Aprendizado de Máquina em Sistemas de Recomendação Baseados em Conteúdo Textual: Uma Revisão Sistemática

Alternative Title: Machine Learning in Textual Content-Based Recommendation Systems: A Systematic Review

Lucas F. Brunialti
Universidade de São Paulo
São Paulo – SP – Brasil
lucas.brunialti@usp.br

Sarajane M. Peres
Universidade de São Paulo
São Paulo – SP – Brasil
sarajane@usp.br

Valdinei Freire
Universidade de São Paulo
São Paulo – SP – Brasil
valdinei.freire@usp.br

Clodoaldo A. M. Lima
Universidade de São Paulo
São Paulo – SP – Brasil
c.lima@usp.br

RESUMO

Sistemas de Recomendação baseados em Conteúdo (SRbC) é uma área em que estratégias de Aprendizado de Máquina (AM) podem ser potencialmente aplicadas com êxito. Contudo, especificamente na área de SRbC textual, o uso de AM não tem sido expressivo nos últimos anos. Neste artigo é apresentada uma Revisão Sistemática para identificação, interpretação e avaliação de como estratégias de AM vêm sendo utilizadas no contexto de SRbC textual a fim de contribuir para a evolução da interseção de tais áreas.

Palavras-Chave

Sistemas de Recomendação baseado em Conteúdo, Conteúdo Textual, Aprendizado de Máquina, Revisão Sistemática

ABSTRACT

Content-based Recommendation Systems (CbRS) is a research area in which Machine Learning (ML) strategies can be applied with success. However, specifically in textual CbRS, the use of ML has not been expressive in recent years. To contribute to the evolution of the intersection of such areas, we present a Systematic Review to identify, interpret and evaluate how the ML strategies have been applied to CbRS.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Miscellaneous; I.2 [**Artificial Intelligence**]: Learning—*Induction, Knowledge Acquisition*

General Terms

Design, Theory

Keywords

Content-Based Recommendation Systems, Textual Content, Machine Learning, Systematic Review

1. INTRODUÇÃO

Um sistema de recomendação (SR) tem o objetivo de sugerir itens de forma a satisfazer sob algum aspecto as necessidades de um usuário. Geralmente, esses sistemas atuam em contextos onde a tomada de decisão sobre a escolha de itens se dá em um conjunto grande de opções, no qual uma busca por meio de mecanismos clássicos como palavras-chave ou termos de interesse tem a chance de retornar resultados insatisfatórios. Os primeiros SR surgiram no início da década de 1990 [1], e um dos primeiros sistemas propostos chamava-se *Tapestry* [20], o qual introduziu a *Filtragem Colaborativa* como uma técnica para implementação da recomendação.

Em [10] é proposta uma taxonomia para tipos básicos de SR, de acordo com a forma como a recomendação é implementada: *colaborativo*, no qual a recomendação se dá com base em itens com os quais usuários interagiram no passado, relacionando as avaliações de usuários sobre um mesmo item; *baseado em conteúdo* (SRbC), no qual as recomendações são geradas com base nas características dos itens e no perfil do usuário; *baseado em conhecimento*, em que sugestões de itens são baseadas em inferências sobre as necessidades e preferências dos usuários. Cada um destes tipos possue vantagens e desvantagens que podem ser superadas por meio de sistemas *híbridos*, os quais são implementados a partir da combinação das estratégias usadas nos tipos básicos.

Os SR são considerados ferramentas de grande utilidade, principalmente para usuários de sistemas *online* [21, 25], os quais se deparam com um montante de informação maior do que conseguem lidar adequadamente; a indústria de software para *web* já considera os SR como um meio de atender melhor as demandas dos clientes e tornar seus sistemas mais lucrativos [12]. Nesse contexto, o processamento de dados textuais (notícias, livros, artigos científicos) representa uma demanda importante e esse tipo de dado é, frequentemente, objeto de processamento de SRbC [5, 7, 13, 30].

Aprendizado de Máquina (AM) é caracterizado pelo de-

senvolvimento de técnicas que objetivam prover os softwares com a habilidade de melhorar seu desempenho em uma tarefa aprendendo através da experiência (aprendizado indutivo) [15]. Diferentes técnicas de AM, ou combinações delas [3], têm sido usadas para análise de dados textuais em diversos contextos [6, 30]. O sucesso obtido nesses contextos mostra que tais técnicas são adequadas para o processamento de dados textuais e, portanto, são adequadas para a construção de SRs que lidam com este tipo de dado.

A presente Revisão Sistemática (RS) foi conduzida com o objetivo de construir um panorama sobre as pesquisas mais recentes na área de SRbC textual construídos com base em técnicas de AM. A RS é apresentada neste artigo como segue: Seção 2 apresenta os conceitos básicos de SRbC e AM; Seção 3 resume alguns trabalhos relacionados a esta RS; Seção 4 traz a metodologia de construção da RS; Seções 5 e 6 apresentam os resultados da revisão e uma discussão sobre eles; a conclusão é apresentada na Seção 7.

2. CONCEITOS FUNDAMENTAIS

Esta seção apresenta os conceitos básicos sobre SRbC e AM, para dar subsídios para o entendimento dos resultados e discussões apresentados das seções seguintes.

2.1 Sistemas de Recomendação baseados em Conteúdo

Para fazer uma recomendanção, os SRbC analisam o conteúdo dos itens candidatos à recomendação por meio da extração de atributos que os descrevem e analisam o perfil do usuário, que pode ser ou não representado por meio dos itens que eles preferem. Conhecendo o perfil dos usuários e o conteúdo referente aos itens, o sistema deve ser capaz de recomendar adequadamente novos itens aos usuários.

