Análise de Recomendação de livros entre SVD e BPR

ISRAEL FELIPE DA SILVA, ICMC - USP, Brasil

Clique aqui para ir ao vídeo com a apresentação.
Um sistema de recomendação é um sistema que combina várias técnicas computacionais para recomendar itens personalizados com base nos interesses e históricos de um usuário. Para tal, existem diversos algoritmos que podem gerar recomendações. A proposta deste trabalho é gerar um recomendador de livros usando o algoritmo SVD otimizado e usando o algoritmo BPR com fatoração de matriz. Por fim, analisar e discutir os resultados obtidos.

ACM Reference Format:
Israel Felipe da Silva. 2022. Análise de Recomendação de livros entre SVD e BPR. 1, 1 (December 2022), 3 pages.

1 INTRODUÇÃO

A proposta deste trabalho é gerar um sistema de recomendação de livros, para isto, a base de dados a ser utilizada será uma das disponibilizadas pela Universidade da Califórnia em San Diego em seu UCSD Book Graph. A UCSD Book Graph é um conjunto de datasets coletados em 2017 do website goodreads.com, e contém um grupo de datasets que possui metadados dos livros, um grupo de datasets que contém interações de usuários com livros e um grupo de datasets que contém informações mais detalhadas das reviews realizadas pelos usuários, bem como o contexto desta review. Além disso, a base de dados contém arquivos que separam livros por gênero, abrindo um leque maior de possibilidades para o sistema de recomendação. É também disponibilizado um link para o github com exemplos de códigos de acesso aos arquivos e notebooks disponibilizados, tudo muito bem documentado e de fácil acesso.

O arquivo de interações utilizado contém as reviews de usuários para determinados livros, além disso, possui um atributo binário que representa se o livro foi revisado pelo usuário, sendo uma alternativa de feedback implícito para ser usada no treinamento. Além disso, existem alguns atributos já oferecidos pelo dataset como por exemplo pontuação média e livros similares. Tais informações nos oferecem uma maneira alternativa para fornecer outras recomendações, que pode resultar em melhores recomendações se usados no treinamento, em contraste às recomendações calculadas utilizando outras métricas, por exemplo, livros similares com similaridade calculada por cosseno.

A proposta principal do projeto é utilizando o arquivo de interações do tipo usuário-livro, usar o algoritmo SVD otimizado para gerar recomendações, em seguida, utilizando uma implementação do algoritmo learning to rank BPR com SVD (fatoração de matriz) gerar também recomendações de livros para os usuários.

Foi desenvolvido em um notebook (.ipynb) todo o código dos recomendadores, desde o download dos datasets mencionados, os algoritmos de treinamento e o código para gerar recomendações

Author's address: Israel Felipe da Silva, israelfelipe@usp.br, ICMC - USP, Brasil.

XXXX-XXXX/2022/12-ART $15.00
https://doi.org/

de livros para um determinado usuário. Também estará no notebook uma analise e discussão dos resultados.

2 SVD

Singular Value Decomposition (SVD) é uma técnica de álgebra linear de fatoração de matrizes, em outras palavras, a fatoração por SVD de uma matriz M gera uma fatoração do tipo $M = U \sum V$. No contexto de sistemas de recomendação, considerando uma matriz M de interação do tipo usuário-livro, a fatoração por SVD gera $M = U \sum V$, onde, U é a matriz de representação por fatores latentes dos usuários, $\sum$ é a matriz diagonal e V é a matriz de representação por fatores latentes dos itens. Estes fatores latentes guardam em suas matrizes características dos usuários e itens, estas características não são facilmente interpretáveis mas mapeiam características não tão triviais.

Apesar de apresentar resultados interessantes, o SVD comum possui alguns problemas que precisam ser contornados em um sistema que precisa de escalabilidade e que trabalha com muitos dados. Devido ao fato de que o SVD comum trabalha com matriz, em uma grande base de dados, com muitos usuários e itens, a memória acaba sendo um grande problema, e além disso, considerando que, em qualquer sistema, os usuários avaliam geralmente uma porcentagem muito baixa dos itens totais, resultando em uma matriz muito esparsa, comprometendo o treinamento do modelo na geração das matrizes.

Solucionando dois dos grandes problemas do SVD tradicional (memória e esparsidade), surge a versão otimizada do SVD, conhecida como FunkSVD, que gera a fatoração de matrizes treinando o modelo através da técnica de gradiente descendente. O gradiente descendente é uma técnica que basicamente considera a derivada da aproximação do erro das predições e treina o modelo baseado nessa informação, visando sempre ajustar os hiperparâmetros que forneçam o menor erro. Além disso, o algoritmo considera apenas as interações reais que aconteceram entre usuário e item, evitando assim, o problema de uso excessivo de memória e esparsidade de matriz.

