Universidade Federal de Pernambuco
Centro de Informática

Recomendação para Grupos Através de Filtragem Colaborativa

Trabalho de Graduação em Inteligência Artificial

Sérgio Ricardo de Melo Queiroz

Orientador: Francisco de Assis Tenório de Carvalho

Co-orientador: Geber Lisboa Ramalho

Recife, abril de 2002

Resumo

Nos últimos anos os sistemas de recomendação têm alcançado grande sucesso. Milhares de recomendações são fornecidas diariamente por *sites* como *Amazon.com* e *CDNow*. No entanto, esses sistemas recomendam itens apenas para usuários individuais. Enquanto isso, muitas atividades do dia-a-dia, como ir ao cinema, são realizadas em grupo, trazendo à tona a necessidade de gerar sugestões que atendam a um grupo de pessoas. Neste trabalho o problema de geração de recomendações para grupos é investigado, a utilização de um método baseado em filtragem colaborativa e escolha de alternativas através de maioria fuzzy é proposto como uma abordagem viável para a geração de recomendação para grupos, e o comportamento deste método é avaliado experimentalmente, utilizando dados reais de um sistema de filtragem colaborativa.

Agradecimentos

Minhas palavras de agradecimento vão para o Prof. Francisco de A. T. de Carvalho, pela solícita orientação; ao Prof. Geber Ramalho pelo apoio e debates proporcionados; a todos os professores do Centro de Informática que contribuíram para minha formação, sem eles este trabalho não seria possível; e, finalmente, mas com não menos importância, à minha família por ser o alicerce com o qual sempre posso contar.

Sérgio Queiroz

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O novo milênio é a era da informação. Nos anos 90, houve uma explosão na quantidade de informações disponível. As pessoas podem escolher entre dúzias de canais de TV, milhares de filmes, milhões de CDs, livros e documentos na Internet. No entanto, a nossa capacidade de tratar as informações continua a mesma. Desta forma, as pessoas se vêem náufragas no meio do oceano de informações.

O que então pode ser feito? Quando temos que fazer escolhas sem ter conhecimento total das alternativas, é comum recorrer às opiniões e experiências de outros, seja um crítico de cinema, um amigo, ou um consultor empresarial.

Este cenário proporcionou o florescimento dos sistemas de recomendação computacionais. Tais sistemas se propõem a automatizar o processo de recomendação.

Atualmente temos, principalmente na Web, diversos sistemas de recomendação. Sistemas de recomendação de sites populares como *Amazon.com* e *CDNow* fornecem milhares de recomendações todos os dias. Esses sistemas estão voltados à recomendação para indivíduos. No entanto, várias atividades do dia-a-dia são realizadas em grupo, tais como:

- Assistir à televisão em casa;
- Ir ao cinema com os amigos;
- Escutar rádio durante uma viagem de carro com a família.

Portanto, se quisermos uma recomendação para ir ao cinema com os amigos, esta recomendação deve agradar, preferencialmente, ao grupo como um todo, e não apenas a um indivíduo.

Surge então a necessidade do desenvolvimento de sistemas de recomendação para grupos, capazes de capturar as preferências de grupos inteiros, e realizar recomendações para estes grupos.

Neste trabalho, exploraremos a área de recomendação para grupos, uma área com muito potencial de crescimento e ainda pouco explorada. No Capítulo 2 conheceremos os sistemas de recomendação: o que são, quais as principais técnicas usadas e como pode ser definido o problema de recomendação para grupos. No Capítulo 3 veremos que a recomendação para grupos está relacionada com diversas outras áreas, notadamente a matemática, economia e psicologia. Alguns importantes resultados dessas áreas serão apresentados. No Capítulo 4 será proposto um método para realizar recomendações para grupos, utilizando uma técnica *fuzzy* existente para a classificação de alternativas segundo diferentes opiniões de especialistas. O Capítulo 5 trata de uma avaliação experimental realizada com o método proposto no Capítulo 4. O Capítulo 6 versa sobre as conclusões gerais deste trabalho. Nos Apêndices são encontrados importantes materiais de referência para a leitura do trabalho.

Capítulo 2

Sistemas de Recomendação

2.1 Visão Geral

Muitos exemplos de recomendação ocorrem no dia-a-dia das pessoas. É comum recorrer à leitura de críticas de filmes para decidir qual deles assistir. Ou então pedir a um livreiro uma recomendação de livro para o seu "amigo secreto" fanático por ficção científica. A percepção de um restaurante sempre lotado indica que pode tratar-se de um bom lugar para fazer refeições, então você decide conhecê-lo.

Esses exemplos ajudam a esclarecer o conceito de recomendação. De uma forma genérica, um indivíduo encontra-se diante de uma *decisão*, uma escolha dentre um universo de alternativas. Tal universo é tipicamente enorme, impossibilitando o indivíduo de até mesmo conhecer quais são todas as alternativas possíveis. Portanto a tarefa de escolher entre essas alternativas é extremamente árdua [29].

Para tratar esse problema de *sobrecarga de informação*, existem diferentes técnicas para encontrar as informações desejadas pelo usuário (*filtering in*) e eliminar as indesejadas (*filtering out*). O termo *filtragem de informação* é utilizado para se referir a esses dois atos.

2.1.1 Categorias de Filtragem de Informação

Malone et al. [17] identificaram três categorias de filtragem de informação: cognitiva, social e econômica.

- **Filtragem cognitiva:** seleciona a informação baseada no seu conteúdo. Por exemplo, pode-se relacionar a presença de palavras-chave num artigo com o perfil do usuário.
- **Filtragem social:** é baseada na relação entre as pessoas e seus julgamentos subjetivos. Um filtro "apagar todas as mensagens cujo remetente é *Fulano*" é um exemplo simples de filtragem social.
- **Filtragem econômica:** é baseada na relação custo/benefício de produção e consumo de um item. Um filtro econômico de e-mail poderia usar as regras: *uma mensagem postada para muitos destinatários tem um baixo custo de produção por endereço, então deve receber baixa prioridade*; *por outro lado, uma mensagem enviada exclusivamente para o endereço de um usuário tem alto custo de produção, então deve receber maior prioridade.*

2.1.2 Sistemas de Recomendação Computacionais

Na década de 90, floresceram os sistemas de recomendação computacionais que automatizam o processo de recomendação através da aplicação de técnicas de filtragem.

Os sistemas de recomendação computacionais (a partir daqui chamados apenas de sistemas de recomendação) baseiam-se principalmente em duas das técnicas de filtragem de informação definidas na Seção 2.1.1: a filtragem cognitiva, ou baseada no conteúdo; e a filtragem social.

Os sistemas baseados no conteúdo utilizam apenas as preferências do usuário; eles tentam recomendar itens que são similares aos que o usuário gostou no passado. O foco desses sistemas é aprender as

preferências do usuário e filtrar dentre os novos itens aqueles que mais se adequam a estas preferências. A Seção 2.2 trata de recomendação baseada no conteúdo.

Os sistemas de filtragem social utilizam a técnica de *filtragem colaborativa*. O foco dessa técnica é encontrar os usuários de gostos similares ao do usuário em questão, para recomendar itens que estes "vizinhos" gostaram. A filtragem colaborativa é atualmente utilizada em muitas aplicações na web. *Sites* de comércio eletrônico como Amazon.com e CDNow possuem áreas de recomendação, onde, além de críticas de especialistas, os usuários podem dar notas a itens e então receber recomendações personalizadas fornecidas por um engenho de filtragem colaborativa. A Seção 2.3 trata de filtragem colaborativa.

2.2 Recomendação Baseada no Conteúdo

Os sistemas de recomendação baseados no conteúdo [16] encontram itens parecidos com os que o usuário gostou no passado. As preferências do usuário são aprendidas através de *feedback* fornecido por ele. O *feedback* pode ser *explícito* (por exemplo, o usuário pode fornecer uma nota ao item) ou *implícito* (por exemplo, o tempo que o usuário passou visualizando uma página pode ser utilizado para medir o seu nível de interesse nela). As preferências são representadas como um *perfil* do usuário, que reflete os seus interesses em determinados tipos de conteúdo. Por exemplo, é comum a representação do perfil como um vetor de palavras com pesos associados. Técnicas de aprendizagem de máquina e recuperação de informação são utilizadas para aprender e representar o perfil do usuário.

As técnicas de recomendação baseadas no conteúdo estão fora do escopo deste trabalho, mas são úteis como contraste para ajudar a esclarecer as técnicas das quais trataremos, os sistemas de recomendação baseados em filtragem colaborativa. Estes criam uma interação (visível ou não) entre o usuário que busca uma recomendação e um conjunto de pessoas que expressaram suas preferências sobre itens.

2.3 Filtragem Colaborativa

Embora a filtragem baseada no conteúdo venha sendo usada com sucesso em vários domínios, essa técnica apresenta uma série de limitações [25]:

- O conteúdo dos itens deve ser passível de manipulação pelo computador (por exemplo, textual), ou atributos devem ser manualmente cadastrados para os itens. Com a tecnologia atual, mídias como som e vídeo apresentam grande dificuldade de serem analisados automaticamente para extração automática de atributos. Muitas vezes não é possível definir os atributos manualmente devido a limitações de recursos.
- As técnicas de filtragem baseadas no conteúdo não têm como encontrar itens que interessariam ao usuário mas não são parecidos (no conteúdo) com outros itens que ele avaliou, isto é, apenas são encontrados os itens parecidos com os já conhecidos pelo usuário.
- Os métodos baseados no conteúdo não são capazes de avaliá-lo quanto a dimensões subjetivas, como qualidade. Por exemplo, não é possível diferenciar um texto mal escrito de um bem escrito com conteúdos muito semelhantes.

A técnica de filtragem colaborativa baseia-se no fato de que as melhores recomendações para um indivíduo são aquelas fornecidas por pessoas que possuem gostos similares aos dele. O processo de filtragem colaborativa pode ser descrito em três passos (Figura 2.1):

- **Representação dos dados de entrada:** o usuário expressa suas preferências fornecendo notas a itens propostos pelo sistema. Estas avaliações (positivas e negativas) indicam os interesses do usuário em itens específicos, e são armazenadas como o perfil do usuário. A forma mais simples de armazenar o perfil é como uma matriz de m itens $\times$ n usuários, onde as células contêm as avaliações fornecidas. Note que as avaliações também podem ser fornecidas de forma implícita, por exemplo um *site* de comércio eletrônico pode considerar que o usuário demonstra interesse por um produto ao comprá-lo.