Formalmente, seja um conjunto de usuários $\mathcal{U} = \{U_1, \ldots, U_n, \ldots, U_N\}$, um conjunto de itens $\mathcal{I} = \{I_1, \ldots, I_m, \ldots, I_M\}$ e o histórico de interações $\mathcal{H}$, no qual uma interação $h \in \mathcal{H}$ é uma tripla $h = (U_n, I_m, \sigma)$ indicando que o usuário U_n interagiu com o item I_m e emitiu a avaliação σ sobre a experiência. Define-se $\mathcal{H}_{U_n} \subset \mathcal{H}$ como o histórico de interações do usuário U_n. O problema de recomendação com base no conteúdo pode ser visto como a definição de uma relação de similaridade $s : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{R}$ que permita aproximar uma função $l_s : 2^{\mathcal{H}} \rightarrow 2^{\mathcal{I}}$, de tal forma que a lista de recomendações gerada para o usuário U_n é dada por $L_{U_n} = l_s(\mathcal{H}_{U_n})$.

Uma arquitetura típica para SRbC, apresentada em [14], é composta por três módulos (Figura 1). O módulo de análise e representação de conteúdo é responsável por receber as descrições dos itens, pré-processá-las a fim de analisar o conteúdo dos mesmos e criar uma representação que possibilite o aprendizado dos perfis dos usuários e a elaboração das recomendações. O módulo de personalização usa tal representação e o histórico de interações do usuário ($\mathcal{H}_{u_n}$) para induzir a função $l_s(\mathcal{H}_{u_n})$. As avaliações da interação entre usuário e item pode ser explícita, em que ele avalia um item, ou implícita, na qual a execução da interação e suas características representam uma avaliação (o usuário leu uma notícia recomendada, ou permaneceu por muito tempo em uma página). O terceiro módulo é composto pelo componente de filtragem. Basicamente esse componente utiliza a saída da função $l_s(\mathcal{H}_{u_n})$ para gerar uma lista de recomendações (L_{u_n}), e a apresenta ao usuário. Tipicamente a lista de recomendações é ordenada com base em um *score* e os items mais relevantes (*top N*) são apresentados.

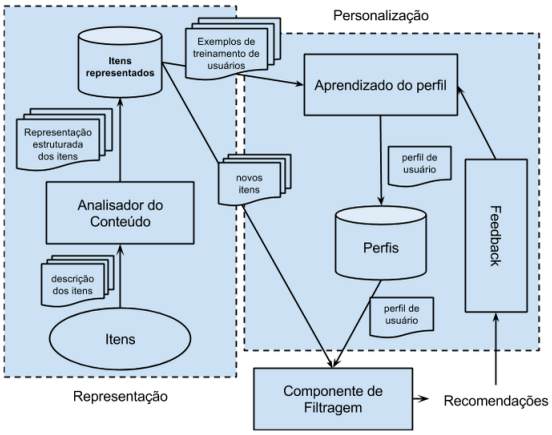

Figura 1: Arquitetura de um SRbC - adap. de [14]

Cada um dos tipos de SR tem vantagens e desvantagens. No caso de SRbC, as seguintes vantagens se destacam [1, 14]: o sistema não necessita de uma base de usuários, visto que suas recomendações são baseadas nas relações entre os itens; itens que nunca foram recomendados e avaliados por usuários (conhecido como problema de *cold-start* de item) podem aparecer na lista de recomendações, dado que é possível descobrir as suas relações com outros itens; e a recomendação de um item pode ser justificada, já que ela é baseada no conteúdo do item e na sua similaridade com itens já acessados pelo usuário. Já em termos de desvantagens, [1, 10, 14] destacam: a análise de conteúdo (textual ou multimídia) é uma tarefa complexa e exige o uso de técnicas sofisticadas para associar conteúdo a interesses de usuários no aprendizado dos perfis; dificuldade de recomendar itens que não pertençam a um mesmo domínio ou domínios similares (dentro do que o sistema aprendeu como perfil de usuário), mas que podem ser de interesse do usuário (conhecido como problema de serendipidade); e interesses de longo prazo para novos usuários podem não ser captados pelo sistema (conhecido como problema de *cold-start* de usuário), uma vez que para novos usuários há pouca informação sobre avaliações de itens.

Nesta RS são analisados estudos que implementam SRbC textual. O desenvolvimento desses SR possui um dificultador que é a necessidade de tratamento de um tipo de dado não estruturado. Isso implica na necessidade de uso de técnicas de pré-processamento de textos, técnicas de representação de texto estruturada, técnicas de redução de dimensionalidade e proposição de tópicos, além do que já é necessário para resolver o problema de recomendação. Detalhes sobre esses tópicos são encontrados em [22, 24, 26, 30].

2.2 Aprendizado de Máquina

A teoria de AM é baseada nos princípios do aprendizado indutivo (AI), ou seja, modelos sao determinados a partir de um conjunto de dados ou representações de experiências [18]. Normalmente, o aprendizado indutivo é implementado por algoritmos que processam um conjunto de dados e extraem um modelo capaz de explicar ou representar os dados sob algum aspecto. Esse modelo pode ser usado para explicar ou representar um novo dado (do mesmo domínio do conjunto de dados inicial), que é apresentado *a posteriori*.

O AI pode ser de três modalidades: supervisionado, não-

supervisionado e semi-supervisionado. Na primeira, os algoritmos ajustam parâmetros de um modelo a partir do erro medido entre respostas obtidas e esperadas. Na segunda, os parâmetros de um modelo são ajustados com base na maximização de medidas de qualidade das respostas obtidas. A terceira é caracterizada pelo uso de algoritmos híbridos, que fazem uso dos recursos de correção de erro e de maximização de medidas de qualidade, conforme necessário.

3. TRABALHOS RELACIONADOS

Os trabalhos relacionados a esta RS tratam de revisões de literatura (RL), contudo nenhum deles é apresentado na forma de RS. Assim, a fim de traçar um paralelo entre a presente RS e essas RLs, foi realizada uma análise livre sobre quais seriam os estudos primários abordados em cada revisão, e assim foi possível inferir o período abordado por elas. Essa análise partiu do pressuposto que os estudos primários estão principalmente descritos nas seções que tratam do tema abordado pela RL, excluindo as citações a estudos que fornecem a base teórica para entendimento do assunto. Por exemplo, em [17], a ênfase é dada em como aplicações de classificação de textos podem ser usadas em SRbC, então foram considerados como estudo primários aqueles que tratavam principalmente de aplicações de SRbC ou classificação de textos, e foram desconsiderados os trabalhos que davam base para entendimento de estratégias ou métodos.

