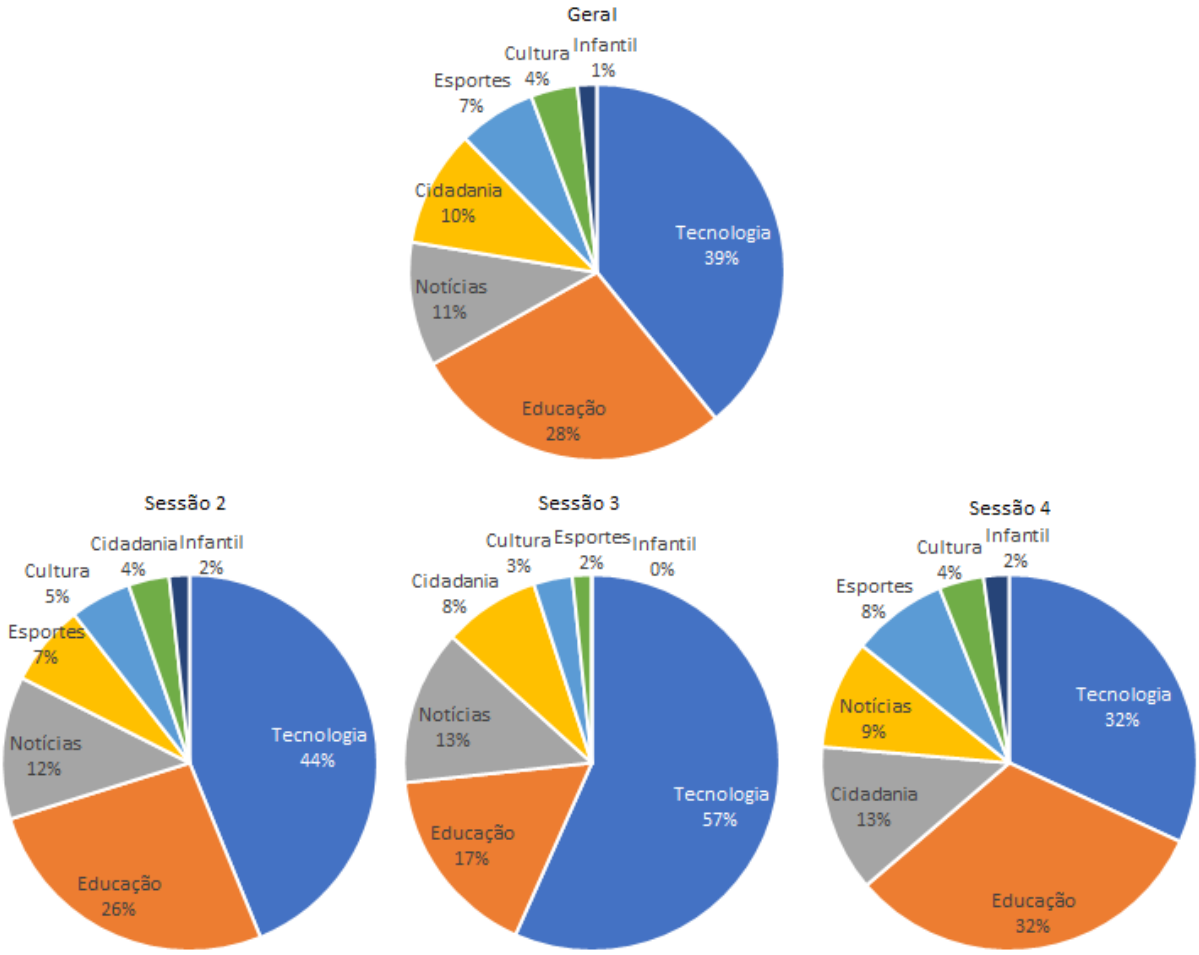
Avaliações sobre as recomendações

As avaliações dos participantes sobre as recomendações de notícias foram organizadas em três indicadores: aceite de recomendação; nota da recomendação em escala 1-5 estrelas; e serendipidade. Estes indicadores e os resultados que eles representam são discutidos nesta seção.

O indicador de aceite de recomendação

O indicador de aceite está relacionado à escolha que o participante faz diante de uma lista de notícias recomendadas pelo sistema. Este indicador é usado para gerar diferentes visualizações dos dados e está associado às avaliações: $\{AV_1\}$, $\{AV_2\}$, $\{AV_5\}$, $\{AV_9\}$, $\{AV_{10}\}$, $\{AV_{11}\}$, $\{AV_{12}\}$, $\{AV_{13}\}$, $\{AV_{14}\}$ e $\{AV_{15}\}$.

Figura 36 – Canais mais curtidos



Fonte: José Luiz Maturana Pagnossim, 2018

O indicador de aceite é definido na equação 7, em que *(NO, ND)* é uma uma dupla formada pela notícia de origem e uma de destino.

$$IndicadorAceite = \frac{quantidadeAceites(NO,\ ND)}{quantidadeRecomendacoes(NO,\ ND)} \tag{7}$$

Por exemplo, se uma dupla ($\mathcal{NO}$, $\mathcal{ND}$) é recomendada 10 vezes em momentos distintos sendo registrado 1 aceite para esta dupla, o valor do indicador é de 10%. Considerando pelo menos 1 aceite para uma determinada dupla, se o indicador ficar próximo de 1% significa que foram feitas muitas sugestões e poucos aceites, representando um resultado ruim para o indicador. Por outro lado, quanto mais próximo de 100%, mais esta recomendação foi aceita, representando um bom resultado para o indicador.

Foram recomendadas 5.444 notícias e registrados 433 aceites considerando as três sessões do experimento. Na tabela 7 está listada, para cada indicador de aceite registrado pelo sistema, a quantidade de aceites em valor absoluto e a representatividade desta

quantidade perante todos os aceites da sessão. As linhas com indicador entre 50% e 100% estão destacadas para facilitar a análise dos gráficos apresentados após as tabelas.

Tabela 7 – Indicador de aceite de recomendação na sessão 2

Indicador	**Quantidade aceites**	**Representatividade**
100%	**49**	**44,5%**
50%	**20**	**18,2%**
40%	4	3,6%
33%	8	7,3%
29%	2	1,8%
25%	11	10,0%
20%	3	2,7%
17%	8	7,3%
14%	1	0,9%
13%	4	3,6%
Total	110	100%

Fonte: José Luiz Maturana Pagnossim, 2018

Destacando a primeira linha da tabela 7, observa-se que 49 aceites atingiram o valor de 100% no indicador, o que significa que todas as vezes que essas notícias foram recomendadas elas foram aceitas. Selecionando a segunda linha para analisar, observa-se que 20 recomendações foram aceitas em 50% das vezes em que foram recomendadas, por exemplo, uma dupla pode ter sido recomendada 10 vezes e aceita 5 vezes, enquanto uma outra dupla pode ter sido recomendada 2 vezes e aceita 1 vez, ou seja, ambas atingiram registraram um indicador de 50%. Este mesmo raciocínio deve ser feito para todas as linhas da tabela. Na tabela 8 são demonstrados os resultados detalhados da sessão 3 do experimento e na tabela 9 são apresentados os resultados da sessão 4.

Uma outra forma de analisar informações de aceite é verificar se o indicador de aceite está mais próximo de 1% ou de 100%. No gráfico ilustrado na figura 37 são apresentados os dados de aceites de recomendações em dois grupos: o primeiro com indicador de aceite maior ou igual a 50%; e o segundo com indicador de aceite maior que zero e menor que 50%. Este indicador é definido como indicador agrupado de aceite de recomendação.

Neste gráfico é importante observar se a maior parte dos aceites está posicionada no primeiro ou no segundo grupo, com destaque para a sessão 4 que registrou um indicador agrupado de 75%.