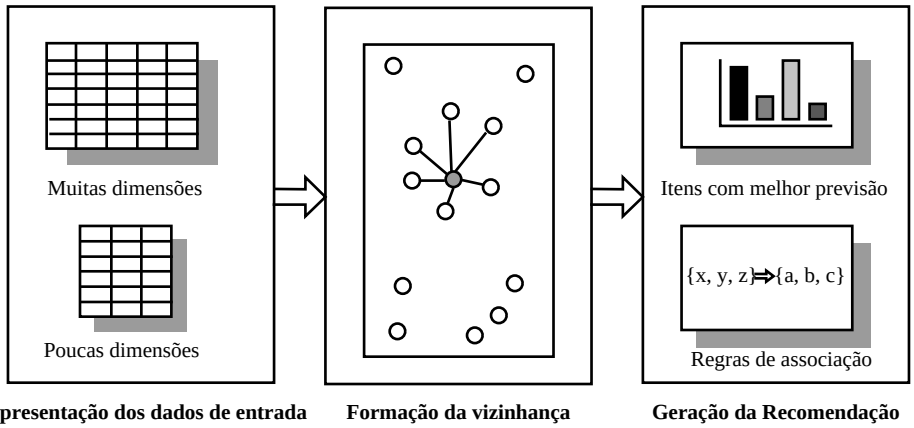

Figura 2.1: Principais passos do processo de recomendação com filtragem colaborativa

Id da mensagem	Ken	Lee	Meg	Nan
1	1	4	2	2
2	5	2	4	4
3			3	
4	2	5		5
5	4	1		1
6	?	2	5	

Figura 2.2: Exemplo de uma matriz de notas do GroupLens

- **Formação da vizinhança:** para fornecer uma recomendação, o sistema compara este perfil com os perfis dos outros usuários, e atribui um peso a estes de acordo com o grau de similaridade com o perfil do usuário que receberá a recomendação. A métrica usada para determinar a similaridade pode variar. O resultado desta comparação é um conjunto dos "vizinhos mais próximos" do usuário. Este conjunto formaliza o conceito de pessoas com gostos similares.
- **Geração da recomendação:** finalmente, o sistema usa a informação contida no conjunto dos vizinhos mais próximos para recomendar itens para o usuário, isto é, os itens mais apreciados pelos vizinhos serão recomendados.

Para ilustrar o processo de filtragem colaborativa, veremos como ele se desenvolve no GroupLens [21].

2.3.1 Caso de Estudo: GroupLens

O GroupLens é um sistema de filtragem colaborativa para Usenet (grupos de notícias da Internet). O seu objetivo é prever o quanto irá interessar ao usuário cada artigo presente num grupo de notícias.

Ao usar um leitor de notícias compatível com o GroupLens, os usuários (identificados por um pseudônimo) podem fornecer notas aos artigos lidos. As notas variam de 1 a 5, onde 1 é a nota mais baixa e 5 a mais alta.

A Figura 2.2 mostra uma matriz de notas fornecidas a 6 mensagens pelos usuários Ken, Lee, Meg e Nan. As células não preenchidas da matriz significam que o usuário não forneceu uma nota para tal mensagem. Prever o quanto um artigo irá interessar a um usuário significa prever a nota que este usuário forneceria a este artigo ainda não visto.

Para implementar a heurística da filtragem colaborativa, isto é, de que as pessoas que demonstraram gostos similares no passado provavelmente irão concordar novamente, o primeiro passo é correlacionar as notas fornecidas anteriormente para determinar os pesos que serão atribuídos a cada pessoa no cálculo da nota prevista.

Assim, é computado o coeficiente de correlação entre o usuário x para o qual se quer prever a nota e cada um dos outros usuários do sistema (y) que avaliou o item em questão. O coeficiente de correlação entre dois usuários é computado como:

$$\begin{aligned} \rho_{xy} &= \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \\ &= \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}} \end{aligned} \tag{2.1}$$

Na fórmula acima, $\overline{x}$ é a média das notas do usuário x. Todas os somatórios e médias da fórmula são feitos apenas para aqueles itens que ambos os usuários avaliaram.

Vamos supor que quiséssemos prever a nota de Ken para a mensagem 6 (célula com "?" da Figura 2.2). Primeiramente calculamos o coeficiente de correlação de Ken com cada um dos outros usuários. A correlação de Ken com Lee de acordo com a equação 2.1 fica:

$$\begin{aligned} \overline{K} &= \frac{1+5+2+4}{4} = 3 \\ \overline{L} &= \frac{4+2+5+1}{4} = 3 \\ \rho_{KL} &= \frac{(1-3)(4-3)+(5-3)(2-3)+(2-3)(5-3)+(4-3)(1-3)}{\sqrt{(1-3)^2+(5-3)^2+(2-3)^2+(4-3)^2}\sqrt{(4-3)^2+(2-3)^2+(5-3)^2+(1-3)^2}} \\ &= \frac{-2-2-2-2}{\sqrt{10}\sqrt{10}} = -0,8 \end{aligned}$$

Similarmente, o coeficiente de correlação de Ken com Meg é $+1$ e com Nan é 0. Isto é, Ken costuma discordar de Lee ($\rho_{KL} = -0,8$) e concordar com Meg ($\rho_{KM} = +1$). As suas avaliações não estão correlacionadas com as de Nan.

Para prever a nota de x para o artigo i, tomamos uma média ponderada de todas as avaliações do artigo i de acordo com a fórmula:

$$x_{i_{\text{Pred}}} = \overline{x} + \frac{\sum_{y \in \text{avaliadores}} (y_i - \overline{y})\rho_{xy}}{\sum_y |\rho_{xy}|} \tag{2.2}$$

No exemplo de Ken, a nota prevista para a mensagem 6 é:

$$K_{6_{\text{Pred}}} = 3 + \frac{2\rho_{KM} - \rho_{KL}}{|\rho_{KM}| + |\rho_{KL}|} = 3 + \frac{2-(-0,8)}{|1| + |-0,8|} = 4,56$$

Esta é uma previsão razoável para Ken, pois como podemos ver na Figura 2.2 a mensagem 6 recebeu uma nota alta de alguém que normalmente concorda com ele e uma nota baixa de alguém que normalmente discorda.

2.3.2 Problemas da Filtragem Colaborativa e Integração com Filtragem Baseada no Conteúdo

Apesar dos elogios recebidos dos usuários de sistemas de recomendação baseados em filtragem colaborativa [25], esta técnica possui algumas limitações, das quais destacam-se:

- **Recomendação de itens novos:** até que um item tenha sido avaliado por um número mínimo de usuários, não é possível recomendá-lo, pois o sistema não tem dados suficientes para prever uma nota para ele.
- **Usuário "ovelha negra":** caso o usuário em busca de recomendações não tenha vizinhos próximos o suficiente, o sistema apresentará um baixo desempenho, pois as recomendações serão baseadas em usuários que não se parecem muito com ele.

Tabela 2.1: Filtragem baseada no conteúdo *versus* filtragem colaborativa

Característica	Filtragem Baseada no Conteúdo	Filtragem Colaborativa
Recomendação de itens novos	Não apresenta problemas, pois usa o conteúdo do item para identificar se o usuário irá apreciá-lo.	Não podem ser recomendados até que sejam avaliados por um número mínimo de usuários.
Usuário "ovelha negra"	Não tem problemas, a recomendação é baseada inteiramente no perfil do usuário.	Baixa performance, pois não será possível encontrar vizinhos próximos o suficiente para gerar boas recomendações.
Pequeno número de usuários	Performance independente do número de usuários.	Baixa performance, difícil de encontrar vizinhos adequados.
Conteúdo não interpretável pelo computador (ex.: multimídia)	Necessidade de cadastrar atributos manualmente, o que pode inviabilizar o sistema de recomendação.	Não tem problemas, a recomendação é totalmente baseada no julgamento das pessoas e as relações entre elas. O conteúdo dos itens não precisa ser conhecido.
Avaliação em dimensões subjetivas	Muito difícil de implementar.	Intrínseco no julgamento das pessoas. Nenhuma necessidade de modificação.
Recomendação de itens interessantes de conteúdo diferente do já visto pelo usuário	Normalmente não ocorre, pois os itens recomendados são aqueles parecidos com os que o usuário já viu no passado.	Fácil de ocorrer. Os itens bem avaliados pelos vizinhos mais próximos podem ser de conteúdo bastante diferente do que o usuário já conhece.

- **Número de usuários insuficiente:** para ter uma boa performance, um sistema de filtragem colaborativa necessita de uma grande comunidade de usuários, pois de outra forma não haverá vizinhos suficientemente próximos de cada usuário.

Para vencer estas limitações, vários sistemas [4, 24, 26] têm adotado com sucesso uma estratégia híbrida, combinando a filtragem colaborativa com a baseada no conteúdo, uma vez que a última não possui essas dificuldades. Na verdade, a filtragem colaborativa é bastante complementar à baseada no conteúdo, como mostra a Tabela 2.1. Desta forma, uma estratégia híbrida pode usar o que cada método tem de melhor.

Embora métodos de correlação estatística (como o visto na Seção 2.3.1) tenham sido usados em todos os sistemas pioneiros de recomendação por filtragem colaborativa, tais como Ringo [25], Bellcore Video Recommender [9] e GroupLens [21]; foram identificadas deficiências nesses métodos [23]:

- **Cobertura reduzida:** sistemas de recomendação comerciais são usados para avaliar um conjunto enorme de produtos (por exemplo, os livros da Amazon.com ou os álbuns da CDNow). Nesses sistemas, mesmos os consumidores mais freqüentes avaliaram bem menos de 1% dos produtos (1% de 2 milhões de livros são 20.000 livros). Desta forma, o sistema de recomendação pode ser completamente incapaz de encontrar qualquer recomendação para um dado usuário (devido à esparsidade das avaliações) ou a qualidade das recomendações pode ser muito baixa. Uma característica que pode claramente melhorar a qualidade das recomendações é a transitividade de vizinhos. No entanto, se os usuários Pedro e Helena tem uma correlação alta, e Helena também tem uma correlação alta com Mateus, não é necessariamente verdade que Pedro e Mateus terão uma correlação alta, pois eles podem ter avaliado poucos itens em comum.

- **Escalabilidade:** para a formação da vizinhança, é necessário um número de operações que cresce

tanto com o número de usuários quanto com o número de itens. Com milhões de usuários e itens, um sistema de recomendação típico sofrerá de sérios problemas de escalabilidade.

- **Sinônimos:** em um cenário real, diferentes nomes de produtos podem se referir a objetos similares. Os métodos de correlação não conseguem enxergar essa associação, considerando cada produto diferentemente. Por exemplo, se o consumidor *A* comprar 10 *cadernos universitários de 12 matérias* e o consumidor *B* comprar 10 *cadernos universitários de 10 matérias*, um sistema de recomendação tradicional não é capaz de descobrir a associação desses itens como "cadernos universitários".

Esses problemas apontam para as limitações dos métodos de filtragem colaborativa baseados em correlação estatística. Para tentar superar esses problemas, novos métodos vêm sendo estudados.

Billsus e Pazzani [5] observaram que o problema de prever itens para um usuário baseado nas notas de outros usuários para esses itens pode ser descrito como um problema de *classificação*, uma tarefa bem estudada na área de aprendizagem de máquina. Desta forma foi possível explorar a "estrutura latente" das avaliações, diminuindo a necessidade dos usuários possuírem muitos itens em comum para prever itens. Os experimentos realizados mostraram que a abordagem por eles utilizada (com redes neurais) obteve uma performance melhor que os métodos tradicionais baseados em correlação.