O resultado da análise livre está na Tabela 1. Nela, a informação sobre número de estudos é apresentada de forma segmentada de acordo com o tema da RL, por exemplo, a revisão [14] apresenta 7 estudos no contexto de SRbC e 10 estudos no contexto de classificação de textos. Nessa tabela também são apresentados os dados referentes a esta RS, para fins de comparação. Os trabalhos relacionados são todos mais abrangentes que a presente RS em termos de tema. Nenhum deles é voltado para o estudo da aplicação de técnicas de AM em uma área específica em SR. Assim, um diferencial da presente RS é o estudo mais detalhado de uma linha de pesquisa dentro de SR, apresentando-se portanto como uma revisão menos abrangente, porém de maior profundidade.

A análise livre também permitiu observar que os estudos primários mais antigos abordados nas RL foram publicados na década de 1990, com exceção de [17] que inclui estudos sobre classificação de textos. Os autores de tais RL, portanto, tinham uma tendência para a apresentação de um quadro histórico dentro dos temas abordados. Diferentemente, a presente RS objetiva fornecer uma visão sobre as preocupações mais recentes existentes na área do tema abordado. A fim de apresentar os trabalhos relacionados com mais detalhes, segue aqui um breve resumo de cada um:

- [1] tem o objetivo de apresentar técnicas que são usadas em cada um dos tipos de SR em geral, assim como as estratégias para contornar as desvantagens de cada tipo de SR. Esta RL também propõem possíveis extensões e linhas para pesquisas futuras em SR.
- [13] tem o objetivo de investigar SRs no domínio de notícias, mostrando os principais problemas relacionados a este domínio. Os autores focam principalmente no problema de escalabilidade. Além disso, propõem estratégias para modelagem de perfil e como relacionar notícias com perfis. Também são realizados experimentos com as técnicas e estratégias estudadas.
- [4] tem como objetivo principal apresentar os SR no domínio de notícias, sumarizando os tipos de SR e as técnicas e estratégias neles utilizadas. O trabalho apresenta uma revisão dos tipos de SR apontando suas vantagens e desvantagens.
- [14] apresenta todos os passos referentes ao desenvolvimento de um SRbC, uma revisão sobre as técnicas utilizadas nos trabalhos, segmentando-as quanto à representação do conteúdo e aprendizado do perfil, e direções futuras para a evolução dos SRbC.
- [17] tem como objetivo principal investigar os algoritmos que são usados em SRbC, e que poderiam ser usados, como os algoritmos comumente usados na área de classificação de textos. O trabalho também discute formas para representar o conteúdo, o perfil do usuário e limitações em SRbC.

4. METODOLOGIA

Uma Revisão Sistemática (RS) permite identificar, interpretar e avaliar a pesquisa relevante para um tópico em particular [11]. Nessa estratégia, quando as contribuições individuais analisadas são chamadas de estudos primários, a RS se constitui como um estudo secundário. A diferença entre uma RS e uma revisão de literatura simples é que a primeira fornece condições de reprodutibilidade, e de avaliação do escopo abrangido na revisão e da qualidade dos estudos nela analisados. Além disso, a sistematização adotada para a elaboração de uma RS fornece condições para maximizar a possibilidade de recuperar um conjunto completo de dados e minimizar a possibilidade de um viés.

A presente RS foi conduzida com bases nas diretrizes propostas por [11], e é constituída das seguintes fases: (i) planejamento da RS; (ii) condução da RS; e (iii) resultados da RS. Nesta seção são detalhados os passos seguidos nas fases (i) e (ii). Os resultados e análises referentes à terceira fase são apresentados nas próximas seções (Seções 5 e 6).

4.1 Planejamento da Revisão Sistemática

O planejamento de uma RS inclui a identificação da justificativa para a revisão, as especificações das questões de pesquisa que se pretende responder e o desenvolvimento de um protocolo para escolha dos estudos que serão incluídos na revisão. Nesta seção, estes passos são explorados.

4.1.1 Justificativa para a Revisão Sistemática

Até onde foi possível identificar (Seção 3), não existem RS publicadas na área de AM aplicada a SRbC textual. Estratégias de AM possuem um alto potencial para fornecer conhecimento importante e útil para SR, e vêm sendo utilizadas para tal. Contudo, não foi encontrado um estudo que apresentasse os avanços que já foram obtidos e as questões em aberto em relação à aplicação de aprendizado supervisionado ou não-supervisionado nesse tipo de sistema.

SR têm sido objeto de interesse na comunidade científica e da indústria. Nos últimos anos, grandes empresas têm enveredado esforços para que soluções cada vez melhores sejam obtidas em termos de recomendação. Porém, ainda há problemas em aberto como otimização da serendipidade, *cold-start* de item e de usuário e tratamento de informação multimídia, nos quais técnicas de AM podem atuar com sucesso. Diante deste cenário, esta RS se justifica como um meio de, sistematicamente, organizar e analisar os mais novos resultados obtidos na área de SR a partir do uso de estratégias

Tabela 1: Comparação de RLs relacionas à presente RS

Ref.	Ano de publicação	Tema	Período	No. de Estudos Primários
[1]	2005	SR em geral	1993-2004	58
[4]	2010	SR, recomendação de notícias	1994-2007	10, 7
[13]	2011	Recomendação de notícias	1994-2011	20
[14]	2013	SRbC	1997-2008	44
[17]	2007	SRbC, classificação de texto	1961-2002	7, 10
Esta RS	2015	SRbC textual com AM	2012-2014	13

de AM. A RS apresentada aqui diz respeito aos estudos publicados nos três últimos anos (2012, 2013 e 2014).

4.1.2 Questões de pesquisa

Diante das justificativas delineadas na Seção 4.1.1 para elaboração desta RS, e seguindo o disposto em [11], um conjunto de questões de pesquisa foi elaborado.