Tabela 8 – Indicador de aceite de recomendação na sessão 3

Indicador	**Quantidade aceites**	**Representatividade**
100%	**49**	**47,6%**
67%	**2**	**1,9%**
50%	**17**	**16,5%**
43%	3	2,9%
38%	3	2,9%
33%	10	9,7%
29%	2	1,9%
25%	7	6,8%
20%	1	1,0%
17%	4	3,9%
14%	1	1,0%
13%	3	2,9%
11%	1	1,0%
Total	103	100%

Fonte: José Luiz Maturana Pagnossim, 2018

Tabela 9 – Indicador de aceite de recomendação na sessão 4

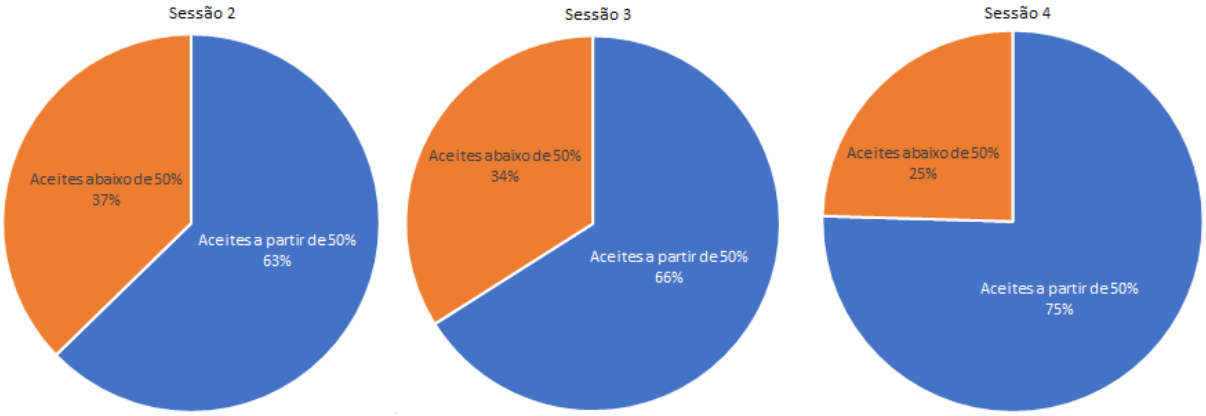
Indicador	**Quantidade aceites**	**Representatividade**
100%	**130**	**59,1%**
67%	**4**	**1,8%**
50%	**32**	**14,5%**
43%	3	1,4%
38%	5	2,3%
33%	19	8,6%
25%	4	1,8%
23%	3	1,4%
20%	7	3,2%
17%	9	4,1%
14%	1	0,5%
11%	1	0,5%
8%	2	1,0%
Total	220	100%

Fonte: José Luiz Maturana Pagnossim, 2018

Na figura 38 é apresentada uma visão comparativa entre as três sessões do experimento, considerando o indicador agrupado entre 50% e 100% e a representatividade para aceite de 100%.

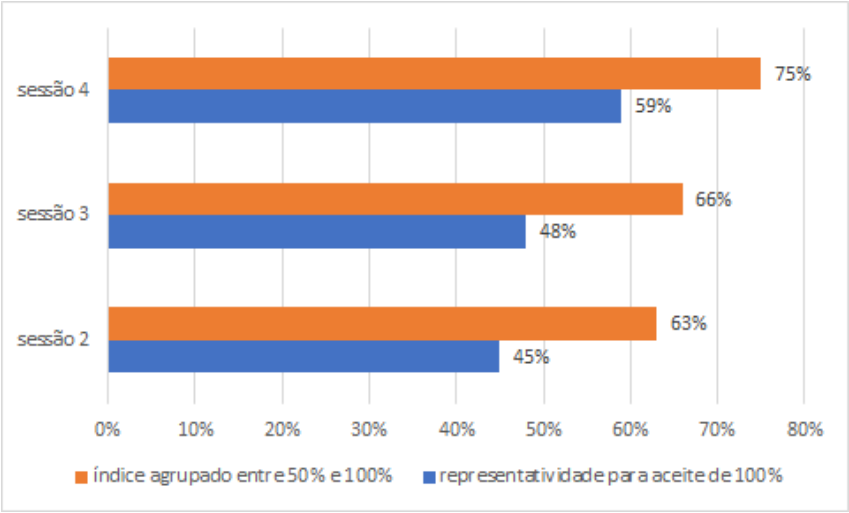
Uma outra visualização a partir do indicador de aceite de recomendação é obtida pelo cruzamento de informações de aceite de recomendação com os critérios de recuperação.

Figura 37 – Indicador agrupado de aceite de recomendação: um comparativo



Fonte: José Luiz Maturana Pagnossim, 2018

Figura 38 – Aceite de recomendação: comparativo entre as sessões do experimento



Fonte: José Luiz Maturana Pagnossim, 2018

As tabelas: 10 (para os dados da sessão 2); 11 (para os dados da sessão 3); e 12 (para os dados da sessão 4) ilustram estas informações.

As comparações apresentadas nestas tabelas servem para mostrar a preferência de aceites dos participantes para cada critério de recuperação da notícia (utilizado pelo algoritmo baseado nos critérios). Nestas tabelas, o caractere # significa quantidade e IAceite significa o índice de aceite de recomendação.

A partir dos dados mostrados nas tabelas 10, 11 e 12 é possível fazer recortes de alguns trechos para auxiliar a análise dos resultados. O primeiro recorte evidencia que os critérios de 9 a 12 foram menos recomendados que os demais, algo esperado, já que os critérios de 9 a 11 dependiam da associação entre $(\mathcal{NO}, \mathcal{ND})$, que neste período de uso

Tabela 10 – Aceite de recomendação por critério - sessão 2

Critério	**sessão 2**		
	#recomend.	**#aceites**	**IAceite**
10 - Maior nota de avaliação $\{AV_{14}\}$	22	4	18%
11 - Maior serendipidade $\{AV_9\}$	8	1	13%
8 - Notícia mais curtida do portal $\{AV_{14}\}$	127	15	12%
6 - Notícia mais lida do portal $\{AV_{14}\}$	122	13	11%
5 - Notícia mais lida do canal $\{AV_{14}\}$	128	13	10%
7 - Notícia mais curtida do canal $\{AV_{14}\}$	127	12	9%
13 - Histórico do usuário $\{AV_1\}$	126	10	8%
1 - Maior similaridade de conteúdo $\{AV_{10}\}$	129	10	8%
3 - Maior similaridade intergrupo $\{AV_{13}\}$	129	9	7%
4 - Aleatória não recomendada $\{AV_6, AV_7\}$	130	8	6%
9 - Mais recomendações aceitas $\{AV_{14}\}$	34	2	6%
12 - Similaridade entre usuários $\{AV_{11}\}$	102	6	6%
2 - Maior similaridade intragrupo $\{AV_{12}\}$	127	7	6%

do sistema ainda estava "frio". Já o critério 12 dependia do participante ter se cadastrado como usuário do sistema.

Fazendo um sub-recorte deste trecho, destacando o critério 11 - Maior serendipidade, nota-se que este critério apresentou valores poucos expressivos, sugerindo que a serendipidade como critério de recomendação é muito particular, ou seja, se algo surpreendeu positivamente uma pessoa, não necessariamente vai surpreender outras pessoas.

Tabela 11 – Aceite de recomendação por critério - sessão 3

Critério	**sessão 3**		
	#recomend.	**#aceites**	**IAceite**
9 - Mais recomendações aceitas $\{AV_{14}\}$	26	4	15%
1 - Maior similaridade de conteúdo $\{AV_{10}\}$	126	17	13%
10 - Maior nota de avaliação $\{AV_{14}\}$	18	2	11%
4 - Aleatória não recomendada SR $\{AV_6, AV_7\}$	131	13	10%
2 - Maior similaridade intragrupo $\{AV_{12}\}$	125	12	10%
8 - Notícia mais curtida do portal $\{AV_{14}\}$	125	12	10%
5 - Notícia mais lida do canal $\{AV_{14}\}$	127	10	8%
7 - Notícia mais curtida do canal $\{AV_{14}\}$	127	10	8%
6 - Notícia mais lida do portal $\{AV_{14}\}$	121	9	7%
3 - Maior similaridade intergrupo $\{AV_{13}\}$	130	8	6%
12 - Similaridade entre usuários $\{AV_{11}\}$	112	3	3%
13 - Histórico do usuário $\{AV_1\}$	125	3	2%
11 - Maior serendipidade $\{AV_9\}$	7	0	0%

Tabela 12 – Aceite de recomendação por critério - sessão 4

Critério	**sessão 4**		
	#recomend.	**#aceites**	**IAceite**
3 - Maior similaridade intergrupo $\{AV_{13}\}$	283	30	11%
13 - Histórico do usuário $\{AV_1\}$	253	25	10%
8 - Notícia mais curtida do portal $\{AV_{14}\}$	246	23	9%
6 - Notícia mais lida do portal $\{AV_{14}\}$	247	23	9%
5 - Notícia mais lida do canal $\{AV_{14}\}$	248	23	9%
7 - Notícia mais curtida do canal $\{AV_{14}\}$	260	21	8%
2 - Maior similaridade intragrupo $\{AV_{12}\}$	290	23	8%
12 - Similaridade entre usuários $\{AV_{11}\}$	206	14	7%
1 - Maior similaridade de conteúdo $\{AV_{10}\}$	281	17	6%
4 - Aleatória não recomendada SR $\{AV_6, AV_7\}$	307	18	6%
9 - Mais recomendações aceitas $\{AV_{14}\}$	100	3	3%
10 - Maior nota de avaliação $\{AV_{14}\}$	69	1	1%
11 - Maior serendipidade $\{AV_9\}$	43	0	0%

Uma possibilidade que pode ser explorada em trabalhos futuros é a combinação do critério de serendipidade com o critério de similaridade entre usuários, algo que não foi testado especificamente por este experimento. Vale notar que o indicador de aceite alternou consideravelmente entre as três sessões e ficou com valores próximos entre os critérios, o que demonstra que não houve uma preferência em comum por algum critério específico, o que permite concluir que o fator diversidade foi bem explorado pelos participantes.

