iRev: Um framework de avaliação de sistemas de recomendação baseados comentários textuais

Guilherme Bittencourt
bittenkourt.gmf@aluno.ufsj.edu.br
UFSJ - Minas Gerais - Brasil

Naan Vasconcelos
naan.vasconcelos@aluno.ufsj.edu.br
UFSJ - Minas Gerais - Brasil

Leonardo Rocha
lcrocha@ufsj.edu.br
UFSJ - Minas Gerais - Brasil

ABSTRACT

Current advances in Recommendation Systems and Natural Language Processing have motivated recent studies to return their interest in Review-Aware Recommendation Systems (RARSs). In this sense, we employ a systematic mapping approach by selecting 117 papers published on the main vehicles of the area, presenting a summary of the advances, highlighting the main proposal algorithms, and detailing the most used datasets and metrics in experimental setups. All the implementations and other artifacts extracted from this study were consolidated into a framework: iREV. In addition, we conduct a comprehensive experimental comparison among state-of-the-art proposals, highlighting the main directions and new perspectives for future developments.

KEYWORDS

Sistemas de recomendação; Comentários textuais de usuários

1 INTRODUÇÃO

Os sistemas de recomendação (SsR) surgiram como uma estratégia eficaz para lidar com a sobrecarga de informações. Nos últimos anos, a literatura tem testemunhado a proposta de inúmeras técnicas de recomendação, buscando constantemente melhorar a eficácia destes métodos. Embora muitas abordagens tenham se concentrado em técnicas que empregam avaliações quantitativas (numéricas), é essencial reconhecer o valor significativo do *feedback do usuário*, geralmente expresso como comentários (textuais) (também conhecidos como revisões), a fim de compreender suas preferências. Neste contexto, vários algoritmos foram desenvolvidos para aproveitar efetivamente os comentários como uma valiosa fonte de informação, conhecidos como Sistemas de Recomendação Review-Aware (RARSs), capazes de gerar recomendações a partir dos comentários dos usuários.

Como primeira contribuição, esse trabalho apresenta um mapeamento sistemático dos estudos sobre sistemas de recomendação *review-aware* com dois objetivos principais: (i) consolidar uma imagem atualizada das principais pesquisas realizadas nessa área recentemente para futuros trabalhos; (ii) destacar as principais limitações, características e orientações que estamos seguindo como comunidade. Identificamos 117 estudos relevantes sobre recomendação *review-aware* publicados nos principais veículos da área (e.g., RecSys, SIGIR, etc.) de 2014 a 2023. Realizamos um estudo detalhado dos principais avanços, dos principais conjuntos de dados e das métricas utilizadas. Observamos que as coleções de dados mais utilizadas dentre os trabalhos relevantes são as bases de dados de produtos da Amazon e de pontos de interesse da Yelp, por disponibilizarem as interações usuário/item e também os comentários provenientes das interações. Em relação às métricas de avaliação, é possível observar que métricas de erro são as formas de metrificação mais utilizadas pelos trabalhos, seguidas pelas métricas de avaliação de ranking. Porém, apesar do consenso na comunidade de recomendação de que é necessário mais do que precisão para avaliar a eficácia dos SsR, a grande maioria dos trabalhos priorizam a precisão sobre outras dimensões de qualidade, tais como serendipidade e diversidade. Outra limitação surge em relação aos algoritmos que são considerados estado-da-arte e suas configurações. Não existe um consenso das linhas de bases a serem consideradas, cada artigo utiliza um conjunto distinto e as configurações dos parâmetros raramente são reportadas. Menos de 50% dos trabalhos analisados disponibiliza código fonte de suas propostas, e menos de 30% fornece as configurações de parâmetros dos algoritmos propostos.