- **Q1** Em quais módulos da construção de um SRbC textual se têm empregado técnicas de AM? *a.* Qual avaliação se faz dos resultados obtidos nessas iniciativas?
- **Q2** Quais tipos de AI em AM são aplicadas especificamente em SRbC textual? *a.* Quais são as vantagens e desvantagens de usar AI em SRbC textual?
- **Q3** Como informações externas ao conteúdo textual sob recomendação podem ser usadas pelas técnicas de AM para melhorar os resultados do sistema?

4.1.3 Protocolo da Revisão Sistemática

O protocolo dessa RS envolveu: (i) escolha de fontes de dados e estratégias de busca; (ii) definição de estratégias para seleção de estudos primários e avaliação de sua qualidade; e (iii) método de extração de dados dos estudos selecionados.

i. ***Fontes de dados e estratégias de busca*** – As fontes de dados escolhidas para serem usadas nesta RS foram: *Scopus*, *ISI Web of Science*, *IEEE Xplore*, *ACM Digital Library* e *SpringerLink*. Cinco fontes de dados foram usadas com o intuito de maximizar o número de estudos candidatos recuperados. Essas fonte de dados foram escolhidas ou por serem temáticas (indexam veículos de disseminação científica da área de Ciência da Computação, Engenharia e áreas correlatas) ou por apresentarem mecanismos eficientes de filtragem para obtenção de estudos de áreas interessantes para o objetivo desta RS. Além disso, são notadamente fontes de dados bastante conhecidas e utilizadas pelos pesquisadores potencialmente interessados no conteúdo desta RS.

A estratégia de busca foi estabelecida a partir de uma *string* genérica (Tabela 2) a ser aplicada à cada uma das fontes de dados. A *string* é uma combinação de palavras-chave que pretende maximizar a abrangência de recuperação e trazer estudos úteis para a composição das respostas para as questões de pesquisa. Assim, a *string* foi composta por termos relacionados apenas à área de aplicação objetivada nesta RS (Sistemas de Recomendação baseados em Conteúdo textual). Já os aspectos referentes à área de AM foram tratados nos critérios de inclusão, uma vez que inúmeros termos deveriam ser usados para abranger a área de AM, pois é frequente que a área de AM esteja presente em um artigo representada apenas pelo nome de um algoritmo.

ii. ***Estratégias de seleção de estudos primários e de avaliação da qualidade*** – Para a seleção dos estudos primários a serem analisados foram estabelecidos critérios de inclusão. Um estudo primário recuperado a partir da aplicação da *string* de busca foi incluído na RS se ele atendeu

Tabela 2: Expressão regular para a *string* de busca genérica aplicada aos mecanismos de busca considerando os campos de indexação: título do estudo, resumo, palavras-chave.

(“content [-]0,1 based filter*” **OR** “content [-]0,1 based recommend*”) **AND** (text* **OR** news **OR** article[s]0,1 **OR** paper[s]0,1 **OR** book[s]0,1)

a todos os critérios de inclusão definidos. Tais critérios têm o objetivo de garantir que os estudos incluídos possuem um nível de qualidade mínimo (CI-1), são primários (CI-2), são acessíveis (CI-3 e CI-4) e são de fato pertinentes ao escopo da RS (CI-5 a CI-7). Os critérios de inclusão aplicados foram:

- **CI-1**: o registro de dados identificado se refere a um estudo científico, publicado através de uma revisão por pares (o registro de dados não se refere a relatórios técnicos, livros, capítulos de livros, prefácios de anais ou editoriais de periódicos, pôsteres e *position papers*);
- **CI-2**: o trabalho é um estudo primário (não se trata de estudo secundário: revisão da literatura ou RS).
- **CI-3**: o trabalho está disponível na *web*;
- **CI-4**: o trabalho é apresentado na língua inglesa;
- **CI-5**: o trabalho é diretamente relacionado com as áreas de Ciência da Computação ou Sistemas de Informação (i.e, o trabalho não é relacionado unicamente com áreas como Marketing ou Ciências Sociais);
- **CI-6**: o trabalho trata, explicitamente, SR para conteúdo textual;
- **CI-7**: o trabalho faz referência explícita ao uso de técnicas de AM para modelagem e/ou implementação de algum módulo do SR.

Os critérios de inclusão foram avaliados a partir da análise dos metadados dos artigos, do título do artigo, do resumo e do texto na íntegra. Além disso, a abordagem *Tollgate* foi aplicada [2], conforme detalhado na Seção 4.1.4. Na abordagem *Tollgate*, diferentes pesquisadores aplicam os critérios de inclusão, e a decisão final sobre a inclusão de um artigo nos resultados da RS é tomada por voto majotário.

iii. ***Método de extração de dados dos estudos selecionados*** – A extração de dados se deu a partir da leitura, na íntegra, dos artigos selecionados para inclusão na RS. Para guiar a extração dos dados foi criado um formulário para anotação de: informações básicas como título, autores, veículo de publicação e ano de publicação; informações sobre o conteúdo do estudo em relação à área de AM e à area de SRs. Na Tabela 3 são listadas as informações sobre conteúdo anotadas nos formulários.

4.1.4 Condução da Revisão Sistemática

Esta seção apresenta a condução da revisão, a qual é dividida em duas etapas: (i) identificação e seleção dos estudos

Tabela 3: Informações sobre conteúdo extraídas dos estudos analisados nesta RS

Tópico	Descrição
Conjunto de Dados	Esse atributo contribui para o entendimento do contexto SRbC textual e dos experimentos realizados.
Análise e Repres. do Conteúdo	Verificação sobre se a análise e repres. do conteúdo foi tratada e se técnicas de AM foram aplicadas nesse tratamento.
Representação textual e pré-processamento	Para análise deste tópico foi verificado como se deu a transformação do conteúdo textual (não-estruturado) em uma representação estruturada.
Estratégia e técnicas de AM	Para esse tópico foi identificada qual o tipo de AI (supervisionada, semi-supervisionada ou não-supervisionada) usado no módulo analisador de conteúdo (Figura 1), e qual técnica foi usada para a implementação do aprendizado.
Personalização	Verificação sobre se a fase de personalização foi tratada e se técnicas de AM foram aplicadas nesse tratamento.
Representação do perfil	Para análise deste tópico foi verificado como se deu a representação do perfil do usuário.
Estratégia e técnicas de AM	Neste quesito foi identificada qual o tipo de AI (supervisionada, semi-supervisionada ou não-supervisionada) usado no Aprendizado do Perfil (Figura 1), e qual técnica foi usada para a implementação do aprendizado.
Recomendações	Verificação sobre se a fase de recomendação foi tratada e se técnicas de AM foram aplicadas nesse tratamento.
Estratégia e técnicas de AM	Neste tópico foi identificado qual o tipo de AI (supervisionada, semi-supervisionada ou não-supervisionada) usado no Componente de Filtragem (Figura 1), e qual técnica foi usada para a implementação do aprendizado.

