XprefRec: minimizando o problema de cold-start de item com mineração de preferências

Cleiane Gonçalves Oliveira[1,2], Sandra de Amo[1]

[1] Universidade Federal de Uberlândia, Brasil
[2] Instituto Federal do Norte de Minas Gerais - Campus Januária, Brasil
cleiane.oliveira@ifnmg.edu.br, deamo@ufu.br

Abstract. Sistemas de recomendação de filtragem colaborativa encontram diversos desafios para alcançar acuradas recomendações. Um deles é o denominado *cold-start* de item que é o fato do sistema não ser capaz de recomendar itens que nunca foram avaliados por outros usuários. Neste artigo apresentamos uma proposta para minimizar esse problema, o ***XPrefRec***. Trata-se de um sistema de recomendação híbrido, seguindo os princípios gerais das abordagens clássicas de *filtragem colaborativa* e *recomendação baseada em conteúdo*, e incorporando técnicas de *mineração de preferências contextuais* e técnicas de *extração de consenso*. Testes realizados em dados reais de filmes mostram resultados bastante promissores de precisão, revocação e cobertura quando comparados aos resultados obtidos utilizando-se um método clássico de recomendação (*Content-boosted collaborative filtering (CBCF)*), também baseado em uma abordagem híbrida.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Data Mining; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.4 [**Information Systems Applications**]: Miscellaneous

Keywords: Mineração de dados, Mineração de preferências, Regra de preferência contextual, Sistemas de recomendação

1. INTRODUÇÃO

Cada vez mais sistemas de recomendação vêm sendo utilizados em diversos cenários devido a sua capacidade de predizer sobre o que os clientes/usuários têm preferência. A técnica mais popular é a denominada *filtragem colaborativa* (CF) introduzida em [Resnick et al. 1994]. Esta técnica recomenda itens a um determinado usuário a partir do histórico de compras/avaliações de outros usuários que tenham gostos semelhantes a ele, isto é, que tenham um histórico semelhante de avaliações de produtos. Os sistemas de recomendação CF partem da ideia de que se um usuário gostou das mesmas coisas que outro no passado, ele tende a continuar gostando das mesmas coisas no futuro. Para prover recomendações é necessário também que o usuário ao qual se deseja recomendar (usuário ativo) possua algum histórico de avaliações para receber recomendações.

Porém, a técnica CF enfrenta alguns desafios. Os usuários do sistema, geralmente, avaliam poucos itens, fazendo com que os dados disponíveis sobre avaliações sejam esparsos. Para que um item seja recomendado é necessário que ele tenha sido avaliado pelos usuários similares ao usuário ativo. Desta forma, um item que nunca recebeu nenhuma avaliação não é recomendado. Da mesma forma, um item novo, recém inserido no sistema e que ainda não possui avaliação, também não é recomendado até que um número expressivo de usuários o avalie. Esta dificuldade é denominada como *cold-start* de item [Adomavicius and Tuzhilin 2005].

Entre os diversos trabalhos que tentam tratar esse problema tem-se o *Content-boosted collaborative filtering* (CBCF), introduzido em [Melville et al. 2002]. Este trabalho apresenta um sistema de recomendação híbrido de CF com *recomendação baseada em conteúdo* (CB). Um sistema de recomendação CB recomenda itens semelhantes ao que o usuário já avaliou. Por exemplo, se o usuário avaliou bem um filme de comédia, o sistema CB recomendará outros filmes de comédia. O CBCF aplica a técnica CB para cada usuário do sistema:

We thank the Brazilian Research Agencies CNPq and FAPEMIG for supporting this work.

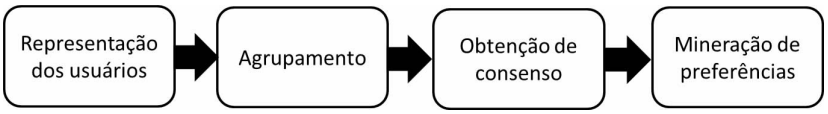

Fig. 1. Fase de processamento *offline* do ***XprefRec***

a partir dos itens avaliados pelo usuário, são preditas avaliações para os itens não avaliados. Desta forma, a técnica CB elimina o problema da esparsidade dos dados fazendo com que todos os usuários tenham avaliado todos os itens. A técnica CF é então aplicada: encontra-se quais são os usuários mais similares ao usuário ativo e a partir de seus históricos são feitas as recomendações. Observa-se que o CBCF consegue tratar o problema de *cold-start* de item, uma vez que todos os itens conseguem ser recomendados porque possuem avaliações de todos os usuários. Porém, quando um novo item é inserido no sistema repete-se todo o processo da técnica CB para predição da avaliação dos usuários àquele novo item, para que assim, em um momento próximo, ele possa ser recomendado.