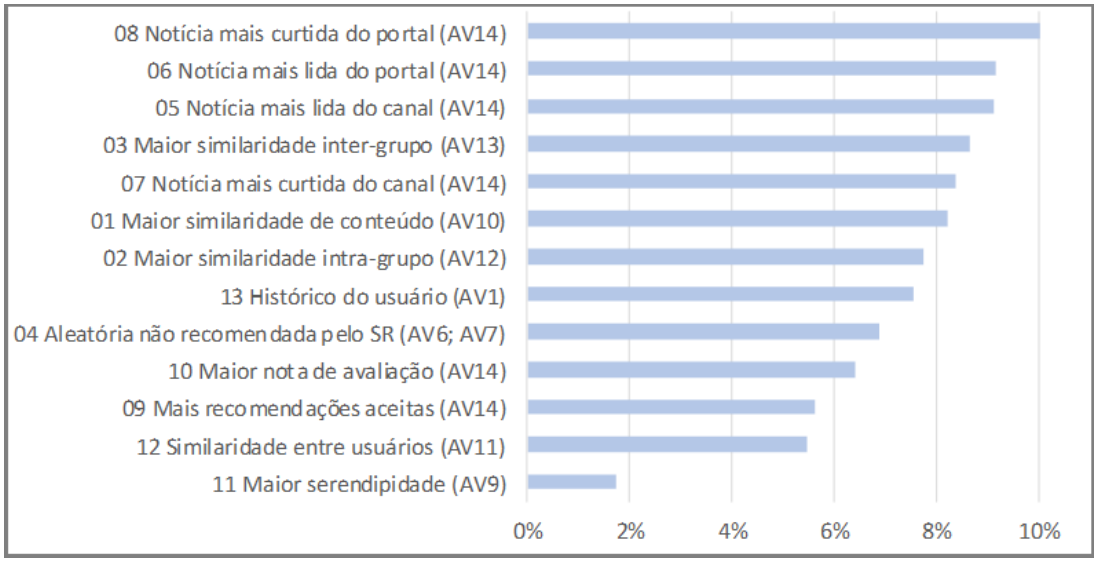
Por meio do indicador de aceite de recomendação, foi possível observar o efeito da diversidade ($\{AV_7\}$) no âmbito geral do experimento. O gráfico ilustrado pela figura 39 apresenta o indicador de aceite de recomendação considerando todas as sessões do experimento.

É possível observar que o indicador de aceite de recomendação teve uma distribuição com valores próximos entre os critérios, exceto o critério de serendipidade que demonstrou uma particularidade já discutida anteriormente.

O indicador de nota da recomendação

O indicador de nota da recomendação ($\{AV_{14}\}$) foi calculado a partir da nota capturada por meio da avaliação do participante em relação à recomendação (o participante

Figura 39 – Indicador de aceite de recomendação por critério: o efeito da diversidade

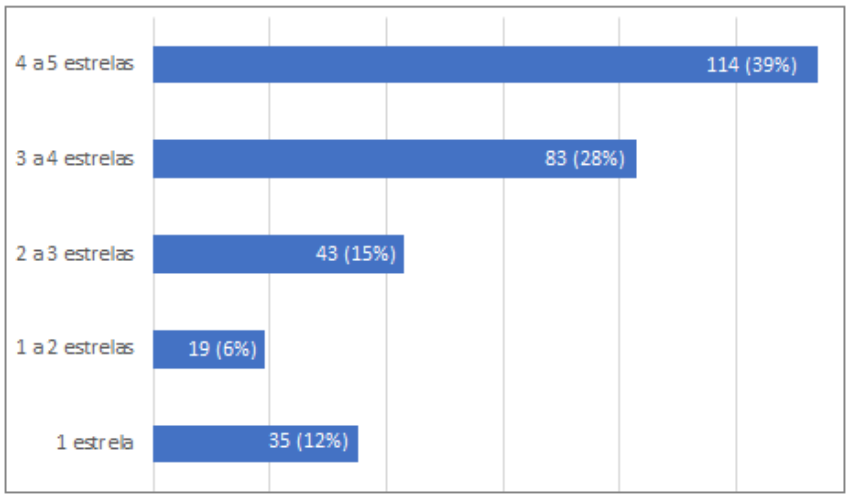


Fonte: José Luiz Maturana Pagnossim, 2018

escolhia uma nota para recomendação na escala 1-5 estrelas). Por ser uma avaliação da recomendação, a nota era atribuída a uma dupla ($\mathcal{NO}$, $\mathcal{ND}$).

Os resultados aqui apresentados para este indicador mostram que foram avaliadas aproximadamente trezentas recomendações que foram agrupadas em cinco faixas de avaliações, sendo que quanto maior o número de estrelas, melhor é a avaliação da recomendação. Na figura 40 é demonstrado o gráfico contendo as cinco faixas de avaliação e os respectivos valores absolutos e relativos de quantidade de avaliações.

Figura 40 – Indicador de nota de recomendação



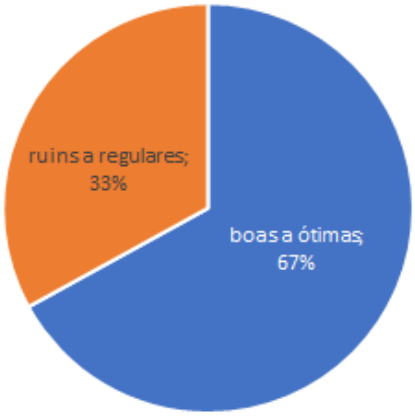
Fonte: José Luiz Maturana Pagnossim, 2018

No gráfico é demonstrado que os participantes aprovaram a maior parte das recomendações, já que um pouco mais de cem avaliações ficaram na faixa entre 4 e 5

estrelas, representando 39% das avaliações. Já as avaliações entre 3 e 4 estrelas totalizaram aproximadamente oitenta avaliações, representando 28% do total de avaliações.

Neste contexto, se as avaliações com notas maiores que três e menores ou iguais a cinco forem consideradas de boas a ótimas, no experimento foi registrado um percentual de 67% de avaliações de boas a ótimas contra 33% de avaliações com notas menores ou iguais a três, que podem ser consideradas de ruins a regulares. Na figura 41 é exibido o gráfico de avaliações agrupadas em ruins a regulares e boas a ótimas.

Figura 41 – Indicador agrupado de nota de recomendação



Fonte: José Luiz Maturana Pagnossim, 2018

O indicador de serendipidade

A serendipidade de uma recomendação foi avaliada por meio de um ícone que representava surpresa acompanhado de um texto que questionava ao participante se a recomendação havia causado surpresa positiva ($\{AV_9\}$). Este indicador ficou associado à relação ($\mathcal{NO}$, $\mathcal{ND}$). Na tabela 13 está demonstrado o o efeito da serendipidade em valores absolutos e relativos (por participante) para cada sessão do experimento. A serendipidade por participação é dada pela fórmula descrita na equação 8.

$$IndicadorSerendipidadeParticipacao = \frac{serendipidadeValorAbsoluto}{qtdParticipacoes} \tag{8}$$

Na figura 42 está demonstrado o gráfico com a evolução da serendipidade ao longo das sessões do experimento em relação à média geral.

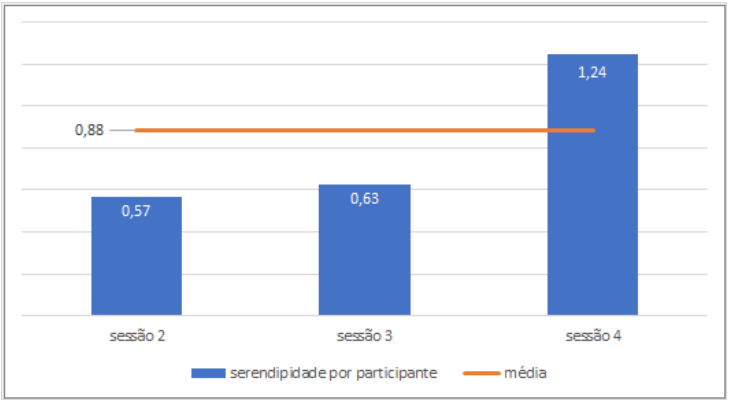
Os resultados relacionados à serendipidade acompanharam os indicadores apresentados anteriormente, com valores próximos nas sessões 2 e 3 e uma média de serendipidade

Tabela 13 – Indicador de serendipidade

indicador	**sessão 2**	**sessão 3**	**sessão 4**	**geral**
serendipidade em valores absolutos	17	20	61	98
quantidade de participações	30	32	49	111
serendipidade por participação	0,57	0,63	1,24	0,88

Fonte: José Luiz Maturana Pagnossim, 2018

Figura 42 – Indicador de serendipidade por participante



Fonte: José Luiz Maturana Pagnossim, 2018

por participante maior na sessão 4. Estes resultados sugerem que quanto mais distante do período de *cold start* do sistema, maior o indicador de serendipidade.