primários e (ii) extração e síntese dos dados desses estudos.

i. ***Identificação e seleção dos estudos primários*** – Para a identificação dos estudos primários, a *string* de busca genérica foi adaptada e executada em cada uma das fontes de dados. A aplicação foi realizada no período de agosto a setembro de 2014. As bases de dados retornaram 97 estudos (para o cômputo desse número, artigos indexados em mais de uma das bases de dados foram contabilizados apenas uma vez) publicados nos anos de 2012, 2013 e 2014. Os critérios de inclusão foram aplicados seguindo os passos: as listas de registros retornados das buscas foram avaliadas para aplicação do critério CI-1; os resumos dos registros restantes foram avaliados e os critérios CI-2 e CI-5 ao CI-7 foram aplicados; os estudos referentes aos registros restantes foram avaliados segundo os critérios CI-3 e CI-4; os estudos restantes foram lidos na íntegra para extração de dados e novamente os critérios CI-6 ao CI-7 foram avaliados. Os critérios CI-1 ao CI-5 foram objeto de atenção de um pesquisador envolvido na condução da RS. Já a primeira aplicação dos critérios CI-6 ao CI-7 foi realizada por três pesquisadores usando uma abordagem *Tollgate*. Ao fim do consenso na abordagem *Tollgate*, a seleção final de artigos a serem incluídos na RS foi realizada. A segunda aplicação dos critérios CI-6 ao CI-7 foi realizada por um pesquisador.

ii. ***Extração e síntese de dados*** – A extração e síntese seguiram o preenchimento de formulários; os resultados estão consolidados e discutidos nas seções 5 e 6.

5. RESULTADOS

Nesta seção são apresentados os resultados da condução do protocolo. A aplicação dos critérios de inclusão resultou na seleção de 13 estudos (13% dos resultados recuperados), sendo que 6 foram publicados em 2012, 5 em 2013 e 2 em 2014. Embora esse resultado indique um uso ainda pequeno de AM em SRbC textual, é preciso considerar a seleção dos artigos de forma mais refinada: 51 estudos (52%) foram considerados atender ao critério CI-6, e 26 (27%) ao critério CI-7. Nesse universo reduzido, a área de AM se mostra um pouco mais presente nas soluções para SRbC textual.

Análises gerais sobre o conjunto de estudos selecionados são apresentas nas Tabelas 4, 5 e 6. Na primeira análise (Tabela 4), os estudos foram classificados quanto à sua relação com as questões de pesquisa, indicando quais estudos contribuiram para a formulação das respostas. A segunda análise (Tabela 5) teve o objetivo de organizar os 13 estudos em relação aos aspectos: em qual módulo de um SRbC a técnica de AM é aplicada; e que tipo de aprendizado é considerado para propor as soluções para os problemas encontrados na concepção desses módulos. Essa análise mostrou uma concentração de estudos que aplicam aprendizado supervisionado no módulo de aprendizado do perfil (8 estudos) e aprendizado não-supervisionado no módulo de representação do conteúdo (7 estudos), sendo que desses últimos, 6 estão preocupados em tratar o problema de redução de dimensionalidade. A terceira análise (Tabela 6) apresenta uma visão sobre os domínios de recomendação sob os quais os SRbC textual construídos com AM têm sido investigados.

Tabela 4: Estudos X questões de pesquisa (QP)

QP	Estudos	#
Q1	[5, 7, 8, 9, 16, 19, 26, 27, 28, 29, 30, 31]	12
Q2	[5, 7, 8, 9, 16, 19, 23, 26, 27, 28, 29, 30, 31]	13
Q3	[5, 7, 16, 19, 23, 29, 31]	7

Tabela 5: Estudos X módulos de um SRbC textual X tipos de AI

Módulo	Aprendizagem		
	sup.	não sup.	semi sup.
Representação de conteúdo	–	[5, 7, 26, 27, 28, 30, 31]	–
Aprendizado de perfil	[7, 8, 9, 16, 19, 23, 27, 29]	–	–
Filtragem	[31]	–	–

Tabela 6: Estudos X domínio de recomendação (DR)

DR	Estudos	DR	Estudos
Notícias	[5, 7, 16, 29]	Artigos	[27, 28]
Ofertas de empregos	[8]	Especialista	[26]
Conteúdo *web*	[9]	Micro-blogs	[30]
Filmes	[19]	Lugares	[23]
e-Commerce	[31]		

Em um trabalho de análise mais específico, cada um dos 13 estudos foi detalhadamente analisado para identificação das soluções propostas para construção dos módulos de um SRbC (conforme Figura 1). Os resultados dessa análise estão organizados por módulo nas três próximas seções.

5.1 Análise e Representação de Conteúdo

Nos estudos analisados, os módulos de Análise e Representação de Conteúdo dos SRbC textual recebem os itens tex-

tuais candidatos à recomendação, geralmente sob uma representação não estruturada, os submetem a um pré-processamento de textos, e os direcionam a técnicas de AM que geram uma representação estruturada para os próximos módulos.