Nesse sentido, como segunda contribuição, implementamos os 10 principais algoritmos de recomendação *review-aware*, consolidando todos os códigos fontes gerados, bem como todos os artefatos levantados durante o mapeamento sistemático (métricas e bases de dados) em um framework aberto e publicamente disponível[1], denominado iRev, com o objetivo de facilitar a pesquisa e comparações entre abordagens na área de *review aware*. Por fim, como terceira contribuição, realizamos uma análise experimental de diferentes algoritmos, com diversas coleções e métricas, destacando as principais direções para desenvolvimentos futuros.

Todas as implementações, execuções e avaliações dos resultados foram realizadas pelo aluno Guilherme Bittencourt, com auxílio do aluno Naan Vasconcelos, sob a orientação do professor Leonardo Rocha.

2 MAPEAMENTO SISTEMÁTICO

Nesta seção abordamos o processo de coleta, filtragem e seleção dos artigos relacionados a RARSs.

2.1 Fase 1: Questões de pesquisa, palavras de busca e fontes digitais

As questões de pesquisa que deverão ser respondidas são:

- **QP1:** Como os algoritmos de recomendação utilizam técnicas de PLN para definir as preferências dos usuários por meio de seus comentários?
- **QP2:** Quais são as bases de dados e métricas mais prevalentes utilizadas na avaliação de algoritmos em estudos relacionados a esse tema?
- **QP3:** Como as avaliações experimentais são conduzidas nos estudos analisados, considerando estados da arte, configurações e parâmetros dos modelos?

[1] https://github.com/guibitten03/iRev

In: IV Concurso de Trabalhos de Iniciação Científica (CTIC 2024). Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (CTIC'2024). Juiz de Fora/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2596-1683

Recorremos ao mecanismo de pesquisa do Google Scholar para realizar as três consultas abaixo:

- **SS-Q1:** ("review based" OR "review aware" OR "review modeling") AND ("recommender systems" OR "recommendation systems" OR "recommender system") AND ("text" OR "textual" OR "review")
- **SS-Q2:** ("review based" OR "review aware" OR "review modeling") AND ("recommender systems" OR "recommendation systems" OR "recommender system") AND ("evaluation" OR "measure" OR "metrics")
- **SS-Q3:** ("review based" OR "review aware" OR "review modeling") AND ("recommender systems" OR "recommendation systems" OR "recommender system") AND ("source code" OR "reproducibility" OR "empirical" OR "experimental")

2.2 Fase 2: Seleção de trabalhos relevantes

Após a coleta, aplicamos um filtro de data entre 2014 e 2023, assegurando um escopo de 10 anos de publicações, acumulando um total de 1.190 artigos. Eliminamos as duplicadas e aplicamos um segundo filtro considerando apenas artigos das 100 conferências de maior fator de impacto de acordo com a research.com), tais como RecSys, WWW, WSDM, SIGIR, etc., resultando em 681 artigos. Por fim, empregamos um filtro avançado com critérios de inclusão e exclusão. Realizamos uma análise manual de cada artigo, classificando-os como relevantes ou não, de acordo com os critérios estabelecidos:

Critérios de Inclusão

- O método principal empregado para realizar as recomendações é o uso dos comentários dos usuários, considerando as avaliações numéricas como um suporte adicional.
- Propõem avanços e inovações no domínio, não se limitando à otimização de algoritmos preexistentes.
- Realizam avaliações experimentais comparativas entre os algoritmos que utilizam comentários do usuário e o método proposto nos artigos correspondentes.

Critérios de Exclusão

- Além dos comentários, empregam outras fontes de informação, como imagens, áudios ou vídeos para prever as preferências do usuário.
- São surveys, casos de estudo, revisões sistemáticas ou experimentais sobre os algoritmos do cenário.
- Utilizam os comentários exclusivamente para justificar as recomendações, focando na explicabilidade..

Após a aplicação desses critérios, restaram 117 artigos que foram identificados como mais relevantes.

2.3 Fase 3: Extração das informações dos artigos

Realizamos uma leitura detalhada dos 117 artigos restantes para identificar as principais características das soluções propostas, suas principais inovações e contribuições para a literatura e as metodologias de avaliação utilizadas. A seguir, detalhamos o resultado dessa análise visando responder as três questões de pesquisas levantadas no início desta seção.