Por fim, com as matrizes de representação por fatores latentes dos usuários e dos itens, é possível obter uma estimativa da nota que um determinado usuário daria para um determinado item simplesmente pelo produto escalar da linha da matriz U pela coluna da matriz transposta V. E através destas estimativas podem ser geradas recomendações para um determinado usuário da base.

Vale ressaltar que a precisão destas estimativas dependem de vários fatores, incluindo: Valor dos hiperparâmetros, tempo de treinamento, conjunto de treinamento, etc...

3 BPR

Bayesian Personalized Ranking (BPR) é um algoritmo Learning to Rank (LTR) do tipo Pairwise. Algoritmos do tipo LTR, são algoritmos que produzem listas ranqueadas de itens para usuários da base, ou seja, dado um usuário, o sistema retorna itens considerados relevantes, do mais relevante para o menos relevante. O ranqueamento é feito usando um modelo de ranqueamento, que é treinado de maneira supervisionada usando um algoritmo LTR, ou seja, para cada usuário da base é necessária uma lista de itens ranqueados por relevância durante o treinamento.

Pairwise significa que o algoritmo usa durante o treinamento uma abordagem do tipo classificador binário: dado dois itens, retorna uma ordenação entre eles.

O BPR pode ser usado com diverso modelos de recomendação, no projeto, usaremos o SVD otimizado (FunkSVD) com fatoração de matrizes. Gerando no fim, um ranqueamento ótimo de itens para cada usuário da base. O BPR utiliza uma tripla (u, i, j) que representa que o usuário u visitou ou comprou o item i mas não o j, e através desta informação o algoritmo sabe que um ranqueamento ótimo para o usuário u é onde o item i vem antes de j no ranking. Em suma, o algoritmo quer maximizar a probabilidade de o sistema retornar uma ordenação que satisfaça todas as preferências de todos os usuários. Abstraindo a parte de cálculo, o BPR também trabalha com a ideia de gradiente descendente e derivadas parciais dos parâmetros.

4 RESULTADOS OBTIDOS

A base de dados contém um número considerável de usuários e livros, com isso, as recomendações acabam sendo de itens não tão comuns (recomendados ou já avaliados) da base, fazendo com que não seja simples de determinar se a recomendação foi eficiente a primeira vista ou não. Um pré processamento dos dados, afim de unir mais informações de usuários com menos linha de interações, visto que o arquivo original de interações tem aproximadamente 4GB, e talvez limitar o número de itens usado nas interações poderiam possivelmente gerar predições melhores.

Contudo, após treinar ambos algoritmos com o arquivo de interações reduzido, gerar as predições para usuários da base e amostrar para os primeiros 50 usuários da base, notamos que ambos algoritmos recomendaram eventualmente itens que de fato o usuário leu e gostou do conjunto de teste, ou seja, itens que não foram usados no treinamento do modelo. Além disso, vale ressaltar que o algoritmo BPR-MF obteve um resultado consideravelmente melhor que o SVD otimizado, isto é natural visto que o SVD otimizado dá o rating e tenta aproximar a nota durante o treinamento, usando fatoração de matriz para isso, enquanto que o BPR-MF usa a mesma fatoração de matriz, mas tenta garantir a ordem (ranqueamento) dos itens, sem se preocupar com a nota, resultando naturalmente numa recomendação mais eficiente. Com isso, o algoritmo Pairwise(BPR) se mostrou mais eficiente que o algoritmo Pointwise(SVD).

5 CONCLUSÃO

De uma maneira geral, não só o projeto final, mas sim a matéria como um todo, foi uma experiência enriquecedora e de muito aprendizado. A área de sistemas de recomendação, tem se mostrado um campo muito fértil para a evolução, e com isso, a matéria foi de fundamental importância para introduzir os alunos no tema. Sobre o projeto, talvez uma das maiores dificuldades é trabalhar com os dados, visto que, dados fornecidos de sistemas reais são geralmente muito volumosos, necessitando de muito tempo de treinamento e poder computacional ou então de um pré-processamento, a fim de compactar as informações sem perder muito conteúdo e cuidando sempre para evitar viéses. Por fim, fica registrado o agradecimento em especial ao Profº Marcelo Manzato por ter dedicado tempo e esforço a propagar conhecimento, conhecimento esse, que sem dúvidas num futuro não distante mostrara seus frutos.