Aggarwal et al. [1] desenvolveram uma nova técnica de filtragem colaborativa baseada em grafos que melhora sensivelmente a qualidade das recomendações na presença de dados esparsos.

Lin, Alvarez e Ruiz [15] propuseram um método de realizar filtragem colaborativa baseado na técnica de regras de associação utilizada em mineração de dados [2]. Segundo os autores, essa técnica é capaz de identificar associações não visíveis para os métodos baseados em correlação. Resultados experimentais mostraram uma melhor performance em relação a estes métodos, no entanto não foi possível concluir se houve melhora em relação ao método proposto por Billsus e Pazzani.

Novos experimentos com essas técnicas poderão trazer contribuições bastante relevantes à área.

2.4 Sistemas de Recomendação para Grupos

Com as técnicas descritas nas Seções 2.2 e 2.3, torna-se possível a recomendação para indivíduos. De fato, como mencionado anteriormente, um bom número de sistemas de recomendação vêm sendo usados com sucesso.

Esses sistemas, no entanto, são usados para recomendar itens para um único usuário por vez. Todavia, muitas atividades do dia-a-dia são realizadas em grupos, tais como:

- Assistir à televisão em casa;
- Ir ao cinema com os amigos;
- Escutar rádio durante uma viagem de carro com a família.

Essas atividades alargam o horizonte dos sistemas de recomendação para a sugestão para grupos. Por exemplo, com o advento da televisão interativa, torna-se possível o processo de transmissão de conteúdo direcionado para cada telespectador, ao invés do processo de *broadcast* atual. Neste novo cenário, a personalização será um componente chave para o sucesso da televisão interativa [27].

Portanto, no caso usual de várias pessoas assistirem à TV, esta deverá ser capaz de filtrar o conteúdo para um grupo, ao invés de um único indivíduo.

2.4.1 Recomendação para Grupos Usando Filtragem Colaborativa

O problema de produzir recomendações para grupos baseando-se em filtragem colaborativa pode ser descrito da seguinte forma:

> Como sugerir (novos) itens que serão apreciados pelo grupo, dado que possuímos um histórico das preferências individuais de membros deste grupo bem como das preferências de outros indivíduos (que não fazem parte do grupo).

Esta sugestão deve ser idealmente a melhor possível para o grupo. Então, duas questões centrais são levantadas:

- O que é a melhor sugestão para um grupo?
- Como alcançar esta sugestão?

No âmbito da filtragem colaborativa, uma forma de transpormos estas questões é:

> Dado que, usando filtragem colaborativa, possuímos a previsão de quanto cada indivíduo irá gostar de cada um dos itens que podem ser recomendados, como agregar essas previsões de forma a obter uma boa sugestão para o grupo?

No próximo capítulo, veremos que esta questão é bastante complexa, e vem sendo estudada a séculos por matemáticos, economistas e psicólogos.

Trabalhos Relacionados

O conceito de fazer recomendações para grupos tem recebido pouca atenção na literatura de sistemas de recomendação. Um dos objetivos na concepção da "comunidade virtual" de Hill et al. [9] é que as recomendações deveriam ser fornecidas para conjuntos de pessoas, não apenas indivíduos. Ainda assim, eles não trataram das dificuldades envolvidas para se alcançar boas recomendações para grupos (isto é, as duas questões fundamentais citadas na Seção 2.4.1.

O Let's Browse [14] é um agente de navegação colaborativa para Web que recomenda páginas de interesse comum àqueles presentes numa sessão de navegação coletiva. Para tal, um perfil constituído de um vetor de palavras com pesos é previamente construído para cada um dos usuários. Este perfil é gerado de forma automática a partir de uma busca em largura (com profundidade pré-estabelecida) iniciada na *homepage* pessoal de cada usuário. É então gerado um perfil para o grupo, que é simplesmente uma combinação linear simples dos perfis dos usuários. Uma página apontada pela página atual de navegação é recomendada se ela obtiver uma similaridade com o perfil do grupo acima de um limiar determinado. Desta forma, o Let's Browse é um sistema de recomendação para grupos baseado no conteúdo, herdando as limitações desta técnica já levantadas. Além disso, possui uma estratégia fixa para gerar recomendações para o grupo de usuários (a combinação linear simples – média – dos perfis). De acordo com os autores, os participantes do sistema forneceram opiniões favoráveis para o sistema, embora não tenham sido conduzidos experimentos controlados.

Capítulo 3

Tomada de Decisão para Grupos

3.1 Visão Geral

Pesquisas em tomada de decisão para grupos têm diversas raízes distintas. Além de abordagens da psicologia e sociologia que estudam como as pessoas chegam a decisões coletivas, a tomada de decisão por grupos é um dos principais tópicos de estudo em uma área interdisciplinar, que combina economia e ciência política, chamada de escolha social (*social choice*). Essas abordagens diferem em como e o que focar (por exemplo, ênfase empírica *versus* analítica, consenso *versus* escolha) e, principalmente no tipo de entrada usado para se chegar à escolha coletiva (por exemplo, ordem de preferências, intensidade de preferências, justificativas, argumentações etc.) [11].

Muito embora os principais desenvolvimentos da psicologia e escolha social na área de tomada de decisão para grupos tenham ocorrido já no século XX, matemáticos motivados pelos problemas envolvidos em votações, vêm estudando este problema há mais de dois séculos.

Mas, afinal, o que pode ser mais simples do que uma votação? Simplesmente contar quantos votos cada candidato recebeu não é o suficiente para declarar o vencedor? O que pode haver de errado em algo tão elementar quanto isso?

Na verdade, durante todo este tempo, matemáticos têm mostrado que quando há pelo menos 3 candidatos – uma situação comum – o vencedor pode não ser necessariamente o preferido pela maioria. E isto não ocorre apenas porque alguns eleitores continuam votando mesmo depois de morrerem; resultados distorcidos também podem ser causados por peculiaridades matemáticas [22].

Para ilustrar o problema, conhecido como "paradoxo da votação" [3], vejamos um exemplo, adaptado de [22]. Um país resolve fazer uma pesquisa para determinar qual iniciativa contra a "nação inimiga do momento" tem maior aprovação da população. Nesta pesquisa são entrevistadas 1500 pessoas que devem indicar sua preferência entre N (Negociações diplomáticas), E (Embargo comercial) e G (Guerra). As preferências obtidas podem ser vistas na Tabela 3.1.

De acordo com a Tabela 3.1, o resultado utilizando-se pluralidade (onde cada pessoa vota na sua medida predileta) é $G \prec E \prec N$ com um resultado de 600:500:400. Aparentemente, Guerra é a escolha da população.

Antes de enviar os fuzileiros, vamos ver se Guerra é realmente a opção preferida pela população. Se isto for verdade, é esperado que a população prefira Guerra a Embargo comercial. Todavia, como mostra a Tabela 3.2, os entrevistados preferem o Embargo comercial à Guerra.

Tabela 3.1: Preferências da população, onde "$\prec$" significa "é preferido a"

Número de Pessoas	Preferências
600	$G \prec N \prec E$
500	$E \prec N \prec G$
400	$N \prec E \prec G$

Tabela 3.2: Comparando Guerra com Embargo comercial segundo as preferências da população

Número de Pessoas	Preferências	Guerra	Embargo comercial
600	$G \prec N \prec E$	600	0
500	$E \prec N \prec G$	0	500
400	$N \prec E \prec G$	0	400
	Total	600	900

Tabela 3.3: Preferências dos membros do departamento

Número de Pessoas	Preferências
5	$A \prec I \prec M$
5	$I \prec M \prec A$
5	$M \prec A \prec I$

Da mesma forma, 900 entrevistados preferem Negociações diplomáticas à Guerra e 1000 preferem Negociações diplomáticas ao Embargo comercial. Isto causa uma contradição em relação ao resultado obtido através da pluralidade, pois essas comparações par a par indicam que a opinião da população é na verdade $N \prec E \prec G$, o *ranking* oposto ao resultado da pluralidade.

Na década de 1780, o matemático, filósofo e político francês Marie-Jean-Antoine-Nicolas de Caritat Condorcet argumentou que os resultados de eleições deveriam ser estritamente estabelecidos por votações par a par. O *vencedor de Condorcet* é o que ganha de todos os outros candidatos nas votações par a par. Com os dados da Tabela 3.1, Negociações diplomáticas é o vencedor de Condorcet, enquanto que Guerra é o perdedor de Condorcet.

O vencedor de Condorcet é bastante aceito como **o** vencedor dentre os candidatos. No entanto, ainda há problemas. Para ilustrar apenas uma dificuldade, suponha que um departamento de computação use votações par a par para decidir que livro de Inteligência Artificial adotar dentre as alternativas $\{A, I, M\}$[1]. Uma maneira natural de selecionar o vencedor de Condorcet é por eliminação, onde após comparar duas escolhas, digamos $\{A, I\}$, o vencedor é comparado com a opção restante, M. Vamos supor que as opiniões dos membros do departamento são como na Tabela 3.3.

Como se vê na Tabela 3.4, A vence a comparação inicial $\{A, I\}$, mas é derrotado na comparação seguinte para M. Em ambas as comparações o vencedor ganha com dois terços dos votos, então parece claro que a opinião do departamento é $M \prec A \prec I$.

Verifiquemos se o resultado é mesmo inquestionável. Nós já verificamos que M vence A e que A vence I. Falta então verificar se M (a "opção preferida") vence I (a "opção menos popular"). Parece óbvio que isso irá ocorrer, mas ao contrário do esperado, I vence M com os mesmos dois terços dos votos. Em outras palavras, essas opiniões definem um resultado de votação cíclico: $A \prec I$, $I \prec M$, $M \prec A$. O candidato que for considerado por último sempre vencerá a eleição. Não há vencedor ou perdedor de Condorcet.

Condorcet identificou a possibilidade de ciclos, demonstrando esse comportamento com os dados do exemplo citado. Um exemplo com esta característica é conhecido como "perfil de Condorcet".

[1] Exemplo retirado de [22].

Tabela 3.4: Comparações par a par dos livros

Número de Pessoas	Preferências	A	I	A	M
5	$A \prec I \prec M$	5	0	5	0
5	$I \prec M \prec A$	0	5	5	0
5	$M \prec A \prec I$	5	0	0	5
	Totais	10	5	5	10

Ciclos tornam impossível a escolha de um "candidato ótimo", e são mais um exemplo que mostra a dificuldade de alcançar-se um ranking de opções ótimo para um grupo.

Arrow, num trabalho histórico [3] chegou a importantes conclusões quanto a impossibilidade de se atingir um resultado ótimo para grupos, estabelecendo assim um marco fundamental da *escolha social*. Na próxima seção veremos os principais resultados alcançados por Arrow.

3.2 Escolha Social

Encontrar uma maneira de agregar escolhas individuais de indivíduos para encontrar a melhor solução para um grupo pode ser visto como um problema de alcançar um máximo social a partir de desejos individuais. Este é o problema central da "economia do bem-estar" (*welfare economics*).