3 AVALIAÇÃO SISTEMÁTICA

Essa seção detalha como os SsR vêm sendo avaliados por meio de uma inspeção das avaliações experimentais dos 117 artigos selecionados.

3.1 Bases de dados

Conforme podemos observar na Figura 1, que as bases de dados da *Amazon* e da *Yelp* são as mais utilizadas. Ambas tratam de cenários clássicos para análises de recomendação *review aware* devido ao grande número de usuários que comentam sobre os itens. A Amazon é composta de subcoleções de acordo com a categoria de item e a Yelp por diferentes cidades. A grande maioria dos trabalhos não especifica qual categoria/cidade utilizada nos experimentos. Outra questão crítica é que essas coleções têm cortes temporais que também não são mencionados. Essas questões impactam negativamente na reprodutibilidade desses trabalhos.

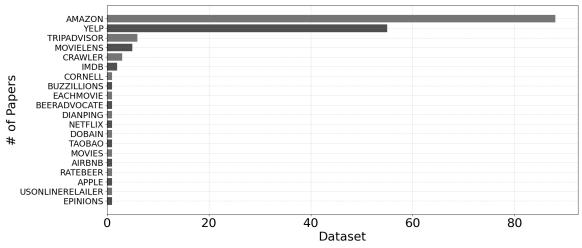

Figure 1: Frequência de bases de dados em experimentações.

3.2 Métricas

Conforme podemos observar na Figura 2, há um constante interesse sob a precisão dos *ratings* preditos. O consenso na comunidade de SR é que a precisão por si só não é suficiente para avaliar a eficácia prática e o valor agregado das recomendações, sendo necessário outras técnicas de avaliação como diversidade e serendipidade. Não identificamos nenhum trabalho dentre os 117 analisado que busca avaliar os modelos nessas dimensões, sendo esse um importante ponto fraco que avaliações futuras precisam considerar.

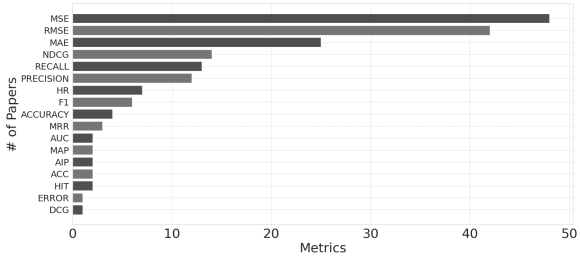

Figure 2: Frequência de métricas em experimentações.

3.3 Algoritmos e Configuração de Parâmetros

Com respeito a disponibilidade do código fonte dos algoritmos propostos, temos que pelo menos 50% dos trabalhos disponibiliza código fonte, o que dificulta a utilização dos mesmos como linhas de base. Um novo algoritmo é comparado, em média, com apenas três outros algoritmos e, no máximo, nove. Outra observação importante é sobre o processo de calibração dos algoritmos, em que menos de 30% dos trabalhos apresentam em detalhes desse processo. Todas essas questões impacta na replicabilidade dos trabalhos.

4 IREV

Nessa seção detalhamos nossa proposta de um framework de avaliação de SsR baseados comentários textuais: iRev (disponível em https://github.com/guibitten03/iRevRS) .

4.1 Algoritmos Implementados

Dos 117 algoritmos examinados, selecionamos todas as abordagens utilizadas em pelo menos dois artigos diferentes. Os 10 algoritmos selecionados são apresentados na Tabela 1, onde a coluna *'Linhas de Base'* representa quantas vezes o algoritmo foi utilizado em outros trabalhos.

Algoritmo	# Linhas de Base	# Citações
DeepCoNN	28	908
Narre	14	455
D-ATTN	9	421
Daml	7	123
MPCN	6	268
CARL	4	163
ANR	3	126
CARP	2	97
HRDR	2	75
RGNN	2	20

Table 1: Algoritmos mais utilizados como linhas de base.