Evidentemente, a fase de pré-processamento de textos está presente nos 13 estudos analisados. De forma geral, os autores usam o Vector Space Model[1] (VSM) [22] em alguma fase do processo. Porém, outros recursos também são usados em alguns estudos para estabelecimento das dimensões, como por exemplo: uso de hiperônimos do dicionário *WordNet*, *POS tag* (*Part-of-speech tagging*) para a extração de substantivos, n-gramas [5]; extração de *tags* relacionadas aos itens sob recomendação [16]. Uma vez escolhidas as dimensões, é preciso atribuir valores a elas, e para isso são empregadas diferentes estratégias: a *binária* [9, 31], em que é atribuído 1 se o valor de uma dimensão ocorre no documento ou 0 se não ocorre; a *frequência de termos* (TF) [5, 7, 8, 16, 19, 23, 26, 27, 28, 29], que considera o número de vezes que o valor de uma dimensão ocorre no documento; e a clássica normalização *TF-IDF* [5, 7, 8, 9, 19, 26, 29], na qual a frequência é relativa em relação ao corpus de documentos. Também na etapa de pré-processamento, são utilizadas técnicas simples para simplificar ou reduzir a dimensionalidade dos vetores de representação, como filtro de *stopwords*, que são palavras que não adicionam informação útil ao VSM [5, 8, 9, 16, 19, 27, 31]; *stemming*, que reduz cada palavra para o seu radical [5, 8, 9, 19, 27, 31]; e filtragem de termos abaixo ou acima de um limiar de frequência [5, 8].

Ainda com o intuito de gerar um modelo de representação estruturada para uso dos próximos módulos do sistema, alguns estudos aplicam técnicas de AM não-supervisionado (veja Tabela 5). Tratam-se, na realidade, de técnicas de redução de dimensionalidade que induzem modelos de representação a partir do conjunto de itens, por meio da implementação de diferentes algoritmos: determinação de variáveis latentes [7]; *Latent Semantic Analysis* (LSA) [26], *Latent Dirichlet Allocation* (LDA) [27, 30], otimização gradiente descendente [31] e *random walks* [28]. Esses algoritmos reduzem a dimensionalidade dos vetores no VSM, criando uma representação compacta dos itens. O desafio nesse contexto é descobrir o número ideal de dimensões a serem usadas na representação. A avaliação sobre o número ideal é geralmente realizada de forma empírica, por meio da realização de testes variando o número de fatores e analisando a qualidade das recomendações geradas ao final do processamento executado pelos SRs [7, 26, 27, 30].

Em [5], os autores aplicam técnicas de agrupamento sobre a representação vetorial original gerando um modelo de representação baseado em grupos de notícias. Nessa estratégia, o algoritmos *k-means* e uma variação dele aplicada a hiperônimos do dicionário *WordNet* geram grupos de notícias que suportam o fornecimento de recomendações extras. Nesse caso, o SR gera recomendações iniciais que são complementadas por outras notícias presentes nos grupos das notícias inicialmente recomendadas. Essa proposta é interessante para minimizar o problema de *cold-start* de usuário.

SRbCs podem usar também, na construção da representação do conteúdo, informação proveniente de outras fontes de dados. Estratégias nessa linha são discutidas nos estudos [7, 30, 31]. Em [7], os autores adicionam informações sobre a categoria da notícia, influenciando no cálculo das probabilidades relacionadas às variáveis latentes. Nos demais artigos, embora acrescentar atributos descritivos externos ao conteúdo influencie na qualidade das recomendações geradas ao final do processo, a maneira como as técnicas de AM são aplicadas para construção do VSM não é alterada pela presença ou ausência de tais atributos.

[1]No VSM, as dimensões do vetor que representa o item (ou documento) são compostas por palavras nele presentes.

5.2 Personalização

O módulo de Personalização, onde ocorre o aprendizado do perfil do usuário, é implementado com técnicas de AM, tipo supervisionadas, em 8 estudos (Tabela 5). Nele, o perfil do usuário pode ser representado no mesmo espaço (VSM) que o item sob recomendação, por meio do uso de termos que ocorrem em documentos que eles acessam, como ocorre nos estudos [7, 8, 9, 16]. Contudo, alguns autores propõem algumas representações alternativas: em [29] os autores incorporam informações demográficas na tentativa de minimizar o problema de *cold-start* de usuário; em [8] informações referentes ao usuários, provenientes de redes sociais, são utilizadas; e, finalmente, em [23], o enriquecimento do perfil do usuário é realizado com a aplicação de AM, representado no referido estudo por *Hidden Markov Models* e Árvores de Decisão, para extração de informações referentes a localização do usuário, malha de transporte e locais de visitação.

Nos estudos que aplicaram AM para personalização foram construídos classificadores que dado um item (como entrada), o classifica como adequado para o usuário ou não. Usualmente um classificador é treinado para cada usuário com base no histórico de iterações entre o usuário e os itens. Experiências positivas entre usuário e item determinam a classe positiva (o item é adequado para o usuário), e experiências negativas determinam a classe negativa. Diversos algoritmos de AM supervisionado são usados nessa estratégia: modelos probabilísticos [7, 29], *Support Vector Machine* (SVM) [8, 9, 27], classificador de Rocchio [8, 9], *k-nearest neighbor* (k-NN) [19] e Regressão Logística (RL) [16].

Em [7] e [29], os autores usam os modelos probabilísticos *Expectation-Maximization* para geração de modelos generativos, e Redes Bayesianas, respectivamente, para indicar se um usuário leria ou não uma notícia. Experimentos em [7] mostraram que a estratégia capta interesses de curto e longo prazo, usando conjunto de dados estratificados de acordo com a variável de tempo, e avaliando o resultado com uso das métricas *P@k* e *Normalized Discounted Cumulative Gain*, as quais analisam a posição em que a notícia apareceu na lista de recomendação. Ainda no contexto de notícias, em [16], os interesses do usuário são modelados com RL.

Classificadores de Rocchio e SVM são usados em [8] e [9], seguindo a estratégia usual de construção de classificadores em SRs. Em ambos os casos, experimentos avaliados sob as clássicas métricas baseadas em matriz de confusão (área sob a curva ROC em [8] e precisão, revocação e *f-score* em [9]) indicaram a superioridade da técnica SVM.