Um pré-requisito para obter uma função de utilidade social é a possibilidade de comparação entre os estados sociais (utilidades individuais) de cada indivíduo. Isto por si só é um ponto de discussão entre os economistas. No entanto, mesmo quando é considerado que esta comparação é possível, há um julgamento de valores implícito quando uma função de utilidade social é utilizada. Por exemplo, esta função pode ser uma soma das utilidades individuais, ou seu produto, ou o produto de seus logaritmos, ou a soma dos produtos dois-a-dois etc. [3].

Tendo em vista este resultado, Arrow não tenta encontrar uma função social ideal Ao invés disso, são identificados valores básicos que uma função social deve atender e estudada a possibilidade destes valores serem alcançados.

3.2.1 Definições

As relações entre alternativas podem ser de preferência ou indiferença. Ao invés de usar duas relações, uma única relação é usada para indicar "preferido ou indiferente". A afirmação "x é preferido ou indiferente a y" é simbolizada como $x\,R\,y$. A notação R_i é usada para representar a relação de ordem do ponto de vista do usuário i sobre o conjunto de alternativas X, enquanto que a relação de ordem da sociedade como um todo será representada por R.

R é uma relação conexa e transitiva. Simbolicamente,

Axioma 1 *Para todo x e y, ou $x\,R\,y$ ou $y\,R\,x$.*

Axioma 2 *Para todo x, y e z, se $x\,R\,y$ e $y\,R\,z$ então $x\,R\,z$.*

R é dita uma relação de ordem fraca. O adjetivo "fraco" significa que a ordem não exclui a possibilidade de indiferença, isto é, os Axiomas 1 e 2 não impedem que para x e y distintos, $x\,R\,y$ e $y\,R\,x$.

P é a relação de preferência. $x\,P\,y$ é definido como não $y\,R\,x$.

Uma função social recebe uma n-tupla de relações de preferência individuais e atribui uma relação de preferência global. Mais formalmente, temos $f : R_1 \times \ldots \times R_n \mapsto R$.

3.2.2 Propriedades Desejadas para uma Função Social

Arrow identificou as seguintes propriedades desejadas para uma função social:

1. **Domínio irrestrito:** f possui domínio irrestrito se, e somente se é definida para todo o produto cartesiano.
2. **Independência de alternativas irrelevantes:** a relação de preferência social entre qualquer par de alternativas x e y depende apenas das relações de preferência individuais entre estas duas alternativas.
3. **Condição Pareto:** se para todos os indivíduos i, $x\,P_i\,y$ então $x\,R\,y$.
4. **Não-ditadura:** para todo x e y em X não há um indivíduo i tal que $x\,R\,y$ se, e somente se $x\,R_i\,y$.

3.2.3 Teorema da Impossibilidade de Arrow

Arrow demonstrou que não existe uma função social com todas as propriedades descritas na Seção 3.2.2. Desta forma, qualquer método de decisão social terá que abdicar de alguma das condições desejáveis.

Não há então uma maneira ideal para se agregar preferências individuais alcançando-se um resultado global. Desta forma, qualquer método apresentará deficiências, como as deficiências dos métodos da pluralidade e de Condorcet, discutidos na Seção 3.1.

3.3 Psicologia Social

Uma área da psicologia, a psicologia social, também vem estudando o problema de tomada de decisão em grupos. Uma de suas principais preocupações é a de entender como características individuais são combinadas para gerar um resultado de grupo [13].

3.3.1 Teoria dos Esquemas de Decisão Social

A teoria dos esquemas de decisão social (*social decision schemes* – SDS) é bastante usada para alcançar respostas de grupo a partir de preferências individuais. Isto envolve três considerações centrais: a distribuição das preferências dos membros do grupo, a regra que combina estas preferências (chamada de esquema de decisão), e a maneira de testar a adequação dos esquemas de decisão na previsão do comportamento observado por uma amostra de decisões de grupos reais (teste do modelo) [10].

Distribuição das Preferências

O modelo SDS genérico assume que cada membro do grupo, e subseqüentemente cada grupo, seleciona uma dentre n alternativas exaustivas e mutuamente excludentes. Para um grupo com r membros (ou residentes), a distribuição destes membros em relação às n alternativas pode ser resumida pelo vetor $(r_1, r_2, \ldots, r_n)$, onde r_i indica o número de elementos do grupo que prefere a alternativa i. Note que esta representação permite que as respostas do grupo sejam distinguidas, mas não os membros individuais do grupo.

Esquemas de Decisão

Com a distribuição dos membros em relação às alternativas (verificada ou prevista a partir da distribuição das preferências da população e considerando uma formação aleatória dos grupos), é possível aplicar uma regra ou procedimento que combine as preferências em uma resposta do grupo. Esta regra ou procedimento é chamado de esquema de decisão.

Esquemas de decisão podem ser construídos para representar vários processos sociais que hipoteticamente dirigiram a decisão do grupo. Nos grupos com maior interação, esses processos de decisão são tipicamente complexos e implícitos. As pesquisas tentam identificar esquemas de decisão plausíveis que possam descrever as regras implícitas de decisão em termos implícitos, normalmente algébricos. Um esquema de decisão reflete as normas ou regras através das quais o consenso e acordo são atingidos num grupo.

Teste do Modelo

O esquema de decisão utilizado pode ser testado comparando-se as decisões observadas nos grupos reais com as obtidas utilizando-se o esquema proposto. Caso a distribuição das respostas previstas para os grupos não diferirem significativamente das respostas reais observadas nos grupos amostrais, o esquema de decisão utilizado pode ser considerado uma descrição plausível do processo pelo qual a decisão dos grupos foi alcançada. No entanto, se as duas distribuições diferirem significativamente, o esquema de decisão é eliminado como uma descrição plausível do processo de decisão de acordo com as condições dos grupos amostrais.

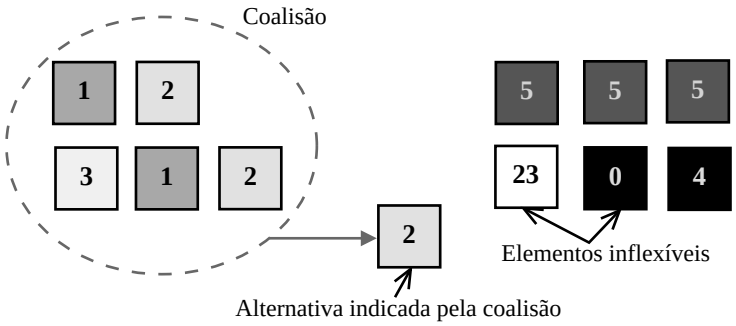

Figura 3.1: Coalisão formada pela maioria dos membros com preferências similares domina o processo de decisão.

Esquemas de Decisão Usados na Psicologia

Resultados empíricos mostram que a adequação de um esquema de decisão é dependente do problema envolvido, das opiniões e características dos membros dos grupos e das informações conhecidas por eles. Uma lista parcial de esquemas de decisão utilizados na psicologia é mostrada a seguir [10].

- **Esquemas de decisão baseados em tendência central**
  - **Média:** uma maneira óbvia de encontrar uma preferência para o grupo é simplesmente tomar a média das preferências individuais. No entanto, esta solução muitas vezes representa uma posição em que todo o grupo está "abrindo mão" de suas preferências, não agradando realmente a ninguém.
  - **Mediana:** similar à média, porém menos sensível a posições extremas.
- **Esquemas de decisão baseados em consenso**
  - **Maioria:** a alternativa escolhida é aquela apoiada por pelo menos a maioria dos membros do grupo. Pesquisas em SDS mostraram um bom suporte à alternativa da maioria. No entanto, também foi verificado que em grupos com opiniões muito variadas ela pode levar a uma escolha inferior, ao invés de uma integração dos interesses [19].
  - **Pluralidade:** quando não existe maioria, seleciona a alternativa com maior apoio.
- **Esquema de decisão "atração por facções":** membros do grupo são atraídos para alternativas que possuam uma porção substancial do grupo suportando-a (facção). À medida que o tamanho da facção favorável a uma alternativa cresce, o impacto da facção no processo de decisão do grupo também aumenta. Uma versão deste esquema de decisão que recebeu algum suporte empírico determina que o impacto da facção no processo de decisão do grupo é o quadrado do número de indivíduos na facção.
- **Esquema de decisão baseado em coalisões**
  - **Coalisão diferença mínima/maioria:** a coalisão formada pela maioria dos membros que têm a menor variabilidade de preferências domina o processo de decisão do grupo. A Figura 3.1 mostra que uma maioria que tinha uma diferença de opiniões pequena, consegue dominar o processo de decisão, formando uma coalisão para indicar a alternativa "2". A alternativa escolhida por pluralidade seria a "5", mas os defensores dessa alternativa não conseguiram formar uma coalisão, pois os defensores das alternativas "23" e "0" eram muito diferentes e inflexíveis.
- **Esquemas de decisão influenciados pela distância**

  - **Diretamente Proporcional:** a influência que um membro do grupo tem na decisão final é proporcional à distância entre as preferências individuais originais dele e a preferência "média" do grupo (quanto mais próximo da média, maior a influência). Baseia-se no fato de que o grupo tem tendência a não dar ouvidos aos membros com preferências mais distantes.
  - **Inversamente Proporcional:** o impacto das preferências de um membro do grupo na decisão final é inversamente proporcional à distância entre as preferências individuais dele e a preferência "média do grupo" (quando mais longe da média, maior a influência). Baseia-se na premissa de que os membros mais extremos são aqueles mais confiantes, dispostos a argumentar e inflexíveis.

- **Esquemas de decisão ditatoriais:** alguns indivíduos ("ditadores") podem ter um tremendo impacto no resultado final do grupo.
  - **Esquema de decisão do membro mais capaz:** em alguns tipos de situação, a presença de um membro mais capaz é decisiva para a resposta do grupo. Por exemplo, na resolução de problemas do tipo *puzzle*, uma vez descoberta a solução por um membro do grupo, todos os outros são facilmente convencidos dela.
  - **Esquema de decisão do membro menos capaz:** em outros tipos de problema, um membro menos capaz consegue levar o grupo para uma alternativa errada, prejudicando a performance do grupo.

3.4 Conseqüências

O teorema da impossibilidade de Arrow (Seção 3.2.3) nos mostra que não existe nenhum método totalmente satisfatório na tomada de decisão para grupos. Já os trabalhos na psicologia demonstram que a adequação de um esquema de decisão a o processo decisório de um grupo é totalmente dependente das características intrínsecas ao grupo (personalidade das pessoas, nível de conhecimento, motivações, julgamentos pessoais) e da natureza do problema (*puzzle*, problema analítico, decisão de júri). Portanto conclui-se que para a recomendação para grupos é necessário um método decisório flexível, que possa ser parametrizado (possivelmente pelo usuário) para melhor se adequar ao grupo, refletindo seus desejos.

Assim, no próximo capítulo, abordaremos um método *fuzzy* originalmente proposto para decisões envolvendo múltiplas pessoas em sistemas especialistas e o traremos para a área de recomendação para grupos. Este método permite um bom grau de flexibilidade quanto à regra de decisão, através do uso de quantificadores *fuzzy*.