Neste artigo apresentamos o ***XprefRec***, um sistema de recomendação híbrido CF e CB que utiliza técnicas de *mineração de preferências contextuais* e técnicas de *extração de consenso* entre usuários, a fim de otimizar a questão da recomendação de itens novos para os usuários. O sistema é capaz de tratar o problema de *cold-start* de item de forma mais eficiente que o CBCF (não necessitando refazer o modelo de recomendação a cada inserção de um novo item) sem prejudicar a acurácia da recomendação.

Este artigo é organizado como se segue: na Seção 2 apresentamos o sistema ***XprefRec***, na Seção 3 os resultados dos experimentos em dados reais e na Seção 4 a conclusão e trabalhos futuros.

2. O SISTEMA *XPREFREC*

O ***XprefRec*** apresenta uma abordagem diferenciada de um sistema de recomendação híbrido CF e CB a partir da *mineração de preferências contextuais* e *extração de consenso* entre os usuários. São definidos grupos entre os usuários do sistema que possuam gostos semelhantes. Para cada grupo define-se um gosto consensual que por sua vez é submetido a um minerador de preferências contextuais. Esse minerador apresenta como resultado um conjunto de *regras de preferências contextuais* que será utilizado para predizer preferências de usuários que se enquadram naquele grupo. Dados dois itens, uma *regra de preferência contextual* é capaz de dizer qual item é preferido, ou se não é possível compará-los, a partir das características dos itens. O ***XprefRec***, no momento da recomendação ao usuário ativo, em vez de compará-lo com todos os usuários do sistema, faz a comparação somente com os grupos, sendo que estes são em número bem menor em relação à quantidade de usuários. As regras daquele grupo são então aplicadas sobre os itens para predizer recomendações. As características da técnica CB estão implícitas no minerador de preferências, já que o modelo de preferências extraído envolve o conteúdo dos itens. Assim, para que um item seja recomendado basta que ele atenda às regras associadas ao usuário ativo. Quando um novo item é inserido no sistema este já pode ser recomendado a um usuário ativo, simplesmente aplicando o modelo de preferências relativo ao grupo no qual este usuário se enquadra.

O ***XprefRec*** pode ser dividido na fase de processamento *offline* e na fase *online*, que consiste na atividade de recomendação. Uma visão geral da fase *offline* é apresentada na Figura 1.

2.1 Representação dos usuários

Diferente da representação dos usuários normalmente utilizada em sistemas de recomendação que representam cada usuário por um vetor de notas (como na Tabela I), apresentamos o conceito de *matriz de preferências* (*MPref*) como uma nova maneira de representar os usuários em sistemas de recomendação. Uma *MPref* é uma matriz de dimensão $n \times n$, sendo n a quantidade de itens no sistema. A cada usuário está associada uma *MPref*. Cada posição (a, b) da *MPref* de um usuário u contém um valor entre 0 e 1 que representa o quanto o usuário u prefere o item i_a ao item i_b. Este valor é calculado utilizando-se uma *relação de preferência fuzzy* ([Chiclana et al. 2001]) dada pela fórmula: $r(n_a, n_b) = \frac{n_a}{n_b}$, onde n_a (resp. n_b) é a nota que o usuário associou ao item i_a (resp. n_b). Esta razão é normalizada utilizando-se a função $h(x) = \frac{x}{x+1}$ que além de produzir um número no intervalo $[0, 1]$ também verifica algumas propriedades importantes quando se trata de *graus de preferência*,