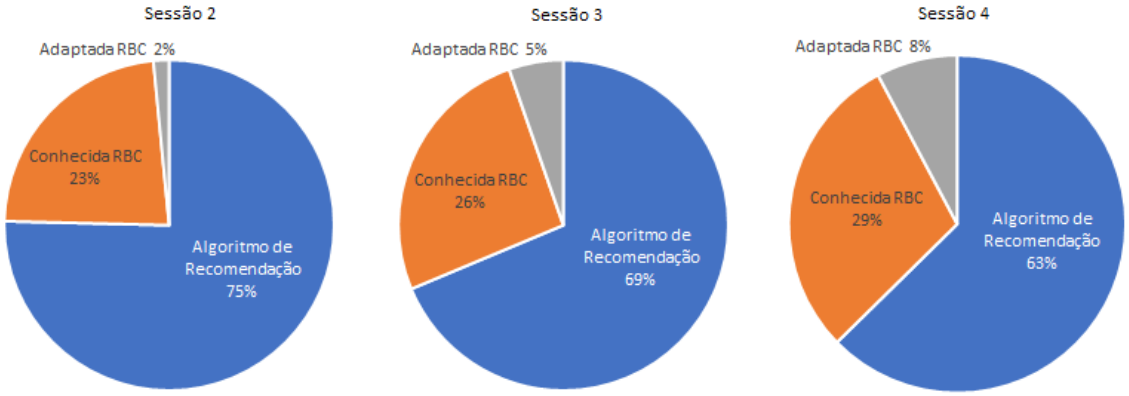
O indicador de estratégia de recomendação

O indicador de estratégia de recomendação apresenta os tipos de soluções usadas pelo sistema para fazer as recomendações ($\{AV_{15}\}$). Foram utilizadas pelo sistema três estratégias de recomendação: o algoritmo de recomendação; solução conhecida *RBC*; e solução adaptada *RBC*. Na figura 43 é demonstrado como o sistema utilizou as estratégias de recomendação ao longo das sessões do experimento.

Os dados demonstrados nos gráficos indicam que a estratégia mais usada foi sugerir itens com base no algoritmo de recomendação. Resultado esperado visto que a base de soluções de *RBC* ainda estava sendo enriquecida.

Observando os gráficos das sessões 3 e 4, nota-se que a representatividade da estratégia baseada no algoritmo de recomendação é reduzida, enquanto as estratégias que utilizam as soluções conhecida *RBC* e adaptada *RBC* tiveram suas representatividades aumentadas.

Figura 43 – Indicador de estratégia de recomendação

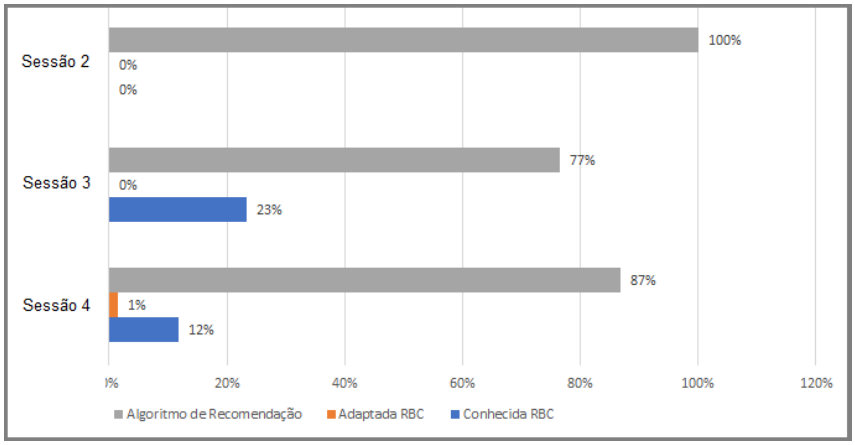


Fonte: José Luiz Maturana Pagnossim, 2018

A regra de utilização das estratégias, implementada especificamente para o experimento desta pesquisa, pode ser facilmente parametrizada e adaptada em trabalhos futuros, possibilitando a exploração de diferentes calibragens na definição das estratégias.

Outra informação derivada da estratégia de recomendação refere-se ao indicador de aceite para cada tipo de solução. Este tipo de análise permite avaliar a utilidade de cada solução ($\{AV_8\}$), conforme pode ser observado nos gráficos da figura 44.

Figura 44 – Indicador de utilidade por estratégia de recomendação



Fonte: José Luiz Maturana Pagnossim, 2018

Com base nas informações ilustradas nos gráficos é possível observar que na sessão 2, somente as recomendações geradas pelo algoritmo de recomendação do sistema obtiveram aceite. Já na sessão 3 é possível observar que o indicador do algoritmo de recomendação reduziu em relação à sessão anterior. Este resultado ocorreu devido ao aumento dos aceites baseados na solução conhecida *RBC* (23% dos aceites). Na sessão 4 é possível observar que a estratégia baseada na solução adaptada *RBC* aparece no indicador de utilidade com

1% de aceite, seguido de 12% da estratégia de solução conhecida *RBC* e 87% do algoritmo de recomendação.

Estes resultados não permitem identificar um padrão ou uma tendência, mas podem ser úteis para futuras calibragens a respeito do uso das estratégias.

Indicador baseado na ($\mathcal{MURR}$)

Este indicador foi gerado com base na da ($\mathcal{MURR}$) (equação 4) e está associado à propriedade de ranqueamento de item ($\{AV_2\}$). O cálculo e registro desta métrica foi feito exclusivamente nas recomendações realizadas com base no algoritmo de recomendação baseado em critérios de recuperação da notícia.

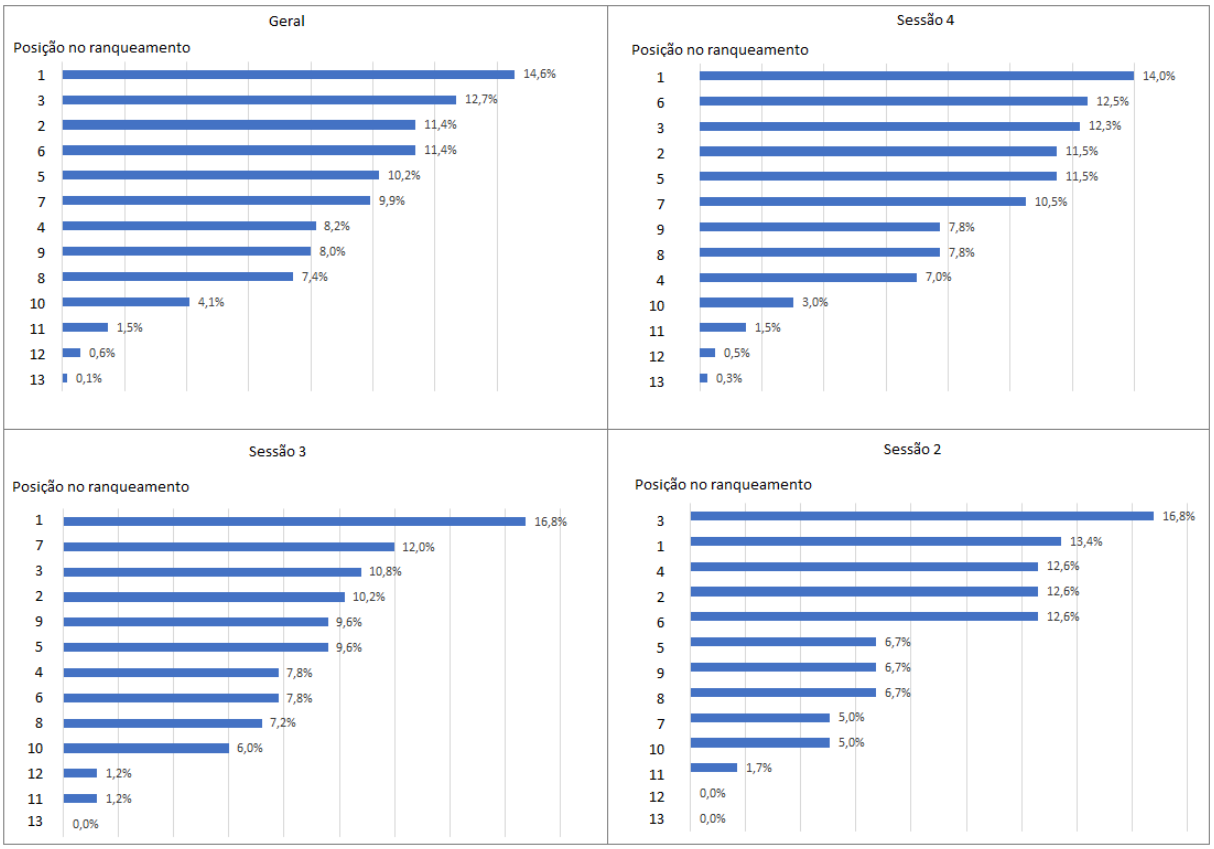
Em um sistema de recomendação convencional, este ranqueamento é utilizado para definir a ordem em que as notícias são listadas na tela do usuário, normalmente elas são apresentadas de cima para baixo, com a notícia mais relevante no topo.