Em seguida, no Capítulo 5 o comportamento deste método será avaliado experimentalmente em relação a diferentes tipos de grupo.

Capítulo 4

Recomendação para Grupos Baseada em Maioria *Fuzzy*

4.1 Visão Geral

Ambientes imprecisos normalmente fornecem uma boa oportunidade para a aplicação de metodologias *fuzzy*. Estas podem ser usadas para transformar especificações vagas de um problema em relações fuzzy (objetivos fuzzy, restrições fuzzy, preferências fuzzy, ...). Dada a natureza imprecisa no processo de escolha para grupos, a aplicação da noção de maioria fuzzy é bastante apropriada. A maioria fuzzy proporciona um *framework* com maior "consistência humana" no processo de escolha, dado o uso de quantificadores fuzzy lingüísticos que expressam o discurso humano.

Neste capítulo vamos inicialmente apresentar o método de classificação de alternativas baseado em maioria fuzzy proposto por Chiclana et al. [7], para o contexto de sistemas especialistas. Em seguida será mostrado através de um exemplo como este método pode ser usado para obter uma recomendação para grupos a partir de sugestões individuais geradas por um sistema de filtragem colaborativa.

4.2 Maioria Fuzzy

Tradicionalmente a maioria é definida como um número limiar da quantidade de indivíduos. Por exemplo, para 10 indivíduos podemos definir o limiar para maioria como 6. A maioria fuzzy, por outro lado, é um conceito mais flexível, manipulado através da lógica fuzzy.

Nesta seção serão apresentados os quantificadores fuzzy, usados para representar o conceito de maioria fuzzy, e o *ordered weighted averaging* (OWA) *operator*, usado para agregar informação. O operador OWA reflete a maioria fuzzy calculando seus pesos através de quantificadores fuzzy.

4.2.1 Quantificadores Fuzzy Lingüísticos

Quantificadores podem ser usados para representar a quantidade de itens que satisfazem um dado predicado. A lógica clássica é restrita ao uso de dois quantificadores, *existe* e *para todo*, que estão relacionados, respectivamente, com os conectivos *ou* e *e*. O discurso humano é muito mais rico e diverso nos seus quantificadores, por exemplo, *por volta de 5, quase todos, alguns, muitos, a maioria, o maior número possível, quase metade, pelo menos a metade* são exemplos de quantificadores familiares ao discurso humano. Na tentativa de acabar com a lacuna entre os sistemas formais e o discurso humano, proporcionando uma forma de representação do conhecimento mais flexível, foi introduzido o conceito de quantificadores lingüísticos.

A semântica dos quantificadores lingüísticos pode ser capturada utilizando subconjuntos fuzzy para sua representação. Existem dois tipos de quantificadores lingüísticos, o *absoluto* e o *proporcional ou relativo*. Os quantificadores absolutos são usados para representar quantidades que são naturalmente

absolutas, tais como *por volta de 2* ou *mais de 5*. Estes quantificadores lingüísticos absolutos estão fortemente relacionados ao conceito de quantidade de elementos. Eles são definidos como subconjuntos fuzzy dos números reais não negativos, $\mathbb{R}^+$. Desta forma, um quantificador absoluto pode ser representado por um subconjunto fuzzy Q, tal que para todo $r \in \mathbb{R}^+$ o grau de pertinência de r em Q, Q(r), indica o grau que a quantidade r é compatível com o quantificador representado por Q. Quantificadores proporcionais tais como *a maioria*, *pelo menos a metade*, podem ser representados por subconjuntos fuzzy do intervalo unitário, $[0, 1]$. Para todo $r \in [0, 1]$, Q(r) indica o grau que a proporção r é compatível com o significado que este quantificador representa. Qualquer quantificador da linguagem natural pode ser representado como um quantificador proporcional ou dada a cardinalidade dos elementos considerados, como um quantificador absoluto. Funcionalmente, os quantificadores lingüísticos são normalmente do tipo *crescente*, *decrescente*, ou *unimodal*. Um quantificador crescente é caracterizado pela relação

$$\mathrm{Q}(r_1) \geq \mathrm{Q}(r_2) \quad \text{se } r_1 > r_2.$$

Um quantificador decrescente é caracterizado pela relação

$$\mathrm{Q}(r_1) \leq \mathrm{Q}(r_2) \quad \text{se } r_1 > r_2.$$

Quantificadores unimodais apresentam a propriedade

$$\mathrm{Q}(a) \leq \mathrm{Q}(b) \leq \mathrm{Q}(c) = 1 \geq \mathrm{Q}(d)$$

para algum $a \leq b \leq c \leq d$.

Um quantificador absoluto $\mathrm{Q} : \mathbb{R}^+ \to [0, 1]$ satisfaz à propriedade:

$$\mathrm{Q}(0) = 0, \text{ e } \exists k \text{ tal que } \mathrm{Q}(k) = 1.$$

Um quantificador relativo, $\mathrm{Q} : [0, 1] \to [0, 1]$, satisfaz à propriedade:

$$\mathrm{Q}(0) = 0, \text{ e } \exists r \in [0, 1] \text{ tal que } \mathrm{Q}(r) = 1.$$

Um quantificador não decrescente satisfaz à propriedade:

$$\forall a, b \text{ se } a > b \text{ então } \mathrm{Q}(a) \geq \mathrm{Q}(b).$$

A função de pertinência de um quantificador relativo não decrescente pode ser representada como

$$\mathrm{Q}(r) = \begin{cases} 0 & \text{se } r < a \\ \frac{r-a}{b-a} & \text{se } a \leq r \leq b \\ 1 & \text{se } r > b \end{cases} \tag{4.1}$$

com $a, b, r \in [0, 1]$.

Neste trabalho foram usados os quantificadores relativos não decrescentes "a maioria", "pelo menos a metade" e "o maior número possível", cujos valores (a, b) são (0,3, 0,8), (0, 0,5) e (0,5, 1), respectivamente (Figura 4.1).

4.2.2 O operador OWA

O operador OWA proporciona uma família de operadores de agregação com o operador *e* em um extremo e o operador *ou* no outro extremo.

Um operador OWA de dimensão n é uma função ϕ,

$$\phi : [0, 1]^n \to [0, 1],$$

que tem um conjunto de pesos associados. Seja $\{a_1, \ldots, a_n\}$ uma lista de valores a agregar, então o operador OWA ϕ é definido como

$$\phi(a_1, \ldots, a_n) = W \cdot B^T = \sum_{i=1}^{n} w_i b_i \tag{4.2}$$

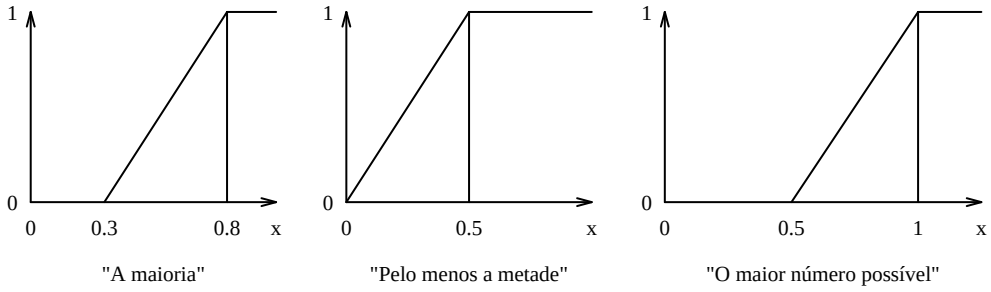

Figura 4.1: Quantificadores lingüísticos relativos fuzzy

onde $W = [w_1, \ldots, w_n]$ é um vetor de pesos, tal que $w_i \in [0, 1]$ e $\sum_i w_i = 1$; e B é o vetor de valores ordenados. Cada elemento $b_i \in B$ é o i-ésimo maior valor na coleção $a_1, \ldots, a_n$ (ordem decrescente).

Conforme mencionado, o operador OWA tem o máximo (ou) em um extremo, o mínimo (e) em outro e outros valores intermediários (como a média) podem ser obtidos, com a escolha apropriada de pesos:

- Para $W = [1, 0, \ldots, 0]$, $\phi(a_1, \ldots, a_n) = \max_i a_i$ (Ou)
- Para $W = [0, 0, \ldots, 1]$, $\phi(a_1, \ldots, a_n) = \min_i a_i$ (E)
- Para $W = [1/m, 1/m, \ldots, 1/m]$, $\phi(a_1, \ldots, a_n) = \text{avg}(a_1, \ldots, a_n)$ $(Média)$

Uma questão natural na definição do operador OWA é como obter o vetor de pesos associados. A primeira abordagem é tentar aprender os pesos através de algum mecanismo de aprendizagem usando exemplos; a segunda é tentar estabelecer uma semântica aos pesos. Esta possibilidade encontrou múltiplas aplicações na área da lógica fuzzy e multivalorada, teoria da evidência, projeto de controladores fuzzy e agregações guiadas por quantificadores.

O interesse deste método é na área de agregações guiadas por quantificadores. A idéia é calcular os pesos das operações de agregação (feitas pelo operador OWA) usando quantificadores lingüísticos que representam o conceito de maioria fuzzy (dando assim um significado humano à agregação). Neste caso, uma maneira usada com sucesso para calcular os pesos do operador OWA baseado em quantificadores lingüísticos, é dada pela fórmula:

$$w_i = \text{Q}(i/n) - \text{Q}((i-1)/n),\ i = 1, \ldots, n. \tag{4.3}$$

Quando um quantificador lingüístico Q é usado para computar os pesos do operador OWA ϕ, este é simbolizado por ϕ_{Q}.

4.3 Processo de Decisão: Método de Classificação das Alternativas

Assume-se que existe um conjunto finito de alternativas $X = \{x_1, \ldots, x_n\}$ bem como um conjunto finito de especialistas $E = \{e_1, \ldots, e_n\}$. Cada especialista $e_k \in E$ fornece sua opinião sobre X na forma de uma ordem individual de preferências $\{x_{o(1)}, \ldots, x_{o(n)}\}$, onde $o(\cdot)$ é a função de permutação sobre o conjunto de índices $\{1, \ldots, n\}$. Cada especialista classifica as alternativas de acordo com uma ordem fraca da melhor para a pior alternativa.

A seguir, é apresentado o processo de agregação a partir do qual uma relação de ordem coletiva é obtida, e então é definido o processo de escolha.

4.3.1 Agregação: a relação de ordem coletiva das preferências

Para cada ordem de preferência é derivada uma relação de preferências, P^k, onde p^k_{ij} reflete a preferência entre as alternativas x_i e x_j para o especialista e_k, $p^k_{ij} \in \{0, 1\}$. Ela assume o valor 1 se x_i é preferido e

0 caso contrário. Desta forma, temos um conjunto de relações de preferência binárias:

$$\{P^1, \ldots, P^n\}.$$

A partir do conjunto de relações de preferência de ordem será obtida a relação de preferência de ordem coletiva, P.

Duas possibilidades serão consideradas quando à intensidade das opiniões dos especialistas, estes podem ser considerados todos iguais (caso homogêneo) ou de importâncias diferentes (caso heterogêneo). O segundo caso é uma generalização do primeiro.