Table I. Avaliações do usuário ***u***

i_1	5
i_2	3
i_3	*
i_4	1

Table II. *MPref* relativa ao usuário ***u***

	i_1	i_2	i_3	i_4
i_1	0.50	0.63	*	0.83
i_2	0.37	0.50	*	0.75
i_3	*	*	0.50	*
i_4	0.17	0.25	*	0.50

dentre elas $h(\frac{1}{x}) = 1 - h(x)$[1]. Assim, o valor final normalizado é dado por $f(a, b) = h(r(n_a, n_b))$. Como exemplo, a Tabela I apresenta um conjunto de avaliações dadas por um usuário ***u***. O símbolo * representa o fato que o usuário ***u*** não avaliou o item correspondente, o que consequentemente não permitirá a comparação desse item com nenhum outro. Calculando a relação de preferência fuzzy entre cada par de item, chega-se à *MPref* apresentada na Tabela II. O valor $f(1, 4) = 0,83$, por exemplo, representa o fato de que o usuário ***u*** prefere o item i_1 ao item i_4 com um grau de preferência $0,83$.

As matrizes de preferências são usadas para determinar a semelhança entre os usuários. Um usuário é semelhante a outro se suas matrizes de preferência forem similares (de acordo com alguma noção de similaridade a ser discutida na Seção 2.2). Assim, usuários semelhantes gostam dos mesmos itens com graus de preferência semelhantes.

2.2 Agrupamento de usuários

Os usuários, representados por suas respectivas matrizes de preferência, são organizados em grupos. Para esta tarefa aplica-se o algoritmo de clusterização DBScan introduzido em [Ester et al. 1996] e a distância do cosseno como medida de similaridade entre as matrizes. Para cada grupo é calculada uma *MPref* consensual. Esta tarefa é discutida na Seção 2.3.

De uma certa maneira, pode-se dizer que a *MPref* consensual de um determinado grupo agrega tanto as preferências comuns de seus usuários quanto preferências específicas de cada um, o que resulta na fase de mineração, em uma maior variedade de regras possibilitando melhores recomendações. Além disso, no momento da recomendação as amostras de preferência do usuário ativo (uma matriz bem esparsa) precisará ser comparada com as matrizes consensuais de cada grupo, cujo número será menor que a quantidade de usuários do sistema, como é feito na técnica CF.

2.3 Técnicas alternativas para a *extração do consenso*

A obtenção do consenso constitui uma etapa muito importante para alcançar boas recomendações. Para tanto, propôs-se o estudo de cinco alternativas para obtenção do consenso que são aplicadas ao conjunto de valores de cada posição da matriz consensual. Cada posição (i, j) da *Mpref* consensual só é calculada se mais da metade dos usuários do grupo informaram algum valor para esta posição, caso contrário tal posição permanece com o símbolo "*".

A primeira alternativa (Alternativa (1)) aplica simplesmente a média aritmética entre os valores das correspondentes posições (i, j) de todas as *Mpref* do grupo. A segunda alternativa (Alternativa (2)) aplica a média ponderada sendo o peso de cada usuário a medida do seu coeficiente de silhueta. Este coeficiente mede o quanto o usuário está coeso ao seu grupo e separado dos outros grupos. Desta forma o usuário que possuir maior coeficiente de silhueta possuirá peso maior no cálculo do consenso.

As alternativas (3), (4) e (5) consistem na aplicação dos quantificadores *fuzzy most*, *at least half*, *as many as possible* respectivamente. Quantificadores *fuzzy* expressam termos da linguagem natural e definem quais valores, dentro de um conjunto de valores, serão considerados no cálculo do consenso ([Chiclana et al. 2001]). Vamos ilustrar estas 3 alterativas através de um exemplo. Suponha que temos 4 usuários dentro de um grupo e desejamos calcular o valor de uma certa posição da matriz de consenso deste grupo. Os valores desta posição nas 4 matrizes de preferência constitui um vetor $V = (0.4, 0.6, 0.5, 0.3)$. A aplicação de um quantificador

[1]Outras funções de normalização podem ser utilizadas, desde que verifiquem certas propriedades especificadas em [Chiclana et al. 2001], dentre elas a propriedade $h(\frac{1}{x}) = 1 - h(x)$.