O DeepCoNN utiliza redes neurais convolucionais para capturar as informações relevantes nos comentários [10]. O MPCN, por sua vez, emprega uma arquitetura de co-atenção multi-pontual para capturar o contexto em diferentes níveis de granularidade [8]. O D-ATTN adota uma rede neural de atenção dupla considerando os comentários e as características do usuário/item [7]. O NARRE utiliza uma abordagem baseada em redes neurais para modelar a atenção e as interações entre aspectos nos comentários [1]. O DAML propõe uma abordagem de aprendizado mútuo de atenção entre avaliações e comentários [4]. O CARL utiliza redes neurais convolucionais em cápsulas para gerar recomendações e fornecer explicações sobre as preferências do usuário [9]. O CARP introduz uma estrutura de rede neural para incorporar a atenção contextual na modelagem de avaliações e comentários [3]. O ANR adota uma abordagem baseada em aspectos para recomendação, capturando a relação entre aspectos e usuários/itens [2]. O HRDR realiza uma abordagem conjunta de representações de aprendizado profundo de avaliações e comentários [5]. Por fim, o RGNN propõe uma representação hierárquica de comentários de avaliações em forma de grafo para aprimorar a precisão das recomendações [6].

4.2 Configuração dos Algoritmos

Utilizamos os códigos dos algoritmos provenientes no GitHub dos respectivos autores e realizamos uma tunagem de parâmetros, de acordo com o apresentado na Tabela 2.

4.3 Coleções de Dados

A Tabela 3 apresenta alguns detalhes sobre as coleções disponibilizadas pelo iRev. Para garantir reprodutibilidade, todas elas são divididas em subconjuntos de treino, teste e validação. O conjunto de treino é composto por 80% dos dados, enquanto os conjuntos de validação e teste possuem 10% cada. Os *reviews* presentes nos dados foram pré-processados utilizando a biblioteca NLTK, que permitiu realizar tratamentos nos textos, tais como remoção de *stopwords* e lematização.

Parâmetros	Valores
Épocas de treinamento	10, 20, 50
Função de perda	MSE
Otimizador	ADAM
Dimensões dos vetores de usuário e item	32
Dimensões dos vetores de palavras	300
Codificadores utilizados	TF-IDF, Word2Vec e FastText.
Taxa de dropout	0.5
Weight decay	$1e^{-3}$
Tamanho do lote	128 a 32
Tamanho máximo dos documentos	500 palavras
Taxa de aprendizado	$2e^{-3}$
# filtros nas camadas convolucionais	100

Table 2: Configurações dos parâmetros dos algoritmos

Coleção	# Usuários	# Itens	Esparsidade
Amazon - Video Games	10.000	17.005	99.99%
Yelp - Tampa	18.437	8.664	99,99%
Yelp - Philadelphia	32.376	14.226	99.99%

Table 3: Visão geral das coleções utilizadas na avaliação.

4.4 Métricas

Para avaliar as recomendações dos algoritmos consideramos quatro métricas de precisão: duas métricas de erro (i.e., MSE e MAE) que avaliam a diferença entre o *rating* real e o previsto pelos algoritmos; e duas de efetividade (i.e. Accuracy e F1 Score) que avaliam o quão bem o algoritmo aprendeu o comportamento do usuário. O consenso na comunidade de SsR é que a precisão por si só não é suficiente para avaliar a eficácia prática e o valor agregado das recomendações. Assim, além da precisão, exclusivamente considerada em praticamente todos revisados neste artigo, consideramos outras duas métricas: serendipidade e diversidade. A serendipidade se refere a descoberta de itens úteis e inesperados e a diversidade aos itens recomendados diferentes do histórico de consumo.

4.5 Avaliação Experimental

Com o objetivo de validar o framework proposto, realizamos uma avaliação experimental de todos os algoritmos implementados, considerando as cinco métricas nas três coleções disponibilizadas e os resultados são apresentados na Tabela 4 apresentamos os resultados obtidos. Observamos que não há um destaque único. Na coleção Amazon, por exemplo, dos 10 algoritmos analisados, cinco deles se destacaram em distintas métricas. Enquanto os algoritmos HRDR, D-ATTN e o NARRE se destacaram em métricas de precisão, o DeepCONN e o ANR se destacaram em diversidade e serendipidade.