SVMs também são usados em [27] para decidir se um artigo é relevante ou não para os usuários do SR, dada a *string* de busca que o usuário entra no sistema. Ainda, [19] aplica como base para recomendação, uma tarefa de regressão que faz o uso do algoritmo k-NN para predizer a nota que o usuário pode dar a um filme.

5.3 Recomendação

O módulo de Recomendação recebe a lista de recomendações gerada no módulo de Personalização e a disponibiliza

ao usuário, refinando-a se necessário. Apenas um estudo [31] realiza processamento de informação nesse módulo usando uma técnica de AM, a k-NN. Em tal proposta, os autores aplicam k-NN na representação do itens, sem que o perfil do usuário seja construído. Os k itens mais similares a um dado item (acesso pelo usuário) entram na lista de recomendação.

6. DISCUSSÕES

Nesta seção são apresentadas respostas para as questões de pesquisa elaboradas para essa revisão sistemática.

Q1. *Em quais módulos da construção de um SRbC textual se têm empregado técnicas de AM?* – A análise dos estudos selecionados nessa revisão sistemática permitiram concluir que, considerando a arquitetura típica de um SRbC (Figura 1), AM oferece recursos para resolução de problemas nos três módulos: de Análise e Representação de Conteúdo, de Personalização e de Recomendação. No caso do primeiro módulo, as técnicas de AM são predominantemente aplicadas com o objetivo de redução de dimensionalidade na representação dos itens (sob recomendação). No segundo módulo, Aprendizado de Perfil, as análises evidenciam que há potencial para aplicação de AM em, pelo menos, duas frentes: para enriquecimento da representação do perfil do usuário; e para aprendizado do perfil. Finalmente, no terceiro módulo, a aplicação de AM foi observada em apenas um estudo, no entanto, trata-se de fato de um módulo bastante simplificado na arquitetura. É interessante observar que a estratégia usada para aplicar AM no terceiro módulo substituiu a necessidade de implementação do segundo módulo (sem aparente prejuízo de desempenho), possibilitando a otimização da arquitetura do sistema.

Q1.a *Qual avaliação se faz dos resultados obtidos nessas iniciativas?* – Todos os estudos apresentaram experimentos cujos resultados foram avaliados por meio de medidas classicamente usadas na área de AM. Os resultados obtidos estão dentro do que se considera resultados aceitáveis na área. Além disso, os experimentos apresentados na maior parte dos artigos também fizeram avaliações empíricas, comparando os resultados das recomendações com dados históricos referentes às interações dos usuários com os itens recomendados. Tais comparações também evidenciaram a adequabilidade das técnicas de AM ou sua superioridade.

Q2. *Quais tipos de AI em AM são aplicadas especificamente em SRbC textual?* – A aplicação de AI não-supervisionados foi observada na implementação de otimizações no módulo de análise e representação de conteúdo. Já a aplicação do AI supervisionado foi observada nos outros dois módulos. Não foi observada o uso de AI semi-supervisionado. Ainda, dentro dos AI não-supervisionado, apenas um estudo utilizou técnicas de quantização do espaço (como o *k-means*). Naturalmente, o tipo de tarefas que parecem ser interessantes para cada um dos módulo levou a esse viés. O AI supervisionado é mais facilmente aplicado no contexto onde se tem uma base de dados histórica que pode ser rotulada, como é caso da tarefa de aprendizado de perfil. Geralmente, há informação sobre que item foi, no passado, acessado por qual usuário, o que permite supervisionar a indução de um modelo que represente essa relação.

Q2.a *Quais são as vantagens e desvantagens de usar AI em SRbC textual?* – A implementação da redução de dimensionalidade na representação dos itens levou a melhorias de desempenho em alguns estudos [7, 26, 27, 28, 30, 31], principalmente por possibilitar a proposição de tópicos para a representação. Entretanto, as técnicas que se baseiam em *Singular Value Decomposition*, como LSA, não permite a interpretação dos fatores latentes, que representariam os tópicos. Os estudos que implementaram classificadores para o aprendizado do perfil do usuário alcançaram bom desempenho, principalmente no uso de SVM. Contudo, seguindo as abordagens propostas nos artigos, pode-se cair em problemas de escalabilidade devido à necessidade da construção de um classificador por usuário. Além disso, a indução desses classificadores geralmente exige uma grande quantidade de exemplos rotulados, o que nem sempre é factível.

Q3. *Como informações externas ao conteúdo textual sob recomendação podem ser usadas pelas técnicas de AM para melhorar os resultados do sistema?* – O uso de informações externas ao conteúdo textual foi observada em sete estudos [5, 7, 16, 19, 23, 29, 31]. Em [7] e [23], informações externas estão diretamente relacionadas à forma como as técnicas de AM são utilizadas. No primeiro caso, probabilidades usadas na construção da representação do conteúdo são influenciadas pelo uso de informações sobre as classes das notícias, e no segundo caso, as técnicas de AM são usadas para obtenção da informação externa. Ainda em [7], informação referente à data de publicação de notícias é usada para estratificar o conjunto de dados de treinamento para indução da recomendação, porém, a forma como a técnica de AM é usada não é alterada. Isso também ocorre nos demais estudos que inserem informações externas como atributos descritivos dos itens de recomendação: [5, 16] complementam a representação com informações de ontologias e de artigos do *wikipedia*, respectivamente; [19, 31] utilizam outros tipos de informações, como imagem e som. Informações externas também são usadas para complementar a representação do perfil dos usuários, [29] aplicam informações demográficas dos usuários para complementar a representação de seus perfis. Melhorias de desempenho observadas em tais estudos são obtidas porque informações externas aumentam o poder descritivo da representação, e não por estarem diretamente ligadas à maneira como as técnicas de AM são aplicadas. Não foi encontrado nenhum estudo que fizesse uso de qualquer conhecimento externo ao conteúdo textual que pudesse ser utilizado pelas técnicas de AM para obter vantagens específicas, como por exemplo, determinar número de grupos ou número de variáveis latentes, ou ainda permitir o uso de estratégias de *boosting* para o algoritmo de aprendizado.