Cada valor $p_{ij} \in [0, 1]$ de P representará o grau que a afirmativa "a alternativa x_i é pelo menos tão boa quanto a alternativa x_j" é verdadeira.

Agregação com Especialistas Homogêneos

Neste caso todas as opiniões dos especialistas são consideradas com a mesma intensidade. Então as relações de preferências são agregadas para obter p_{ij} a partir de $\{p_{ij}^1, \ldots, p_{ij}^n\}$ para todo i, j. Isto é feito usando a maioria fuzzy. Através do quantificador fuzzy lingüístico escolhido para esta fase, é calculado o vetor de pesos do operador OWA (de acordo com a Equação 4.3). O operador OWA é então usado para obter a relação de preferência de ordem coletiva P como

$$P = \phi_{\mathrm{Q}}(P^1, \ldots, P^n)$$

onde $p_{ij} = \phi_{\mathrm{Q}}(p_{ij}^1, \ldots, p_{ij}^n)$, e a agregação realizada de acordo com a Equação 4.2.

Agregação com Especialistas Heterogêneos

Neste caso, associado aos especialistas temos os seus respectivos *graus de importância* como um subconjunto fuzzy, tal que, $\mu_E(k) \in [0, 1]$ denota o grau de importância do especialista e_k.

Assumindo que neste contexto cada valor $\mu_E(k)$ é um peso indicando a importância do especialista no processo de agregação, o procedimento geral para incluir a importância na agregação envolve a transformação dos valores de preferência sob os graus de importância. Esta transformação segue a seguinte expressão:

$$p_{ij}^{k'} = g(p_{ij}^k, \mu_E(k)).$$

Como operação padrão para g pode-se usar o *Mínimo*, que é a implementação padrão para a intersecção de conjuntos fuzzy. Portanto obtemos,

$$p_{ij}^{k'} = \min(p_{ij}^k, \mu_E(k)).$$

Quando todos os especialistas são igualmente importantes ($\mu_E(k) = 1$ para todo $k \in \{1, \ldots, n\}$), $p_{ij}^{k'}$ é reduzido a p_{ij}^k.

4.3.2 Exploração: o Processo de Escolha

Serão definidos dois graus de escolha para as alternativas, um grau de dominância e um grau de não-dominância. Estes graus de escolha atuarão sobre a relação de preferência de ordem coletiva. A sua aplicação poderá ocorrer de acordo com dois diferentes processos de seleção que serão definidos, um *processo seqüencial de seleção* e um *processo de seleção conjuntivo*.

Graus de escolha de alternativas

Serão apresentados dois graus de escolha para as alternativas, baseados no conceito de maioria fuzzy: um grau de dominância e um grau de não dominância. Ambos os graus são baseados no uso do operador OWA cujos pesos são calculados de acordo com o quantificador escolhido para representar a maioria fuzzy nesta fase.

- **Grau de dominância guiado por quantificador**

  O grau de dominância guiado por quantificador (*quantifier guided dominance degree* – QGDD(·)) é usado para quantificar o grau de dominância que uma alternativa tem sobre todas as outras do ponto de vista da maioria fuzzy. O QGDD atua sobre o conjunto de alternativas como:

  $$\mathrm{QGDD}(x_i) = \phi_{\mathrm{Q}}(p_{ij}, j = 1, \ldots, n, j \neq i) \tag{4.4}$$

  onde ϕ_{Q} é um operador OWA cujos pesos são definidos usando o quantificador relativo Q, e cujos componentes (a agregar) são os elementos da linha correspondente de P, isto é, para x_i o conjunto de $n-1$ valores $\{p_{ij} | j = 1, \ldots, n \text{ e } i \neq j\}$.

  Os elementos do conjunto

  $$X^{\mathrm{QGDD}} = \{x | x \in X, \mathrm{QGDD}(x) \geq \mathrm{QGDD}(z), \text{ para todo } z \in X\}$$

  são chamados de elementos de dominância maximal da maioria fuzzy de X quantificada por Q.

- **Grau de não-dominância guiado por quantificador**

  O grau de não-dominância guiado por quantificador (*quantifier guided non-dominance degree* – QGNDD(·)) exprime em que grau uma alternativa não é dominada por uma maioria fuzzy das outras alternativas. Ele é definido pela seguinte expressão:

  $$\mathrm{QGNDD}(x_i) = \phi_{\mathrm{Q}}(1 - p^s_{ji}, j = 1, \ldots, n, j \neq i) \tag{4.5}$$

  onde

  $$p^s_{ji} = \max\{p_{ji} - p_{ij}, 0\}$$

  representa o grau em que x_i é estritamente dominado por x_j.

  Os elementos do conjunto

  $$X^{\mathrm{QGNDD}} = \{x | x \in X, \mathrm{QGNDD}(x) \geq \mathrm{QGNDD}(z), \text{ para todo } z \in X\}$$

  são chamados de elementos maximais não-dominados pela maioria fuzzy de X quantificada por Q.

Processo de seleção

A aplicação dos graus de escolha de alternativas pode ser realizada de duas formas:

- **Política seqüencial:** escolher um dos dois graus de escolha, aplicando-o para obter o conjunto de elementos maximais. Se houver mais de um elemento no conjunto, o segundo grau de escolha pode ser aplicado sobre os elementos desse conjunto.
- **Política conjuntiva:** aplicar os dois graus de escolha, obtendo os conjuntos X^{QGDD} e X^{QGNDD}. A seleção final é a intersecção desses dois conjuntos.

A Figura 4.2 representa graficamente todo o processo de escolha de alternativas baseado em maioria fuzzy.

4.4 Exemplo: obtendo sugestões para um grupo através do método fuzzy

Suponha que temos um grupo de 6 pessoas ($n = 6$) e que os itens a serem recomendados sejam $X = \{x_1, x_2, x_3, x_4\}$. Vamos considerar que todas as pessoas têm a mesma importância (homogêneo).

O primeiro passo para a recomendação é utilizar a filtragem colaborativa para cada pessoa a fim de prever as suas notas para cada um dos itens. Uma maneira de se fazer isso é como na Seção 2.3.1.

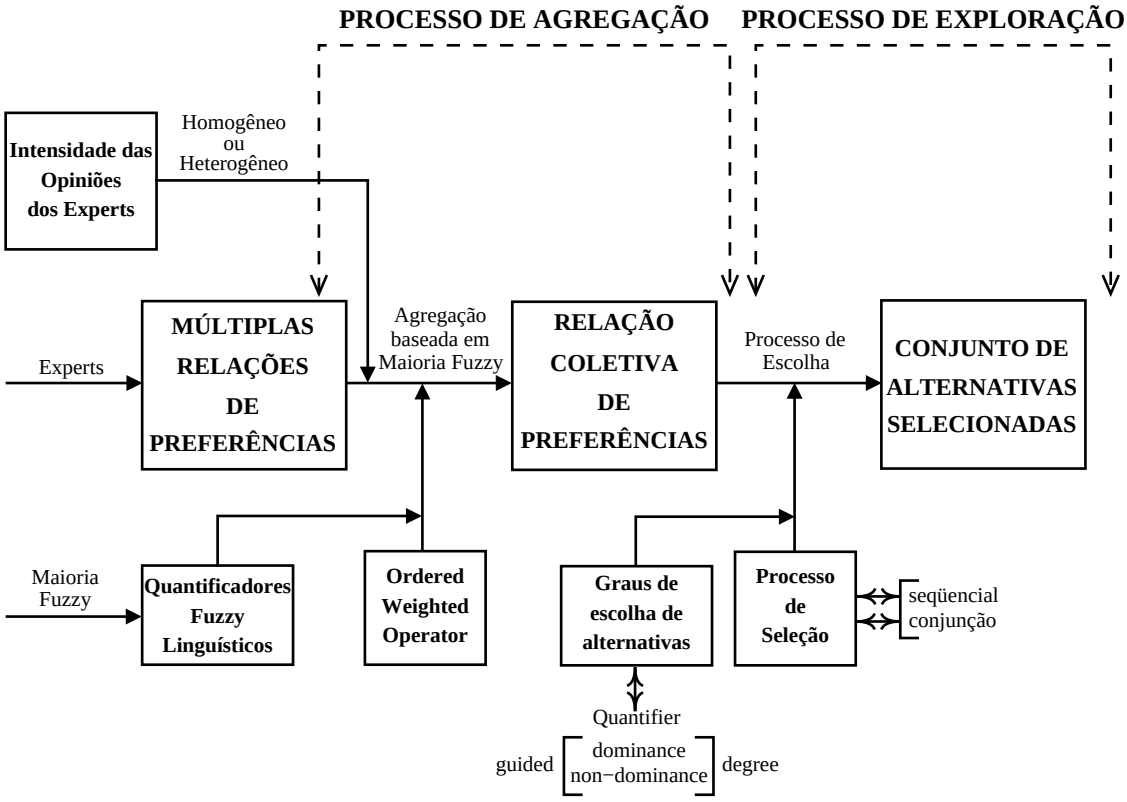

Figura 4.2: Processo de classificação de alternativas baseado em maioria fuzzy

Admita que após usar a filtragem colaborativa, obteve-se as seguintes notas:

$$\begin{aligned} N^1 &= (4.7, 3.8, 3.5, 2.1) \\ N^2 &= (4.8, 4.0, 1.1, 3.2) \\ N^3 &= (2.2, 2.5, 1.8, 1.4) \\ N^4 &= (3.2, 4.8, 2.3, 3.0) \\ N^5 &= (2.4, 2.0, 4.6, 1.5) \\ N^6 &= (3.0, 2.2, 4.1, 2.4) \end{aligned}$$

onde N^k indica as notas previstas para o usuário k. A primeira nota se refere ao item x_1, a segunda ao item x_2 e assim por diante.