Os resultados também variam de acordo com as coleções. Na Yelp - Tampa, o segundo algoritmo mais recente proposto, HRDR, obteve melhores resultados nas métricas de precisão, corroborando com os experimentos mencionados no artigo original. Além disso, observou-se que os algoritmos CARP, CARL e MPCN não obtiveram resultados significativos nessas mesmas métricas, o que contradiz as afirmações de seus respectivos artigos. Por outro lado, esses algoritmos foram destaque em termos de serendipidade e diversidade. O ANR não obteve o melhor resultado em nenhuma métrica, mas

Coleção	Amazon - Video Games						Yelp - Tampa						Yelp - Philadelphia					
Measure	MSE	MAE	Acc	F1@10	Ser	Div	MSE	MAE	Acc	F1@10	Ser	Div	MSE	MAE	Acc	F1@10	Ser	Div
DeepCoNN	1.541	0.928	0.206	0.323	0.141	**0.197**▲	1.337	0.892	**0.361**▲	0.216	0.147	0.097	1.561	0.994	0.300	0.125	**0.159**●	**0.224**●
D ATTN	1.127	**0.727**▲	0.228	0.428	0.048	0.060	1.346	0.912	0.329	0.197	0.159	0.088	1.207	0.871	0.343	0.198	0.141	0.099
MPCN	1.636	0.993	0.121	0.262	0.069	0.1598	1.447	0.965	0.288	0.124	**0.164**●	0.163	1.322	0.913	0.323	0.121	0.145	0.125
NARRE	1.075	0.691	0.255	**0.459**▲	0.061	0.141	1.302	0.892	0.348	0.218	0.147	0.049	1.172	0.844	0.364	0.218	0.146	0.063
DAML	1.149	0.744	0.234	0.411	0.035	0.094	1.364	0.935	0.308	0.177	0.146	0.037	1.270	0.899	0.329	0.173	0.152	0.037
CARL	1.286	0.839	0.326	0.152	0.021	0.081	1.525	0.995	0.293	0.134	**0.166**●	0.125	1.306	0.921	0.332	0.171	**0.153**●	**0.224**●
CARP	1.262	0.824	0.340	0.170	0.213	0.124	1.583	0.987	0.285	0.107	0.147	**0.225**▲	1.336	0.902	0.324	0.136	0.152	0,172
ANR	1.171	0.780	0.381	0.226	**0.242**▲	0.071	1.288	0.899	0.338	0.215	0.148	0.061	**1.115**▲	**0.813**▲	**0.397**▲	**0.267**▲	0.124	0.050
HRDR	**1.039**▲	0.751	**0.456**▲	0.248	0.082	0.081	**1.257**▲	**0.879**▲	0.355	**0.249**▲	0.146	0.050	1.482	0.964	0.324	0.209	0.146	0.057
RGNN	1.179	0.792	0.297	0.150	0.101	0.148	1.364	0.916	0.314	0.186	0.142	0.124	1.296	0.896	0.287	0.128	0.150	0.075

Table 4: Resultados validados com o teste de Wilcoxon com um valor p = 0,05. ▲ representa ganhos significativos e ● empates estatísticos.

apresentou resultados consistentes e estáveis em termos de desempenho. O algoritmo RGNN, embora seja o mais recente em termos de proposta, apresentou resultados inferiores a muitas outras estratégias avaliadas. Na coleção Yelp - Philadelphia, o algoritmo ANR mostrou os melhores resultados nas quatro métricas relacionadas à efetividade, mais uma vez alinhados com o que foi apresentado no artigo original. O algoritmo CARL foi melhor nessa coleção em comparação com a anterior, obtendo os melhores resultados de serendipidade e diversidade, juntamente com o DeepCoNN.