7. CONCLUSÃO

Neste artigo foi apresentada uma Revisão Sistemática sobre a intersecção das áreas: SRbC textual e AM. Com a condução da revisão foi observado que iniciativas usando AM em SRbC textual ainda são em pequeno número. Entretanto, na análise dos estudos selecionados percebeu-se que as iniciativas na interseção dessas áreas têm produzido resultados promissores, fornecendo uma motivação para que mais esforços de pesquisa sejam empregados nessa área. Nesse contexto, foi ainda possível perceber que há uma carência de iniciativas voltadas para estudar a adequabilidade de AI não-supervisionado na tarefa de Aprendizado do Perfil do usuário. Técnicas de aprendizado não-supervisionado podem contribuir nesta tarefa resolvendo desvantagens encontradas no uso de AI supervisionado (escalabilidade e necessi-

dade de dados rotulados ou *cold-start* de usuário), ou produzindo conhecimento inesperado em uma recomendação, minimizando o problema de serendipidade, comum em SRbC.

8. REFERÊNCIAS

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. 17(6):734–749, 2005.

[2] W. Afzal, R. Torkar, and R. Feldt. A systematic review of search-based testing for non-functional system properties. 51(6):957–976, 2009.

[3] D. Bell, J. Guan, and Y. Bi. On combining classifier mass functions for text categorization. 17(10):1307–1319, Oct 2005.

[4] H. L. Borges and A. C. Lorena. A survey on recommender systems for news data. In E. Szczerbicki and N. Nguyen, editors, *Smart Inf. and Knowledge Management*, volume 260 of *Studies in Comput. Int.*, pages 129–151. Springer, 2010.

[5] C. Bouras and V. Tsogkas. Assisting cluster coherency via n-grams and clustering as a tool to deal with the new user problem. pages 1–14, 2014.

[6] D. Cai and X. He. Manifold adaptive experimental design for text categorization. 24(4):707–719, April 2012.

[7] S. Cleger-Tamayo, J. M. Fernández-Luna, and J. F. Huete. Top-n news recommendations in digital newspapers. 27:180–189, 2012.

[8] M. Diaby, E. Viennet, and T. Launay. Exploration of methodologies to improve job recommender systems on social networks. 4(1), 2014.

[9] D. Godoy. Comparing one-class classification algorithms for finding interesting resources in social bookmarking systems. In *Resource Discovery*, volume 6799 of *LNCS*, pages 88–103. Springer, 2012.

[10] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems An Introduction*. Cambridge Univ. Press, 2011.

[11] B. Kitchenham. Guidelines for performing systematic literature reviews in software engeneering. Technical Report EBSE-2007-01, Keele Univ., UK, 2007.

[12] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. 42(8):30–37, Aug. 2009.

[13] L. Li, D. Wang, S. Zhu, and T. Li. Personalized news recommendation: A review and an experimental investigation. 26(5):754–766, 2011.

[14] P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 73–105. Springer, 2011.

[15] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[16] F. Narducci, C. Musto, G. Semeraro, P. Lops, and M. de Gemmis. Exploiting big data for enhanced representations in content-based recommender systems. 152:182–193, 2013.

[17] M. J. Pazzani and D. Billsus. The adaptive web. chapter Content-based Recommendation Systems, pages 325–341. Springer, Berlin, Heidelberg, 2007.

[18] S. M. Peres, T. Rocha, M. R. C. B. Bíscaro, H. H., and C. Boscarioli. Tutorial sobre fuzzy-c-means e fuzzy learning vector quantizations: Abordagens híbridas para tarefas de agrupamento e classificação. 19(1):120–163, 2012.

[19] W. Qu, K.-S. Song, Y.-F. Zhang, S. Feng, D.-L. Wang, and G. Yu. A novel approach based on multi-view content analysis and semi-supervised enrichment for movie recommendation. 28(5):776–787, 2013.

[20] P. Resnick and H. R. Varian. Recommender systems. 40:56–58, March 1997.

[21] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 1–35. Springer, 2011.

[22] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. 18(11):613–620, 1975.

[23] S. Savage, M. Baranski, N. E. Chavez, and T. Hollerer. I'm feeling loco: A location based context aware recommendation system. In *Advances in Location-Based Services: 8th International Symposium on Location-Based Services, Vienna 2011*, LNGC. Springer, 2011.

[24] F. Sebastiani. Machine learning in automated text categorization. 34(1):1–47, 2002.

[25] S. Senecal and J. Nantel. The influence of online product recommendations on consumers' online choices. 80(2):159–169, 2004.

[26] A. Spaeth and M. Desmarais. Combining collaborative filtering and text similarity for expert profile recommendations in social websites. In *User Modeling, Adaptation, and Personalization*, volume 7899 of *LNCS*, pages 178–189. Springer, 2013.

[27] S. Tantanasiriwong. A comparison of clustering algorithms in article recommendation system. In *Proc. of SPIE - The Int. Soc. for Optical Eng.*, volume 8349, Singapore, 2012.

[28] Y. Wang, J. Liu, X. Dong, T. Liu, and Y. Huang. Personalized paper recommendation based on user historical behavior. In M. Zhou, G. Zhou, D. Zhao, Q. Liu, and L. Zou, editors, *Natural Language Processing and Chinese Comp.*, volume 333 of *Commun. in Comp. and Inf. Sci.*, pages 1–12. Springer, 2012.

[29] K. F. Yeung and Y. Yang. A proactive personalized mobile news recommendation system. In *J. of Internet Services Applications (2012)*, pages 207–212. IEEE, Sept. 2012.

[30] J. Yu, Y. Shen, and J. Xie. Mining user interest and its evolution for recommendation on the micro-blogging system. In J. Wang, H. Xiong, Y. Ishikawa, J. Xu, and J. Zhou, editors, *Web-Age Information Management*, volume 7923 of *LNCS*, pages 679–690. Springer, 2013.

[31] Q. Zhu, Z. Li, H. Wang, Y. Yang, and M.-L. Shyu. Multimodal sparse linear integration for content-based item recommendation. In *Multimedia (ISM), 2013 IEEE Int. Symp. on*, pages 187–194, Dec 2013.