Já no contexto do experimento projetado para esta pesquisa, a posição de cada notícia dentro deste ranqueamento foi registrada pelo sistema porém não foi mostrada ao participante, já que conhecidamente há uma maior probabilidade de aceite nos itens que estão no topo da lista. Para que esta possível vantagem dos itens posicionados no topo da lista não influenciasse a decisão de aceite do participante, ao montar a lista de recomendação para exibição na tela do portal de notícias, o sistema ordenou a lista aleatoriamente, ou seja, o participante não tinha conhecimento em que posição da lista estavam as melhores recomendações.

Ao relacionar os aceites das recomendações com as informações de ranqueamento, foi possível calcular o percentual de aceite de recomendação que cada posição no ranqueamento obteve. Na figura 45 são apresentados os gráficos de cada sessão, assim como a visão geral que consolida os dados de todas as sessões do experimento. Estes gráficos visam auxiliar na análise da eficácia da fórmula utilizada para definição da $\mathcal{MURR}$.

No gráfico da sessão 2 está demonstrado que 16.8% dos aceites foram feitos em notícias que estavam na posição 3 do ranqueamento, algo que se aproxima do esperado. Já as notícias recomendadas que estavam na posição 11 tiveram apenas 1.7% de aceites (também era esperado que as notícias no final do ranqueamento seriam menos aceitas).

Figura 45 – Indicador baseado na $\mathcal{MURR}$



Fonte: José Luiz Maturana Pagnossim, 2018

Nota-se na sessão 2 a ausência das posições 12 e 13, causado pela inexistência de aceites em notícias que estavam colocados nestas posições.

Embora a sequência de aceites das quatro primeiras colocações tenha ficado, respectivamente, nas posições 3, 1, 4 e 2, a representatividade destas quatro posições ficou em 55,5% do total, o que indica que mais da metade dos aceites ocorreram em notícias que estavam colocadas nas primeiras posições do ranqueamento.

No gráfico da sessão 3, a maior parte dos aceites ficou concentrado nas notícias recomendadas na posição 1 do ranqueamento, na sequência ficaram as posições 7, 3 e 2 (resultado próximo do esperado). Nas três últimas colocações ficaram as posições 10, 12 e 11 (resultado próximo do esperado). Nesta sessão, não houve aceite para notícias recomendadas na posição 13.

Na sessão 4, a posição 1 voltou a ficar com maior indicador de aceite (resultado esperado), seguido pelas posições 6, 3 e 2. Nas colocações finais desta sessão ficaram as posições 10, 11, 12 e 13, respectivamente (resultado esperado). Nesta sessão foi registrado 1 aceite na posição 13, o que representou um indicador de aceite de 0,3%.

Na visão consolidada do indicador (incluindo todas as sessões), o resultado também foi o esperado, colocando a posição 1 do ranqueamento com maior indicador de aceite, seguido pelas posições 3 e 2 (estas três posições somadas representaram 38,6% sobre o total de aceites). Nas últimas colocações, apareceram as posições 11, 12 e 13, com uma representatividade de 2,2%, considerando a soma destas três posições (resultado esperado).

Por meio dos resultados gerados pelo experimento foi possível observar que as recomendações mais relevantes foram as mais aceitas pelos participantes. Já as recomendações menos relevantes foram as menos aceitas, mesmo com a proposital ordenação aleatória na apresentação da lista de notícias recomendadas.

4.3.2 Considerações sobre os resultados

A partir das análises dos dados produzidos pelo experimento foi possível sumarizar informações a fim de fornecer uma visão objetiva sobre as discussões apresentadas neste capítulo:

- Os participantes do experimento preferiram navegar como leitores anônimos (52%) a ter que cadastrar um usuário (48%).
- Os resultados das sessões 2 e 3 apontam para um cenário típico de *cold start* de item.
- Os canais de notícias favoritos dos participantes foram em primeiro lugar tecnologia e em segundo educação.
- A propriedade da diversidade foi aplicada na recomendação e também foi evidenciada pela análise do indicador de aceite que ficou distribuído de forma equilibrada em relação aos critérios de recuperação utilizados.
- Dois terços dos usuários ficaram satisfeitos com as recomendações.
- O algoritmo de distribuição das estratégias de recomendação atingiu o resultado esperado, balanceando automaticamente as estratégias de recomendação ao longo das sessões do experimento.
- Por meio do indicador de utilidade foi possível observar um maior índice de aceite a favor do algoritmo de recomendação com base em critérios de recuperação.
- A quarta sessão do experimento apresentou melhores resultados em relação às sessões anteriores, o que demonstra que o sistema evoluiu após duas sessões de experimentação. Indicadores que apresentaram melhores resultados na sessão 4:

1. Média de leituras por participação (37% maior que as sessões anteriores).
2. Média de curtidas por participação (95% maior que as sessões anteriores).
3. Indicador agrupado de aceite (16% maior que as sessões anteriores).
4. Indicador de serendipidade (107% maior que as sessões anteriores).

4.3.3 Avaliação das funcionalidades da arquitetura

Ao considerar a arquitetura como uma contribuição técnica desta pesquisa, torna-se relevante avaliá-la em termos de atendimento aos requisitos essenciais para seu funcionamento em um ambiente de prova de conceito. Esta avaliação está organizada seguindo a fundamentação de *SR* que considera: funções dos *SR*; tarefas dos *SR*; tipos de *SR*; uso de tarefas e procedimentos da mineração de dados; e as estratégias para recomendação.

Para identificar se a arquitetura atendeu aos requisitos, são definidos os possíveis valores para o indicador de avaliação da arquitetura (*IAA*): T - atendeu totalmente ao requisito; P - atendeu parcialmente ao requisito; e N -não atendeu ao requisito. Este indicador foi atribuído pelo autor durante o processo de acompanhamento do experimento *online* e apuração dos resultados, considerando o funcionamento técnico e funcional da arquitetura.

O quadro 6 ilustra a avaliação da arquitetura em relação às funções de *SR*.

Quadro 6 – Avaliação da arquitetura: funções dos *SR*

Requisito	***IAA***
$\{R_1\}$ Aumentar o número de notícias lidas	T
$\{R_2\}$ Ter notícias mais diversificadas sendo aceitas	T
$\{R_3\}$ Aumentar a satisfação do leitor e do usuário	P
$\{R_4\}$ Aumentar a fidelização de clientes	T
$\{R_5\}$ Melhorar o conhecimento das preferências do usuário	T

Fonte: José Luiz Maturana Pagnossim, 2018

A arquitetura se mostrou capaz de registrar dados de notícias lidas, possibilitando a evolução deste número de leituras ao longo das sessões do experimento ($\{R_1\}$). Também foi possível observar pelos resultados do experimento que a arquitetura fez recomendações diversificadas (até 13 itens de diferentes critérios) e isso refletiu na diversidade de aceites de recomendação por parte dos leitores e usuários($\{R_2\}$). A satisfação do leitor e do usuário foi medida no experimento por meio de um indicador de aceite das recomendações e por uma nota da recomendação, desta forma, este indicador foi atribuído como P - parcial

pelo fato da nota de avaliação ter sido calculada por meio de uma média que atualizava este indicador sempre que uma nova nota era atribuída à recomendação, impossibilitando um histórico comparativo dessas notas ($\{R_3\}$). A fidelização dos clientes foi viabilizada na arquitetura por meio do cadastro de usuários e pela possibilidade deste usuário retornar ao sistema e ser identificado ($\{R_4\}$). O conhecimento das preferências do usuário foi validada pelo histórico de navegação do usuário e pelas recomendações feitas pelo sistema com base neste tipo de conhecimento ($\{R_5\}$).

O quadro 7 ilustra a avaliação da arquitetura em relação às tarefas dos *SR*.

Quadro 7 – Avaliação da arquitetura: tarefas dos *SR*

Requisito	***IAA***
$\{R_6\}$ Encontrar alguns itens bons (itens que outros usuários mais gostam)	T
$\{R_7\}$ Anotações no contexto (histórico em segundo plano)	T
$\{R_8\}$ Melhorar o perfil (incluindo o que gosta e não gosta)	P

Fonte: José Luiz Maturana Pagnossim, 2018

O sistema se mostrou capaz de recomendar itens com base em similaridade usuário-usuário ($\{R_6\}$). As anotações em segundo plano foram feitas para registrar o histórico de interação do leitor e também por meio do registro de um controle mais detalhado que registrou as ações executadas pelo participante independente se seria ou não usado para avaliação ($\{R_7\}$). Este controle foi disponibilizado a partir da segunda sessão do experimento. A melhoria do perfil do usuário foi considerada parcial, pois apesar da recomendação com base no histórico de navegação, o sistema não registrou dados cadastrais que pudessem ser usados em recomendação. Também não foi feita uma classificação das preferências do usuário em rótulos do tipo: "gosto" ou "não gosto" ($\{R_8\}$).