Grande parte dos algoritmos apresentaram resultados consistentes com seus estudos originais. No entanto, alguns algoritmos não obtiveram bons resultados em comparação com o que foi descrito pelos autores. Algoritmos como CARL, CARP, MPCN e RGNN, que teoricamente deveriam superar metodologias como DATTN, ANR e NARRE, não tiveram sucesso em nossa avaliação empírica. Esses resultados reforçam a importância do iRev por avançar na questão da reprodutibilidade, por ser um repositório público não apenas de coleções de dados, como também dos próprios algoritmos e seus processos de tunagem de parâmetros.

5 CONCLUSÕES E TRABALHOS FUTUROS

Como primeira contribuição, esse trabalho apresenta um mapeamento sistemático dos estudos sobre sistemas de recomendação *review-aware* (RARSs) selecionando e investigando os 117 artigos relevantes publicados nos principais veículos da área (e.g., RecSys, SIGIR, WWW, etc.), identificando esforços, resultados, contribuições e limitações relevantes. A partir desse levantamento, propomos e disponibilizamos um framework, denominado iRev, contendo a implementação dos 10 principais RARSs, bem como todos os artefatos levantados durante o mapeamento sistemático (métricas e bases de dados) com o intuito de mitigar a limitação atual de falta de reprodutibilidade devido à ausência de códigos fontes e de confiabilidade devido à ausência de distintas métricas de avaliação. Para validar o iRev, realizamos uma avaliação completa das principais abordagens, considerando diferentes coleções de dados e métricas. Nossos resultados mostram que as SsR baseadas em redes neurais, especialmente as que utilizam mecanismos de extração de atenção e aspecto, obtiveram os resultados mais competitivos. Por outro lado, tais resultados também reforçam que não há um único algoritmo que se destaque de forma absoluta, deixando claro que ainda há espaço de melhora considerável a ser explorado por novas estratégias. Como trabalhos futuros, visamos complementar o iRev com a implementação de outras estratégias, tornando o repositório uma referência para que pesquisadores da área.

AGRADECIMENTOS

Este trabalho foi financiado por CNPq, CAPES, Fapemig, FAPESP e AWS.

REFERENCES

[1] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 world wide web conference*. 1583–1592.
[2] Jin Yao Chin, Kaiqi Zhao, Shafiq Joty, and Gao Cong. 2018. ANR: Aspect-based neural recommender. In *27th ACM CIKM*. 147–156.
[3] Chenliang Li, Cong Quan, Li Peng, Yunwei Qi, Yuming Deng, and Libing Wu. 2019. A capsule network for recommendation and explaining what you like and dislike. In *42nd ACM SIGIR*. 275–284.
[4] Donghua Liu, Jing Li, Bo Du, Jun Chang, and Rong Gao. 2019. Daml: Dual attention mutual learning between ratings and reviews for item recommendation. In *25th ACM SIGKDD*. 344–352.
[5] H. Liu, Y. Wang, Q. Peng, F. Wu, L. Gan, L. Pan, and P. Jiao. 2020. Hybrid neural recommendation with joint deep representation learning of ratings and reviews. *Neurocomputing* 374 (2020), 77–85.
[6] Yong Liu, Susen Yang, Yinan Zhang, Chunyan Miao, Zaiqing Nie, and Juyong Zhang. 2021. Learning hierarchical review graph representations for recommendation. *IEEE TKDE* 35, 1 (2021), 658–671.
[7] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *11st ACM RecSys*. 297–305.
[8] Y. Tay, A. Tuan Luu, and S. Hui. 2018. Multi-pointer co-attention networks for recommendation. In *24th ACM SIGKDD*. 2309–2318.
[9] Libing Wu, Cong Quan, Chenliang Li, Qian Wang, Bolong Zheng, and Xiangyang Luo. 2019. A context-aware user-item representation learning for item recommendation. *ACM TOIS* 37, 2 (2019), 1–29.
[10] L. Zheng, V. Noroozi, and P. Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *ACM WSDM*. 425–434.