Agora vamos transformar as notas em ordens de preferência, para cada um dos usuários:

$$\begin{aligned} O^1 &= (x_1, x_2, x_3, x_4) \\ O^2 &= (x_1, x_2, x_4, x_3) \\ O^3 &= (x_2, x_1, x_3, x_4) \\ O^4 &= (x_2, x_1, x_4, x_3) \\ O^5 &= (x_3, x_1, x_2, x_4) \\ O^6 &= (x_3, x_1, x_4, x_2) \end{aligned}$$

Inicia-se então a fase de agregação. Primeiramente será obtido o conjunto de relações de preferências individuais, $\{P^1, \ldots, P^n\}$:

P^1	x_1	x_2	x_3	x_4
x_1	-	1	1	1
x_2	0	-	1	1
x_3	0	0	-	1
x_4	0	0	0	-

P^2	x_1	x_2	x_3	x_4
x_1	-	1	1	1
x_2	0	-	1	1
x_3	0	0	-	0
x_4	0	0	1	-

P^3	x_1	x_2	x_3	x_4
x_1	-	0	1	1
x_2	1	-	1	1
x_3	0	0	-	1
x_4	0	0	0	-

P^4	x_1	x_2	x_3	x_4
x_1	-	0	1	1
x_2	1	-	1	1
x_3	0	0	-	0
x_4	0	0	1	-

P^5	x_1	x_2	x_3	x_4
x_1	-	1	0	1
x_2	0	-	0	1
x_3	1	1	-	1
x_4	0	0	0	-

P^6	x_1	x_2	x_3	x_4
x_1	-	1	0	1
x_2	0	-	0	0
x_3	1	1	-	1
x_4	0	1	0	-

Para obter a relação de preferência de ordem coletiva, P, escolheremos usar o quantificador lingüístico "o maior número possível", que possui $a = 0,5$ e $b = 1$ (ver Figura 4.1). Primeiramente temos que calcular os pesos para o operador OWA baseado no quantificador escolhido (Equações 4.3 e 4.1):

$$\begin{aligned} w_1 &= \mathrm{Q}(1/6) - \mathrm{Q}(0) = 0 - 0 = 0 \\ w_2 &= \mathrm{Q}(2/6) - \mathrm{Q}(1/6) = 0 - 0 = 0 \\ w_3 &= \mathrm{Q}(3/6) - \mathrm{Q}(2/6) = 0 - 0 = 0 \\ w_4 &= \mathrm{Q}(4/6) - \mathrm{Q}(3/6) = 0.33 - 0 = 0.33 \\ w_5 &= \mathrm{Q}(5/6) - \mathrm{Q}(4/6) = 0.67 - 0.33 = 0.34\ ^1 \\ w_6 &= \mathrm{Q}(6/6) - \mathrm{Q}(5/6) = 1 - 0.67 = 0.33 \end{aligned}$$

[1] Note que se fosse utilizada a precisão completa para os cálculos, os pesos w_4, w_5 e w_6 teriam valores iguais a $0.33\overline{3}$.

O operador OWA é então usado para calcular cada p_{ij} de P:

$$\begin{aligned}
p_{12} &= [0,0,0,0.33,0.34,0.33] \times [1,1,1,1,0,0]^T = 0.33 \\
p_{13} &= [0,0,0,0.33,0.34,0.33] \times [1,1,1,1,0,0]^T = 0.33 \\
p_{14} &= [0,0,0,0.33,0.34,0.33] \times [1,1,1,1,1,1]^T = 1.00 \\
p_{21} &= [0,0,0,0.33,0.34,0.33] \times [1,1,0,0,0,0]^T = 0.00 \\
p_{23} &= [0,0,0,0.33,0.34,0.33] \times [1,1,1,1,0,0]^T = 0.33 \\
p_{24} &= [0,0,0,0.33,0.34,0.33] \times [1,1,1,1,1,0]^T = 0.67 \\
p_{31} &= [0,0,0,0.33,0.34,0.33] \times [1,1,0,0,0,0]^T = 0.00 \\
p_{32} &= [0,0,0,0.33,0.34,0.33] \times [1,1,0,0,0,0]^T = 0.00 \\
p_{34} &= [0,0,0,0.33,0.34,0.33] \times [1,1,1,1,0,0]^T = 0.33 \\
p_{41} &= [0,0,0,0.33,0.34,0.33] \times [0,0,0,0,0,0]^T = 0.00 \\
p_{42} &= [0,0,0,0.33,0.34,0.33] \times [1,0,0,0,0,0]^T = 0.00 \\
p_{43} &= [0,0,0,0.33,0.34,0.33] \times [1,1,0,0,0,0]^T = 0.00
\end{aligned}$$

Representando P na forma de matriz, temos:

P	x_1	x_2	x_3	x_4
x_1	-	0.33	0.33	1.00
x_2	0.00	-	0.33	0.67
x_3	0.00	0.00	-	0.33
x_4	0.00	0.00	0.00	-

Inicia-se agora a fase de exploração. Para cada item, calcularemos o QGDD e o QGNDD. Para o cálculo de ambos os graus, o quantificador lingüístico usado será o "a maioria", que possui $a = 0.3$ e $b = 0.8$ (ver Figura 4.1). Note que nesta fase, o operador OWA será usado para agregar 3 valores (de um item em relação aos outros), e não 6 (número de pessoas, na fase de agregação). Calculando os pesos para o operador OWA baseado no quantificador escolhido (Equações 4.3 e 4.1):

$$\begin{aligned}
w_1 &= \mathrm{Q}(1/3) - \mathrm{Q}(0) = 0.07 - 0 = 0.07 \\
w_2 &= \mathrm{Q}(2/3) - \mathrm{Q}(1/3) = 0.73 - 0.07 = 0.66 \\
w_3 &= \mathrm{Q}(3/3) - \mathrm{Q}(2/3) = 1 - 0.73 = 0.27
\end{aligned}$$

Calculando o QGDD dos itens (Equação 4.4):

$$\begin{aligned}
\mathrm{QGDD}(x_1) &= [0.07,0.66,0.27] \times [1.00,0.33,0.33]^T = 0.38 \\
\mathrm{QGDD}(x_2) &= [0.07,0.66,0.27] \times [0.67,0.33,0.00]^T = 0.26 \\
\mathrm{QGDD}(x_3) &= [0.07,0.66,0.27] \times [0.33,0.00,0.00]^T = 0.02 \\
\mathrm{QGDD}(x_4) &= [0.07,0.66,0.27] \times [0.00,0.00,0.00]^T = 0.00
\end{aligned}$$

Calculando o QGNDD (Equação 4.5):

- x_1:

  $p^s_{21} = \max(0.00 - 0.33, 0) = 0.00$, $p^s_{31} = \max(0.00 - 0.33, 0) = 0.00$, $p^s_{41} \max(0.00 - 1.00, 0) = 0.00$.
  Valores a agregar: $\{1.00, 1.00, 1.00\}$.

  $\mathrm{QGNDD}(x_1) = [0.07,0.66,0.27] \times [1.00,1.00,1.00]^T = 1.00$.

- x_2:

  $p^s_{12} = \max(0.33 - 0.00, 0) = 0.33$, $p^s_{32} = \max(0.00 - 0.33, 0) = 0.00$, $p^s_{42} = \max(0.00 - 0.67, 0) = 0.00$.
  Valores a agregar: $\{0.77, 1.00, 1.00\}$.

  $\mathrm{QGNDD}(x_2) = [0.07,0.66,0.27] \times [1.00,1.00,0.77]^T = 0.94$.

- x_3:

  $p^s_{13} = \max(0.33 - 0.00, 0) = 0.33$, $p^s_{23} = \max(0.33 - 0.00, 0) = 0.33$, $p^s_{43} = \max(0.00 - 0.33, 0) = 0.00$.
  Valores a agregar: $\{0.77, 0.77, 1.00\}$.

  $\text{QGNDD}(x_3) = [0.07, 0.66, 0.27] \times [1.00, 0.77, 0.77]^T = 0.79$.

- x_4:

  $p^s_{14} = \max(1.00 - 0.00, 0) = 1.00$, $p^s_{24} = \max(0.67 - 0.00, 0) = 0.67$, $p^s_{34} = \max(0.33 - 0.00, 0) = 0.33$.
  Valores a agregar: $\{0.00, 0.33, 0.67\}$.

  $\text{QGNDD}(x_4) = [0.07, 0.66, 0.27] \times [0.67, 0.33, 0.00]^T = 0.26$.

Sumarizando os resultados obtidos:

	x_1	x_2	x_3	x_4
QGDD	0.38	0.26	0.02	0.00
QGNDD	1.00	0.94	0.79	0.26

Estes valores representam, respectivamente, o grau de dominância que uma alternativa tem sobre "a maioria" (quantificador usado na comparação das alternativas) das outras de acordo com "o maior número possível" (quantificador usado para agregar as preferências individuais) de pessoas do grupo; e o grau de não-dominância que uma alternativa não é dominada por "a maioria" das outras de acordo com "o maior número possível" de pessoas.

Os conjuntos maximais são claramente:

$$X^{\text{QGDD}} = \{x_1\} \text{ e } X^{\text{QGNDD}} = \{x_1\},$$

portanto para ambos os processos de seleção do método fuzzy, x_1 seria a alternativa escolhida.

No entanto, quando estamos realizando uma recomendação, não precisamos ser tão rígidos ao ponto de oferecer como sugestão apenas as alternativas maximais. Muitas vezes o interesse é de *rankear* as alternativas, ou escolher as m melhores, onde m é o número de sugestões que serão fornecidas. Nesse contexto podemos selecionar um dos graus de escolha (QGDD ou QGNDD) para ordenar as alternativas, e então sugerir as m de melhor grau. O outro grau de escolha pode ser usado para "desempate", por exemplo.

Assim, se desejássemos ordenar as alternativas $\{x_1, x_2, x_3, x_4\}$ poderíamos usar o QGDD para isso, e o QGNDD como desempate. No exemplo fornecido não há empates, e a ordem obtida se usado o QGDD ou o QGNDD é a mesma: (x_1, x_2, x_3, x_4).

Nos experimentos que realizaremos no próximo capítulo, o QGDD será usado como critério para ordenação das alternativas, enquanto que o QGNDD será usado como desempate.

Capítulo 5

Avaliação Experimental de Estratégias *Fuzzy* na Recomendação para Grupos

5.1 Visão geral

Grupos de pessoas podem se apresentar das mais diversas formas: grupos pequenos, ou grandes; de pessoas extremamente afins, ou com opiniões bastante antagônicas. Deste modo, é importante que o comportamento de uma estratégia de recomendação para grupos seja avaliada quanto a essas variáveis, isto é, quanto ao:

- **tamanho do grupo:** o tamanho do grupo influencia no comportamento da estratégia adotada?
- **grau de homogeneidade:** qual a importância da afinidade entre as pessoas do grupo para a geração de uma recomendação?

Neste capítulo o método de recomendação para grupos baseado em maioria fuzzy descrito no Capítulo 4 será avaliado experimentalmente quanto aos dois fatores supracitados. Diferentes estratégias de recomendação (quantificadores lingüísticos fuzzy aplicados) serão utilizadas.

5.2 O Conjunto de Dados Eachmovie

Os experimentos foram realizados utilizando os dados do *EachMovie*. O *EachMovie* era um serviço de recomendação por filtragem colaborativa que funcionou durante 18 meses (até Setembro de 1997) como parte de um projeto de pesquisa do *Compaq Systems Research Center*[1]. Neste período, 72.916 usuários forneceram 2.811.983 avaliações para 1.628 filmes diferentes. As notas dos usuários foram registradas numa escala numérica de 6 níveis (0.0, 0.2, 0.4, 0.6, 0.8, 1.0). O conjunto de dados está disponível para uso não comercial, podendo ser solicitado à Compaq Computer Corporation [6].

Embora estejam disponíveis avaliações de 72.916 usuários, os experimentos foram restritos àqueles que possuíam um número de avaliações maior ou igual a 150 (2.551 usuários). Esta restrição foi adotada para permitir uma intersecção (dos filmes avaliados) de tamanho razoável entre cada par de usuários, para que se possa formar grupos considerados homogêneos ou heterogêneos com maior confiabilidade.