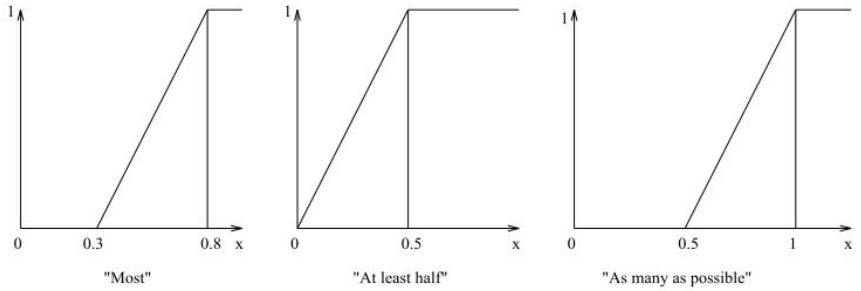

Fig. 2. Quantificadores *fuzzy* e seus parâmetros

fuzzy para obter o valor consensual envolve os seguintes passos: (1) ordenação dos valores do vetor de forma decrescente; (2) cálculo do peso de cada posição do vetor a partir dos parâmetros a e b de cada quantificador (por exemplo, na Figura 2 os parâmetros do quantificador *most* são a = 0.3 e b = 0.8); (3) cálculo da média ponderada dos valores do vetor com seus respectivos pesos. O peso de cada posição i ($i = 1, ..., 4$) é calculado por

$$w_i = Q(i/n) - Q((i-1)/n), \text{ onde } Q(r) = \begin{cases} 0 \text{ se } r < a \\ \frac{r-a}{b-a} \text{ se } a \leq r \leq b \\ 1 \text{ if } r > b \end{cases}$$

Assim, o consenso dos valores do vetor V utilizando-se os quantificadores *most*, *at least half* e *as many as possible* são 0.43, 0.55 e 0.35 respectivamente.

2.4 Mineração de preferências

Sobre a *MPref* consensual de cada grupo é aplicado o minerador de preferências contextuais *CPrefMiner* ([de Amo et al. 2012], [de Amo et al. 2013]). Este minerador explora as características dos itens sobre os quais o grupo tem preferência e retorna um conjunto de *regras de preferências contextuais* que permitem dizer, dados dois itens, qual deles o usuário prefere.

As *regras de preferências contextuais* são mineradas no formato contexto->preferência, exemplo: Diretor="Woody Allen"→ (Categoria=comédia) > (Categoria=drama). A parte à esquerda de "→"é o contexto da regra e a parte à direita é a preferência neste contexto. Esse exemplo, em outras palavras, diz que o usuário prefere filmes de comédia a filmes de drama quando o diretor é Woody Allen.

O uso das regras de preferências contextuais se torna benéfico no tratamento do *cold-start* de item. Ao associar o conjunto de regras do grupo ao usuário ativo, um novo item adicionado ao sistema de recomendação poderá ser comparado, mesmo nunca tendo sido avaliado por outro usuário, se ele puder ser ordenado pelo modelo de preferências (as regras) do grupo.

2.5 Recomendação

A fase *online* do ***XprefRec*** é a tarefa de recomendação. Dado um usuário ativo e seu histórico de avaliação, constrói-se sua matriz de preferência e a compara com as matrizes consensuais de cada grupo do ***XprefRec***. Definido o grupo mais semelhante, o seu respectivo conjunto de regras de preferências contextuais é aplicado sobre os itens para gerar as recomendações. Como as regras são aplicadas a pares de itens, são formados pares dos itens ainda não avaliados pelo usuário ativo para saber quais são recomendados. Se for desejado um *ranking* de recomendações em ordem decrescente de relevância, pode-se aplicar algoritmos que extraem um *ranking* a partir de um conjunto de pares de itens fornecido como entrada (por exemplo, o algoritmo proposto em [Cohen et al. 1999]).

3. RESULTADOS EXPERIMENTAIS

A fim de avaliar o ***XprefRec*** sobre dados reais usamos a base disponível no projeto GroupLens [2] que apresenta dados de avaliações de filmes no formato (IdUsuario,AtributosFilme, Nota). Para compor a base de teste foram escolhidos usuários que avaliaram muitos itens e itens que foram avaliados por muitos destes usuários selecionados. Assim, selecionou-se uma base de 296 usuários, 262 itens e todas as avaliações entre estes usuários e itens, no total de 67.971 avaliações. Este banco é o denominado BD_100. Os itens são representados pelos atributos dos filmes (AtributosFilme): gênero, ator principal, categoria, diretor, ano e idioma. Para comparar o comportamento do sistema em bancos mais esparsos, o conjunto de avaliações foi sendo decrementado de forma estratificada, construindo-se mais 5 bancos de testes. Assim, o banco BD_90 possui 10% a menos de avaliações em relação ao banco original, e assim sucessivamente. A Tabela III apresenta as características de cada banco.