O quadro 8 ilustra a avaliação da arquitetura em relação aos tipos de recomendação.

Quadro 8 – Avaliação da arquitetura: tipos de recomendação em *SR*

Requisito	***IAA***
$\{R_9\}$ Recomendação baseada em conteúdo	T
$\{R_{10}\}$ Recomendação baseada em conhecimento	T
$\{R_{11}\}$ Recomendação baseada em filtro colaborativo	T
$\{R_{12}\}$ Recomendação baseada em casos	P

Fonte: José Luiz Maturana Pagnossim, 2018

O sistema se mostrou capaz de recomendar itens similares utilizando a medida de distância cosseno extraída a partir dos conteúdos das notícias ($\{R_9\}$). O sistema fez recomendações com base em conhecimento por meio de: dados de domínio como o canal

de notícia para fazer recomendação; dados estruturados obtidos da interação do usuário (indicadores de popularidade); dados de preferência do usuário ($\{R_{10}\}$). O modelo de vizinhança usuário-usuário foi implementado possibilitando fazer recomendações baseadas em filtro colaborativo ($\{R_{11}\}$). A *RBC* foi usada com base em soluções conhecidas e também foi capaz de adaptar soluções ($\{R_{12}\}$). Apesar destes recursos, o ciclo de raciocínio baseado em casos não foi totalmente utilizado neste trabalho[12], podendo ser objeto de estudos futuros.

O quadro 9 ilustra a avaliação da arquitetura em relação às tarefas e procedimentos de mineração de dados.

Quadro 9 – Avaliação da arquitetura: mineração de dados

Requisito	***IAA***
$\{R_{13}\}$ Pré-Processamento	T
$\{R_{14}\}$ Agrupamento	T

Fonte: José Luiz Maturana Pagnossim, 2018

A rotina de pré-processamento foi responsável por fornecer à arquitetura o conjunto de dados que registrou a matriz de distância, estrutura essencial na recomendação baseada em conteúdo e também utilizada como insumo para a tarefa de agrupamento ($\{R_{13}\}$). A recomendação com base em agrupamento foi implementada e usada no experimento por meio de similaridade intragrupo e intergrupos ($\{R_{14}\}$).

O quadro 10 ilustra a avaliação da arquitetura em relação às estratégias utilizadas para recomendação.

Quadro 10 – Avaliação da arquitetura: estratégias para recomendação

Requisito	***IAA***
$\{R_{15}\}$ Algoritmo de recomendação com base em critérios de recuperação da notícia	T
$\{R_{16}\}$ *RBC* por meio de uma solução conhecida	T
$\{R_{17}\}$ *RBC* por meio de uma solução adaptada	T

Fonte: José Luiz Maturana Pagnossim, 2018

O algoritmo implementado para equilibrar o uso das estratégias de recomendação funcionou conforme esperado, ajustando as estratégias conforme a base de soluções da *RBC* era incrementada.

[12] Faltou a etapa de revisão que testa e repara os "casos" resultando em uma solução confirmada

4.4 Limitações e ameaças

Embora o estudo para seleção do *corpus* tenha sido criterioso, o recorte de notícias foi obtido em um intervalo de tempo que gerou uma defasagem de até 8 meses entre a data de publicação da notícia na internet e o período de realização experimento. Outras limitações sobre o *corpus* tem a ver com: o tamanho da amostra (1097); os assuntos ou canais de notícias obtidos (que foram reportados por alguns participantes como pouco interessantes); e os dados faltantes (sendo necessários tratamentos para padronizar estes dados, como a retirada dos atributos foto e autor da notícia, já que nem todas notícias tinham esses dados.

O tamanho e o tipo da amostra de pessoas que participaram do experimento também configura uma limitação, já que contou com 111 pessoas do entorno acadêmico e profissional dos pesquisadores (por uma questão de viabilidade de execução do projeto). Constituindo portanto uma amostra com poucos elementos e que possuem características comportamentais similares, representando um recorte do universo de usuários de *SR* de notícias. No que se refere à avaliações qualitativas previstas para essa pesquisa, pode ter ocorrido algum tipo de viés de avaliação devido à questões subjetivas inerentes ao comportamento humano. Uma evidência observada nos resultados está relacionada à preferência dos usuários pelos canais de notícias, em que o canal preferido foi o de tecnologia, seguido pelo de educação, coincidindo com o perfil do público selecionado.

Ainda sob a perspectiva do viés relacionado às avaliações qualitativas (e subjetivas) dos usuários, surge uma ameaça associada à propriedade robustez. Esta propriedade descreve situações em que o sistema pode ser manipulado por usuários que inserem dados falsos ou ainda fornecem avaliações positivas ou negativas em grande escala. Este comportamento afeta o núcleo de recomendação de *SR* que se baseiam, principalmente, em indicadores de popularidade. Sistemas mais robustos fazem controles proativos e reativos para minimizar os efeitos destes tipos de ataques. Estes controles não foram implementados neste trabalho e portanto a pesquisa não ficou isenta deste tipo de comportamento.

A implementação do cálculo da média de avaliação da recomendação apresentou uma limitação que impediu a extração de resultados separados por sessão. Esta limitação foi identificada na fase apuração dos resultados, impossibilitando a adequação deste cálculo a tempo de ser aplicado no experimento.

5 Conclusão

Esta pesquisa levantou temas relevantes para a área de *SR* no que ser refere ao uso de recomendação baseada em: conteúdo, conhecimento, filtro colaborativo e casos, sendo esta última adaptada de conceitos advindos do raciocínio baseado em casos. O trabalho implementou uma arquitetura fundamentada nos conceitos de *SR* e apresentou a fórmula de uma métrica unificada para ranqueamento da recomendação.

Por meio da análise dos dados gerados pela realização de um experimento *online* foi possível concluir que a pesquisa confirmou a hipótese delineada no que se refere à privilegiar recomendações com base em: similaridade; popularidade; diversidade; novidade; e serendipidade.

A hipótese também foi confirmada por meio da análise da evolução dos indicadores de leitura, curtida, aceite e serendipidade. Tal evolução foi observada comparando a sessão 4 do experimento com as sessões anteriores. A evolução pode ser atribuída à capacidade do protótipo do sistema de recomendação ter acumulando dados históricos de preferências dos usuários e construído uma base histórica de "casos" (conhecidos e adaptados), permitindo recomendações mais diversificadas e portanto, mais adequadas.

A satisfação do usuário foi confirmada por meio da análise do indicador de nota de avaliação da recomendação (escala 1-5 estrelas), resultando em mais de dois terços dos participantes satisfeitos com as recomendações.

Os resultados do experimento também permitiram evidenciar a eficácia da métrica unificada para ranqueamento da recomendação. Foi verificado que as notícias melhores colocadas no ranqueamento foram as mais aceitas pelos participantes, enquanto que as notícias colocadas nas últimas posições do ranqueamento foram as menos aceitas.

A arquitetura apresentada e implementada nesta pesquisa se mostrou robusta e funcional, já que atendeu adequadamente aos requisitos técnicos e também forneceu funcionalidades que permitiram a interação *online* dos participantes.

Por meio do experimento foi possível observar o efeito do *cold start* de item, indicando que o sistema enfrentou este problema entre as sessões 2 e 3 e o superou na sessão 4 (embora não seja possível apontar o momento exato em que o sistema deixou de ser impactado pelos itens considerados "frios"). Para superar o período de *cold start* de item esta pesquisa utilizou de diferentes estratégias, como a recomendação baseada em

conteúdo, fez uso de similaridade por agrupamento de dados e também inseriu um fator de aleatoriedade[1].

O algoritmo de distribuição das estratégias de recomendação atingiu o resultado esperado, balanceando automaticamente a utilização destas estratégias ao longo das sessões do experimento.

5.1 Contribuições adicionais

A abordagem híbrida implementada neste trabalho apresentou contribuições técnicas, como: um *corpus* com notícias reais em idioma português do Brasil, extraídas no período de abril de 2017 até setembro de 2017; um conjunto de dados com indicadores registrados durante um experimento *online* que contou mais de cem participações; e uma arquitetura para sistemas de recomendação de notícias.

A arquitetura de recomendação apresentada nesta pesquisa envolveu atividades de especificação, construção e integração dos seguintes módulos: o *corpus* de notícias; procedimentos e tarefas de mineração de dados; mecanismo de carga para uma base de caso; modelagem da base de casos sob o conceito de *RBC*; construção do protótipo de sistema de recomendação; uma camada de integração com a base de casos; e um núcleo de recomendação (contendo os algoritmos e estratégias utilizadas para a recomendação).

