Ruído Natural na Filtragem Colaborativa

[Trabalho Final]

Paulo Xavier
contato.pauloxavier@gmail.com

Gabriel Segobia
segobia.gos@gmail.com

Fellipe Bravo
fellipe.bravo@gmail.com

1. ABSTRACT

Em SR é assumido que as avaliações nos *datasets* estão livres de irregularidades. Foi mostrado recentemente que os usuários podem ser inconsistentes quando eles elicitam avaliações para itens, expondo os dados usados nos SR à inconsistências. Esse trabalho busca quantificar o impacto do ruído na predição por modelos (usando rsvd), e o desempenho dos algoritmos de detecção de ruído propostos por *O'Mahony* e *Toledo*.

2. INTRODUÇÃO

Sistemas de Recomendações(SR) são ferramentas de software e técnicas que fornecem sugestões de itens que sejam de uso ao usuário. Existem diversas abordagens que podem ser usadas em SR, e uma das mais populares é a *Collaborative Filtering(CF)*, que produz recomendações específicas a um usuário baseado em padrões de avaliação ou uso de itens.

Em SR é assumido que as avaliações nos *datasets* estão livres de irregularidades. Foi mostrado recentemente que os usuários podem ser inconsistentes quando eles elicitam avaliações para itens, expondo os dados usados nos SR à inconsistências. Esse tipo de inconsistência é conhecida como *ruído natural*. Abordagens que buscam *detectar* e *corrigir* essas avaliações inconsistentes surgiram para resolver esse problema em aberto.

O artigo será organizado da seguinte forma: Seção 3 faz uma revisão de filtragem colaborativa. Seção 4 elicita o problema e faz uma proposta para quantificar o impacto do ruído nos modelos e qual o desempenho dos algoritmos *Mahony* e *Toledo* na detecção de ruído natural. Seção 5 faz uma discussão sobre os experimentos feitos para avaliar o impacto do ruído nos modelos. Seção 6 explica o *dataset* utilizado para o artigo. Seção 7 apresenta a metodologia usada para obtenção dos resultados obtidos na Seção 8. A seção 9 apresenta uma conclusão sobre os resultados obtidos e propostas para trabalhos futuros sobre o tema.

3. FILTRAGEM COLABORATIVA

Collaborative Filtering é um algoritmo de recomendação popular que baseia suas predições e recomendações nas avaliações ou comportamento de outros usuários no sistema. A principal suposição por trás desse método é que as opiniões de outros uários podem ser selecionadas eagregadas de maneira que é possivel ter uma predição razoável das preferências um usuário alvo.

Intuitivamente, assume-se também, que se os usuários concordam sobre a qualidade ou relevância de alguns itens, então eles também vão provavelmente concordar sobre outros itens.

A grande maioria dos algoritmos de CF usados atualmente, operam inicialmente gerando predições de preferência usuário e então produzindo as recomendações deles ranqueando os itens candidatos por preferências previstas.

3.1 Vantagens e Desvantagens

Os CFs possuem algumas vantagens notáveis:

- Não necessitam de informações sobre conteúdo ou usuários. As *abordagens puras* utilizam apenas as avaliações para fazer a predição;
- Consegue avaliar a experiencia, qualidade e ponto de vista de outras pessoas na predição;
- Consegue sugerir *serendipitous items* através da análise de pessoas do comportamento de usuários semelhantes

Possui também desvantagens, dentre elas:

- Baixa acurácia quando possui poucas informações sobre as avaliações dos usuários(*Cold Start*);
- Ratings são dados explicitamente por usuários, gerando uma base de dados ruidosa por natureza (*ruído natural e/ou ruído malicioso*).

4. PROBLEMA/PROPOSTA

Em SR é assumido que as avaliações nos *datasets* estão livres de irregularidades. Foi mostrado recentemente que os usuários podem ser inconsistentes quando eles elicitam avaliações para itens, expondo os dados usados nos SR à inconsistências.

Ruído em SR pode ser classificado em 2 categorias principais:

1. *Ruído Malicioso*, associado ao ruído intencionalmente introduzido por um agente externo para influenciar os resultados de um recomendador, e

2. *Ruído Natural*, involutariamente introduzido por usuários, e que também poderia afetar o resultado da recomendação.

O *ruído natural* é inserido sem intenção maliciosa, e, ao contrário do *ruído malicioso*, que já é estudado na literatura há tempos, é um tópico recente.

A identificação do *ruído natural* é mais difícil, pois ele tende a aparecer de diversas formas (ao contrário do malicioso, que costuma estar associado à alguns padrões nos perfis dos usuários).

Neste artigo serão usados dois algoritmos para quantificação do impacto do ruído no modelo

-Proposta para o problema

5. EXPERIMENTOS

6. DATASET

Para este artigo, foi utilizado o dataset *movielens*, composto por 943 usuários e 1982 itens(filmes), num total de 100 mil avaliações (*movielens 100k*). Essas avaliações são compostas de notas de 1 a 5.

7. METODOLOGIA

10% treinamento -> gerar ruidos aqui

90% teste

Usando Precision recall f1 (comparando a previsão sem ruído/com ruído e checar a % de ruidos)

Algoritmos: Mahony e Toledo

8. RESULTADOS

9. CONCLUSÃO E TRABALHOS FUTUROS

10. REFERENCES