Para validar o sistema proposto usamos a técnica de *k-cross-validation*, para $k = 5$: primeiramente dividimos o conjunto de usuários em 5 partes, sendo que a cada iteração quatro partes são usadas para treinamento e a outra para teste. Para cada usuário de teste também foi realizado o *5-cross-validation* dividindo o conjunto de itens em 5 partes, de tal forma que os itens avaliados por esse usuário estavam divididos de forma estratificada em relação às notas recebidas.

Como as regras de preferências contextuais são aplicadas sobre pares de itens, são definidos sobre o conjunto dos 262 itens todos os pares que se sabe que o usuário avaliou. Como exemplo da Tabela I seriam gerados os pares $(i_1, i_2), (i_1, i_4), (i_2, i_4)$ de forma que o primeiro termo é preferido ao segundo. Estes pares são denominados *pares reais*.

Após o processamento *offline* do ***XprefRec***, as regras são aplicadas sobre os pares reais. Elas podem acertar indicando a ordem correta do par; errar, quando invertem a ordem; ou ser indiferente quando não é capaz de definir a ordem entre os dois itens. A partir disso são definidas as medidas de validação *precisão*(p) e *revocação*(r) bem como a média harmônica entre tais medidas:

$$\text{p} = \frac{acertos}{acertos+erros}; \text{r} = \frac{acertos}{|paresreais|}; \text{F1} = \frac{2}{\frac{1}{r}+\frac{1}{p}}.$$

Para verificar como o ***XprefRec*** se comporta em relação ao *cold-start* de item definimos a medida de *cobertura*. Esta medida calcula a capacidade do sistema para recomendar itens que não foram avaliados ainda, tendo somente suas características. Assim, para cada usuário teste foram selecionados os itens que ele avaliou e que não estavam entre os 262 selecionados. Entre estes itens foram escolhidos 20% de forma estratificada em relação as notas dadas. Com esses itens foram definidos pares e sobre eles aplicadas as regras geradas pelo sistema. As regras podem acertar (acertos_cobertura), errar (erros_cobertura) ou serem indiferentes. A medida de cobertura é definida de forma semelhante à precisão: a razão entre os acertos de cobertura e a soma entre os acertos e erros de cobertura.

O primeiro teste teve como objetivo apresentar quais das alternativas de *extração de consenso* consegue melhor resultado sobre o banco BD_100. A Tabela IV apresenta os resultados obtidos. Como pode ser observado, a alternativa (1) e (2) alcançaram melhores resultados tanto na acurácia da recomendação (precisão, revocação e f1), como na cobertura. Entre as duas alternativas, por ser levemente superior, a alternativa (1) é a escolhida para os testes de comparação com o baseline CBCF apresentados na Tabela V e Tabela VI. Nestas tabelas também são apresentados as medidas de desvio padrão σ_p, σ_r e σ_c para a precisão, revocação e cobertura respectivamente.

Em comparação com o *baseline CBCF* o ***XprefRec*** apresenta resultados superiores confirmando ser esssa uma abordagem promissora para sistemas de recomendação. Observa-se também que a precisão não decresce muito à medida que a esparsidade dos dados aumenta. Quanto à cobertura, verifica-se que o ***XprefRec*** apresentou mesmo um aumento no valor da cobertura à medida que a esparsidade aumentou, o inverso acontecendo com o *baseline*.