O projeto de pesquisa foi publicado na trilha regular (artigo completo) de uma conferência internacional da área de sistemas de informação. Nesta mesma área, o projeto de pesquisa foi apresentado em um *workshop*. Uma última contribuição refere-se à publicação de um relatório técnico na área de mineração de dados.

- Pagnossim, J. L. M.; Peres, S. M.. Uma arquitetura híbrida para sistemas de recomendação de notícias baseada em casos. 14th CONTECSI - International Conference on Information Systems and Technology Management. São Paulo-SP, 2017, p. 4650-4672.
- Pagnossim, J. L. M.; Peres, S. M.. Arquitetura para Sistemas de Recomendação de Notícias: Uma Abordagem Híbrida e Baseada em Casos. WTDSI 2017 - X Workshop de teses e dissertações em sistemas de informação. Lavras-MG, 2017.

[1] A aleatoriedade foi combinada com a diversidade, uma vez que o recomendador sugeria uma notícia aleatória que ainda não tivesse sido recomendada anteriormente pelo sistema.

- A. Diaz, A. Lima, A. Silva, F. Costa, J. Pagnossim, S. Peres. Uma análise comparativa das ferramentas de pré-processamento de dados textuais: NLTK, PreTexT e R[2]. Relatórios Técnicos, EACH-USP. São Paulo-SP, 2018.

5.2 Trabalhos futuros

A arquitetura apresentada pode ser generalizada a outros domínios de dados e áreas de aplicação, entre as quais destacam-se: recomendação de filmes e vídeos; recomendação de imagens; recomendação de músicas e áudios; e recomendação de outras fontes textuais (artigos científicos; repositório de documentos corporativos; e logs de sistemas).

Importante considerar quanto à defasagem das notícias que uma alternativa é a construção de algoritmos capazes de obter as notícias de forma *online*, carregando-as automaticamente para uso da arquitetura do sistema de recomendação. Este tipo de funcionalidade possibilita considerar aspectos relacionados à temporalidade da notícia, aplicando de forma mais abrangente a novidade como propriedade de recomendação e possibilitando o estudo da influência da temporalidade na serendipidade. Convém também a extração de mais metadados das notícias, como: o autor da notícia; imagens, *hiperlinks* e palavras-chaves, ampliando a capacidade de recomendação do sistema.

Sobre o ciclo do raciocínio baseado em casos, sugere-se seu uso completo, considerando a etapa "3. Revisa", que confronta a solução recuperada com a base de casos. Esta etapa testa e repara o "caso" proporcionando maior inteligencia ao recomendador que contaria com todas as etapas do método de raciocínio baseado em casos. A inclusão desta etapa proporciona a recomendação de uma solução considerada "confirmada" que pode medida em termos da propriedade de utilidade que compara os resultados advindos de diferentes métodos ou algoritmos de recomendação.

[2] (DIAZ et al., 2018)

Referências[3]

AAMODT, A.; PLAZA, E. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, p. 39–59, 1994. Citado na página 32.

AMATRIAIN, X. et al. Data mining methods for recommender systems. In: RICCI, F. et al. (Ed.). *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 39–71. Citado 2 vezes nas páginas 27 e 28.

BERMINGHAM, A. et al. Automatically recommending multimedia content for use in group reminiscence therapy. *Health Monitoring and Personalized Feedback using Multimedia Data*, p. 215–244, 2015. Citado na página 44.

BOBADILLA, J. et al. Recommender systems survey. *Knowledge-Based Systems*, Elsevier, v. 46, p. 109–132, 2013. Citado 2 vezes nas páginas 31 e 43.

BRUNIALTI, L. F. et al. Aprendizado de máquina em sistemas de recomendação baseados em conteúdo textual: Uma revisão sistemática. In: *Anais do XI Simpósio Brasileiro de Sistemas de Informação (SBSI)*. [S.l.: s.n.], 2015. p. 203–210. Citado na página 18.

BRUSILOVSKY. *The adaptative Web Methods and Strategies of Web Personalization*. [S.l.]: Springer, 2007. Citado na página 19.

DIAZ, A. et al. *Uma análise comparativa das ferramentas de pré-processamento de dados textuais: NLTK, PreTexT e R*. [S.l.], 2018. Citado 3 vezes nas páginas 28, 53 e 111.

DONG, R.; O'MAHONY, M. P.; SMYTH, B. Further experiments in opinionated product recommendation. *Case-Based Reasoning Research and Development Lecture Notes in Computer Science*, p. 110–124, 2014. Citado na página 44.

FEINERER, I.; HORNIK, K. *tm: Text Mining Package*. [S.l.], 2017. R package version 0.7-1. Disponível em: ⟨https://CRAN.R-project.org/package=tm⟩. Citado na página 28.

GUNAWARDANA, S. Evaluating recomender systems. In: RICCI, F. e. a. (Ed.). *Recommender Systems Handbook Second Edition*. [S.l.]: Springer, 2015. p. 265–308. Citado 2 vezes nas páginas 72 e 76.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3. ed. [S.l.]: Morgan Kaufmann, 2011. Citado 4 vezes nas páginas 20, 28, 29 e 36.

HERLOCKER, J. et al. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, p. 5–53, 2004. Citado na página 25.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of Classification*, v. 2, n. 1, p. 193–218, 1985. Citado na página 30.

JONNALAGEDDA, N. et al. Incorporating popularity in a personalized news recommender system. *PeerJ Computer Science*, v. 2, Jun 2016. Citado 2 vezes nas páginas 45 e 47.

[3] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

KERMANY, N. R.; ALIZADEH, S. H. A hybrid multi-criteria recommender system using ontology and neuro-fuzzy techniques. *Electronic Commerce Research and Applications*, v. 21, p. 50–64, 2017. Citado na página 44.

KILGARRIFF, A. Web as corpus. *Proceedings of the Corpus Linguistics 2001 Conference*, p. 342–344, 2001. Citado na página 52.

KITCHENHAM, B. *Procedures for Performing Systematic Reviews: Keele University Technical Report*. [S.l.], 2004. Citado na página 42.

KOLODNER, J. *Case-Based Reasoning*. 1. ed. [S.l.]: Morgan Kaufmann Publishers Inc., 1993. Citado na página 20.

KOREN, Y.; BELL, R. Advances in collaborative filtering. *Recommender Systems Handbook*, p. 77–118, 2015. Citado 2 vezes nas páginas 27 e 37.

KUNAVER, M.; POžRL, T. A hybrid recommendation system for news in a mobile environment. *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics - WIMS 16*, 2016. Citado 2 vezes nas páginas 18 e 46.

KUNAVER, M.; POžRL, T. Diversity in recommender systems – a survey. *Knowledge-Based Systems*, v. 123, p. 154–162, 2017. Citado na página 45.

LENZ, M. et al. *Case-Based Reasoning Technology: From Foundations to Applications*. [S.l.]: Springer-Verlag, 1998. Citado 2 vezes nas páginas 20 e 31.

LEWIS, D. D. et al. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, v. 5, p. 361–397, 2004. Cited By :1161. Disponível em: ⟨www.scopus.com⟩. Citado na página 46.

LIAO, C.-L.; LEE, S.-J. A clustering based approach to improving the efficiency of collaborative filtering recommendation. *Electronic Commerce Research and Applications*, Elsevier, v. 18, p. 1–9, 2016. Citado 2 vezes nas páginas 31 e 45.

LIU, Y. et al. Understanding of internal clustering validation measures. *2010 IEEE International Conference on Data Mining*, 2010. Citado na página 30.

LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*. [S.l.], 2002. p. 63–70. Citado na página 28.

LOPS, P.; GEMMIS, M. de; SEMERARO, G. Content-based recommender systems: State of the art and trends. In: RICCI, F. et al. (Ed.). *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 73–105. Citado na página 18.

LU, J. et al. A web-based personalized business partner recommendation system using fuzzy semantic techniques. *Computational Intelligence*, v. 29, n. 1, p. 37–69, 2012. Citado na página 44.

MALIK, Z. K.; FYFE, C. Review of web personalization. *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE*, School of Computing, University of The West of Scotland, v. 4, n. 3, p. 285–296, 2012. Citado na página 44.

NAVATHE, S. B.; ELSMARI, R. *Sistemas de Banco de Dados*. 6. ed. [S.l.]: Pearson Brasil, 2013. Citado na página 60.

RAND, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, v. 66, n. 336, p. 846, 1971. Citado na página 30.

REZENDE, O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Revista de Sistemas de Informacao da FSMA*, FMSA, p. 7–11, 2011. Citado na página 37.

RIBEIRO-NETO, B.; BAEZA-YATES, R. *Modern Information Retrieval*. 1. ed. [S.l.]: Addison-Wesley, 2011. Citado na página 27.

RICCI, F.; ROKACH, L.; SHAPIRA, B. *Recommender Systems Handbook Second Edition*. [S.l.]: Springer, 2015. Citado na página 26.

RICCI, F. et al. *Recommender Systems Handbook*. [S.l.]: Springer, 2011. Citado 6 vezes nas páginas 18, 24, 25, 26, 41 e 82.

ROUSSEEUW, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, v. 20, n. 1, p. 53–65, 1987. Citado na página 30.

SAQUIB; SIDDIQUI, J.; ALI, R. Classifications of recommender systems a review. *Journal of Engineering Science and Technology Review*, Elsevier, v. 18, p. 132–153, 2017. Citado 2 vezes nas páginas 20 e 43.

SHANI, G.; GUNAWARDANA, A. Evaluating recommendation systems. In: RICCI, F. et al. (Ed.). *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 257–297. Citado na página 34.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. *Introdução à Mineração de Dados com Aplicações em R*. 1. ed. [S.l.]: Elsevier, 2016. Citado 2 vezes nas páginas 20 e 27.

SMYTH, B. Case-based recommendation. In: BRUSILOVSKY, P.; KOBSA, A.; NEJDL, W. (Ed.). *The Adaptive Web*. [S.l.]: Springer, 2007. p. 342–376. Citado 3 vezes nas páginas 20, 31 e 33.

SON, L. H. Dealing with the new user cold-start problem in recommender systems: A comparative review. *Information Systems*, v. 58, p. 87–104, 2016. Citado na página 45.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. [S.l.]: Addison-Wesley Longman Publishing, 2005. Citado na página 37.

TATAR, A. et al. From popularity prediction to ranking online news. *Social Network Analysis and Mining*, v. 4, n. 1, Dec 2014. Citado 3 vezes nas páginas 39, 40 e 47.

VINH, N. X.; EPPS, J.; BAILEY, J. Information theoretic measures for clusterings comparison. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML 09*, 2009. Citado na página 30.

Apêndice A – Página de Instruções Iniciais

Obrigado por participar desse experimento, sua atuação consiste em ler as notícias de seu interesse, navegar pelo portal e interagir com as notícias e as recomendações.

Você pode colaborar como um leitor anônimo, ou ainda, se cadastrar (opcionalmente) no portal de notícias, como forma de personalizar a recomendação das notícias de acordo com seu perfil.

Não há um tempo mínimo nem máximo para sua participação, podendo sair da pesquisa a qualquer momento.

As notícias são reais, obtidas do portal EBC ⟨www.ebc.com.br⟩ que concedeu permissão para uso das notícias.

A seguir é apresentado um Termo de Consentimento que deve ser lido e aceitado para que possa acessar o sistema.

Preencha os dados a seguir e clique no botão de Prosseguir:

Nome Completo:

Sexo:

Idade:

Clique Aqui para Prosseguir

Apêndice B – Termo de Consentimento Livre e Esclarecido

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Você está sendo convidado (a) para participar da pesquisa intitulada Avaliação de sistema de recomendação de notícias por meio de interações de usuários sob a responsabilidade dos pesquisadores José Luiz Maturana Pagnossim e Sarajane Marques Peres. Nesta pesquisa nós estamos buscando coletar interações entre indivíduos e um sistema de recomendação de notícias que foram carregadas do site da Empresa Brasileira de Comunicação - EBC ⟨www.ebc.com.br⟩. A pesquisa tem interesse em coletar interações de notícias lidas, aceite das recomendações, curtidas e avaliações das recomendações. A pesquisa tem interesse também em diferenciar um leitor anônimo de um usuário cadastrado no sistema. O objetivo é melhorar a qualidade das recomendações de notícias para leitores e usuários, disponibilizar um corpus com notícias reais em idioma português brasileiro e os indicadores de interação dos participantes com o sistema, registrados durante o período do experimento. Você pode acessar o experimento por meio de um computador pessoal, notebook, celular ou tablet, desde que conectado à internet e de posse do link para o protótipo do portal de notícias. A coleta poderá ocorrer nos turnos da manhã, tarde ou noite a depender da disponibilidade do participante e durante o período estabelecido de cada sessão ocorrida no experimento. Após esse período, os dados serão explorados pelo pesquisador para análise dos resultados e conclusões da pesquisa. Na sua participação você acessará o sistema, através de um endereço de site na internet, fornecido pelo pesquisador, e poderá navegar entre os canais de leitura, ler as notícias de seu interesse, aceitar as recomendações sugeridas pelo sistema, avaliar a notícia através de curtida, avaliar as recomendações Você poderá ainda se cadastrar no sistema, informando alguns dados relacionados à sua preferência de leitura. Os dados pessoais cadastrados não serão divulgados, para garantia da privacidade do participante. Você não terá nenhum gasto e ganho financeiro por participar na pesquisa. Os riscos mais prováveis consistem cansaço da vista e / ou fadiga, devido à necessidade de ler notícias através de um computador ou outro dispositivo móvel, durante o tempo que o participante estiver usando o sistema, tempo este determinado pelo próprio indivíduo, realizado em um único dia ou em mais dias, desde que dentro do período de vigência da pesquisa. Os benefícios desta pesquisa podem envolver diretamente os participantes caso ele seja um pesquisador e pretenda usar

o *corpus* disponibilizado ou ainda se o sistema for viabilizado por algum portal brasileiro de notícias, contribuindo com a melhoria na qualidade da recomendação. Você é livre para deixar de participar da pesquisa a qualquer momento sem nenhum prejuízo ou coação. Se esse termo não for aceito, o sistema é encerrado e nenhum registro de sua participação será armazenado. Por outro lado, optando em prosseguir, você aceita os termos deste documento e colabora com o projeto de pesquisa em questão. Caso queira uma cópia desse termo, você pode salvar ou imprimir uma via do Termo de Consentimento Livre e Esclarecido. Qualquer dúvida a respeito da pesquisa, você poderá entrar em contato com: José Luiz Maturana Pagnossim, telefone: (11) 98058-7052; Sarajane Marques Peres, telefone: ou (11)3091-8897; Rua Arlindo Bettio, 1000 - CEP: 03828-000, Vila Guaraciaba, São Paulo-SP. Poderá também entrar em contato com o Comitê de Ética na Pesquisa com Seres-Humanos – Escola de Artes, Ciências e Humanidades – Universidade de São Paulo : Rua Arlindo Bettio, 1000 - CEP: 03828-000, Vila Guaraciaba, São Paulo-SP.

São Paulo, dia (DD) do mês (MM) do ano (AAAA)

Assinatura digital dos pesquisadores: _____

Eu, (Nome do Participante) , sexo, (Sexo do Participante), Idade, (Idade do Participante), aceito participar do projeto citado acima, voluntariamente, após ter sido devidamente esclarecido.

Clique Aqui para Aceitar e Prosseguir com o Experimento

Apêndice C – Modo de navegação

Defina qual modo o seu modo preferido de navegação

1. Leitor anônimo: Nesse modo você tem acesso à todas as notícias do portal sem a necessidade de cadastrar um usuário. Por outro lado, o sistema não gera recomendações com base no perfil do usuário.

Acessar como leitor anônimo

2. Usuário cadastrado: Nesse modo você também tem acesso à todas as notícias e deve se cadastrar como um usuário do sistema para que receba recomendações com base no seu perfil. Você pode ainda acessar o sistema em outro momento e se identificar com seu usuário previamente cadastrado.

E-mail do usuário

Nome

Senha

Senha (Confirmação)

Cadastrar usuário

3. Já sou cadastrado: Nesse modo, basta identificar usuário e senha previamente cadastrados para ter acesso ao sistema e receber recomendações com base no seu perfil.

E-mail/Usuário

Senha

Acessar

Apêndice D – Protocolo para disponibilização dos dados

Este trabalho recebeu dois tipos de autorização por parte da empresa detentora dos direitos autorais das notícias (portal EBC): de utilização e de disponibilização. A autorização de utilização tem relação com o uso do conteúdo das notícias desde sua obtenção para montagem de um *corpus* (usado na pesquisa na etapa de mineração de texto), até a apresentação das notícias no portal utilizado como ambiente de um experimento *online*. Com relação à disponibilização dos dados, estes podem ser obtidos pelos interessados mediante solicitação aos pesquisadores deste trabalho. O formato de disponibilização destes dados envolve um *corpus* e um conjunto de indicadores:

- *Corpus* de notícias - construído como uma etapa preliminar do experimento, denominado *corpus-noticias-ebc-2017*. Este *corpus* contempla 1.097 arquivos texto, cada um contendo o conteúdo de uma notícia, e um arquivo texto separado por vírgula que consolida todas as notícias. O conteúdo das notícias foi extraído portal EBC de abril de 2017 até setembro de 2017.
- Conjunto de indicadores - resultantes da navegação e interação dos participantes com o portal de notícias utilizado no experimento. Os dados foram organizados um arquivo texto separado por vírgula, denominado *conjunto-indicadores-ebc-2017*, que relaciona os dados de navegação e interação dos participantes com as notícias contidas no *corpus-noticias-ebc-2017*. Vale ressaltar que este conjunto de dados não disponibiliza informações cadastrais dos participantes.