[2] http://www.grouplens.org/

Table III. Bancos de validação

BD	| Avaliações |	Esparsidade
BD_100	67.971	12,35%
BD_90	61.143	21,16%
BD_80	54.423	29,82%
BD_70	47.464	38,80%
BD_60	40.831	47,35%
BD_50	33.344	57,00%

Table IV. Teste das alternativas de obtenção de consenso

Alternativa	Precisão	Revocação	F1	Cobertura
(1)	**76.06%**	**73.80%**	**74.91%**	**63.25%**
(2)	76.04%	73.79%	74.90%	63.17%
(3)	63.58%	60.12%	61.80%	56,92%
(4)	63.31%	60.02%	61.62%	57.83%
(5)	63.83%	60.53%	62.13%	59.05%

Table V. Resultados CBCF

BD	Precisão	σ_p	Revogação	σ_r	F1	Cobertura	σ_c
BD_100	70.64%	7.41%	70.64%	7.41%	70.64%	61.19%	5.35%
BD_90	69.59%	7.18%	69.59%	7.18%	69.59%	61.60%	5.39%
BD_80	64.88%	5.13%	64.88%	5.13%	64.88%	61.25%	4.99%
BD_70	58.46%	4.43%	58.46%	4.43%	58.46%	61.36%	4.65%
BD_60	52.07%	4.75%	52.07%	4.75%	52.07%	60.89%	4.65%
BD_50	50.71%	5.94%	50.71%	5.94%	50.71%	58.55%	4.68%

Table VI. Resultados ***XPrefRec***

BD	Precisão	σ_p	Revogação	σ_r	F1	Cobertura	σ_c
BD_100	76.06%	6.46%	73.80%	8.86%	74.91%	63.22%	4.63%
BD_90	75.38%	6.70%	71.43%	10.65%	73.35%	63.64%	4.73%
BD_80	74.01%	4.73%	65.28%	5.31%	69.37%	67.81%	4.17%
BD_70	71.80%	3.87%	60.80%	4.12%	65.84%	68.26%	4.96%
BD_60	73.94%	3.52%	60.19%	3.61%	66.36%	71.14%	4.02%
BD_50	72.98%	3.24%	59.27%	3.48%	65.41%	74.36%	4.04%

4. CONCLUSÃO

O ***XprefRec*** apresenta uma nova abordagem híbrida de sistema de recomendação trazendo como diferencial alternativas de *extração de consenso* e aplicação de um *minerador de preferências contextuais*. Os testes realizados apontam esta proposta como promissora ao obter resultados de precisão, revocação e cobertura superiores ao *baseline CBCF* utilizado. Como continuação deste trabalho propõe-se testes com outros algoritmos de clusterização bem como uma otimização do algoritmo de mineração de preferências. A otimização, já em fase de implementação, se baseia na aplicação da técnica de *Range Voting* ([Woodall 1994]) após a mineração das preferências. Tal técnica permite uma melhora sensível na revocação de algoritmos de mineração de preferências. Desta forma, acreditamos que melhorando a revocação do *Cprefminer* possamos melhorar a revocação do ***XprefRec***.

REFERENCES

ADOMAVICIUS, G. AND TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.* 17 (6): 734–749, jun, 2005.

CHICLANA, F., HERRERA, F., AND HERRERA-VIEDMA, E. Integrating multiplicative preference relations in a multipurpose decision-making model based on fuzzy preference relations. *Fuzzy Sets and Systems* 122 (2): 277–291, 2001.

COHEN, W. W., SCHAPIRE, R. E., AND SINGER, Y. Learning to order things. *J. Artif. Intell. Res. (JAIR)* vol. 10, pp. 243–270, 1999.

DE AMO, S., BUENO, M. L. P., ALVES, G., AND DA SILVA, N. F. F. Cprefminer: An algorithm for mining user contextual preferences based on bayesian networks. In *ICTAI*. pp. 114–121, 2012.

DE AMO, S., BUENO, M. L. P., ALVES, G., AND DA SILVA, N. F. F. Mining user contextual preferences. *JIDM* 4 (1): 37–46, 2013.

ESTER, M., KRIEGEL, H.-P., SANDER, J., AND XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*. pp. 226–231, 1996.

MELVILLE, P., MOONEY, R. J., AND NAGARAJAN, R. Content-boosted collaborative filtering for improved recommendations. In *Eighteenth national conference on Artificial intelligence*. American Association for Artificial Intelligence, Menlo Park, CA, USA, pp. 187–192, 2002.

RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTROM, P., AND RIEDL, J. Grouplens: An open architecture for collaborative filtering of netnews. In *CSCW*. pp. 175–186, 1994.

WOODALL, D. R. Properties of preferential election rules, 1994.