5.3 Preparação dos dados: a formação dos grupos

Para a realização dos experimentos, era necessária a existência de grupos de usuários com diferentes tamanhos e graus de homogeneidade. Nos dados do *EachMovie* não há nenhuma noção de grupos de

[1] Na época, *Digital Equipment Corporation (DEC) Systems Research Center*.

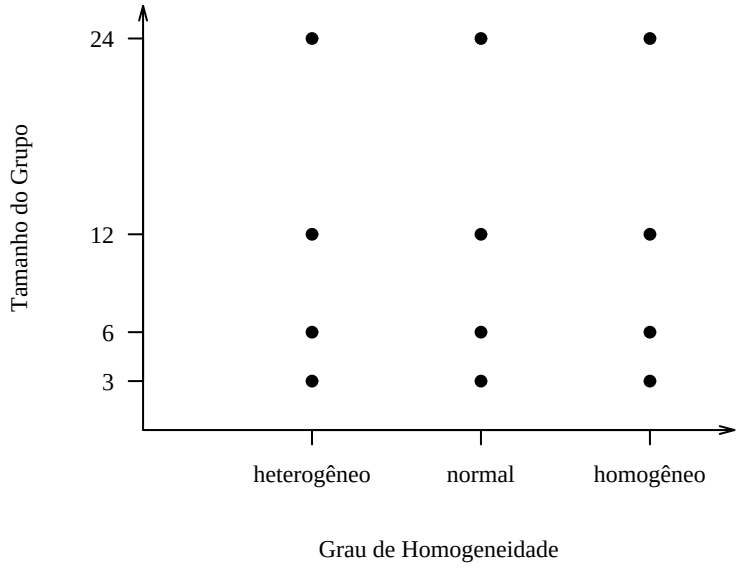

Figura 5.1: Grupos de diferentes tamanhos e graus de homogeneidade. Cada ponto na figura corresponde a 100 grupos formados

pessoas, então é necessário inicialmente que os grupos sejam formados.

Foram definidos 4 níveis para o fator *tamanho do grupo*: 3, 6, 12 e 24 pessoas. Assim estão compreendidos desde grupos bem pequenos (3 pessoas) até grupos relativamente grandes (24 pessoas). Acreditamos que essa faixa de tamanhos abrange a maioria dos cenários nos quais a recomendação para grupos poderá ser usada (Seção 2.4). Já para o fator *grau de homogeneidade* foram utilizados 3 níveis: grupos *homogêneos*, *normais* e *heterogêneos*. Foram gerados 100 grupos para cada combinação "tamanho do grupo" × "grau de homogeneidade". Estas informações estão sumarizadas na Figura 5.1. No nosso contexto, os grupos não precisam ser uma partição do conjunto de usuários, isto é, um mesmo usuário pode pertencer a mais de um grupo diferente[2]. A metodologia utilizada para a formação dos grupos será descrita nas próximas seções.

5.3.1 Obtenção da Matriz de Dissimilaridade

O primeiro passo para a definição dos grupos foi a obtenção de uma matriz de dissimilaridade dos usuários. Isto é, uma matriz m de tamanho $n \times n$ (n é o número de usuários) onde cada m_{ij} contém o valor de dissimilaridade entre os usuários i e j. Para construir essa matriz foi calculada a dissimilaridade de cada usuário em relação a todos os outros. Como a dissimilaridade é simétrica, isto é, $\text{dissim}(x, y) = \text{dissim}(y, x)$, só é necessário calcular uma diagonal da matriz (superior ou inferior), que contém $n(n-1)/2$ valores.

As dissimilaridades entre os usuários serão usadas posteriormente na formação dos grupos com os três graus de homogeneidade desejados. Para obter a dissimilaridade entre dois usuários, os seguintes passos são seguidos:

1. Calcula-se o coeficiente de correlação entre os dois usuários (Equação 2.1). No nosso contexto, o coeficiente de correlação pode ser interpretado como uma medida de similaridade, variando de -1 (menor similaridade) até $+1$ (maior similaridade).

2. O resultado do coeficiente de correlação é transformado num valor de dissimilaridade, que varia

[2]Isto acontece na vida real, por exemplo, a mesma pessoa vai ao cinema com grupos de amigos diferentes.

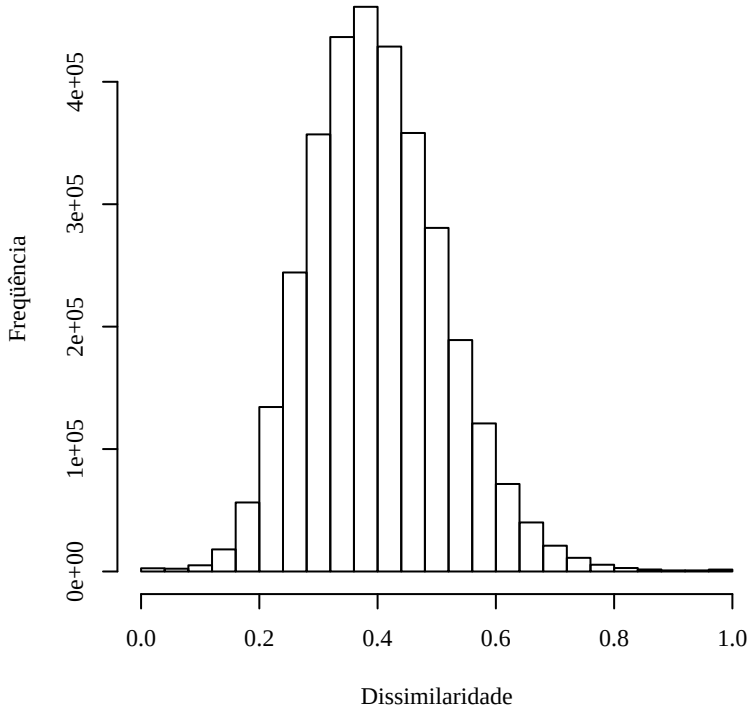

Figura 5.2: Histograma das dissimilaridades entre os usuários

entre 0 (menor dissimilaridade) e 1 (dissimilaridade máxima). Esta transformação é definida como:

$$\text{dissim}(x, y) = 1 - \frac{\rho_{xy} + 1}{2}.$$

Para o experimento, os filmes do *EachMovie* foram separados aleatoriamente em dois conjuntos iguais. Somente as avaliações dos usuários que se referem a elementos do primeiro conjunto (chamado de *conjunto profile*) foram utilizadas para o cálculo das correlações entre os usuários na obtenção da matriz de dissimilaridade. As avaliações que se referem a filmes do outro conjunto (chamado de *conjunto de teste*) não foram usadas neste estágio.

O motivo para esse procedimento, é que os filmes que fazem parte do conjunto de testes serão os itens que poderão ser recomendados para os grupos, assumindo-se então que eles são desconhecidos para o grupo que receberá a recomendação. Desta forma, não deve ser permitido que a avaliação real de membros dos grupos sobre esses filmes do conjunto de teste influencie na determinação do grau de homogeneidade dos grupos que, como veremos a seguir, é dependente da matriz de dissimilaridade dos usuários.

5.3.2 Formação dos Grupos

A formação dos grupos com diferentes graus de homogeneidade demonstrou ser a tarefa mais difícil na preparação do experimento. De posse da matriz de dissimilaridades temos condições de determinar se usuários dois-a-dois são parecidos ou não, mas como estender essa noção para um grupo?

Critério da Dissimilaridade Estrita

Podemos considerar a dissimilaridade entre usuários como uma variável aleatória. Conforme observa-se na Figura 5.2, as dissimilaridades se distribuem de maneira aproximadamente normal. A Tabela 5.1 mostra a média, desvio padrão e o resumo de 5 números de Tukey para a dissimilaridade.

Uma maneira de gerar grupos homogêneos é a seguinte:

Tabela 5.1: Média, desvio padrão e resumo de 5 números de Tukey para a dissimilaridade entre os usuários

	Mínimo	1º quartil	Mediana	Média	Desvio padrão	3º quartil	Máximo
Dissimilaridade	0.0000	0.3193	0.3920	0.3996	0.1150	0.4719	1.0000

Usuário	a	b	c	d
a	-			
b	0.7	-		
c	0.2	0.5	-	
d	0.1	0.1	0.3	-

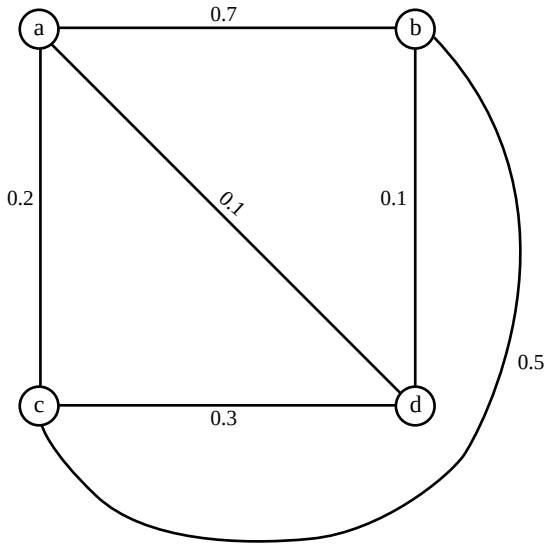

Figura 5.3: Representando uma matriz de dissimilaridades como um grafo completo não direcionado

1. Estabelecer um limiar máximo de dissimilaridade. Por exemplo, poderíamos usar média − desvio padrão.

2. Os grupos homogêneos de tamanho k são formados por k usuários entre os quais todas as dissimilaridades ($k(k-1)/2$) são menores que o limiar estabelecido.

Para encontrar grupos heterogêneos procede-se da mesma forma, só que desta vez é determinado um limiar mínimo de dissimilaridade. Os grupos normais são encontrados limitando-se o valor mínimo e o máximo de dissimilaridade.

À primeira vista, esta estratégia parece simples de implementar. Vamos analisar um pouco mais profundamente a sua complexidade.

O problema que estamos tentando resolver pode ser facilmente expresso como um problema em grafos. Se considerarmos cada usuário como um *vértice* do grafo e a dissimilaridade entre um usuário a e um usuário b como o peso da aresta não direcionada entre a e b, transformamos a matriz de dissimilaridade em um grafo completo[3] não direcionado com pesos nas arestas. A Figura 5.3 mostra como uma matriz de dissimilaridade hipotética entre os usuários {a, b, c, d} pode ser representada como um grafo.

Assim, o problema se resume a encontrar os subgrafos completos (também chamados de *cliques*) de k vértices, de forma que todas as arestas destes subgrafos tenham pesos menores (no caso homogêneo) que um limiar. Dessa forma para resolver o problema podemos em primeiro lugar remover todas as arestas do grafo maiores que o limiar e depois encontrar os cliques de tamanho k. A Figura 5.4 mostra um clique de tamanho 3 (correspondendo a um grupo homogêneo de tamanho 3) encontrado no grafo da Figura 5.3, considerando-se um limiar hipotético de 0.4.

No entanto, o problema "dado um grafo não direcionado e um inteiro k, determinar se o grafo contém um clique de tamanho $\geq k$" (problema *clique*) é reconhecidamente NP-completo [18]. Portanto não existe algoritmo que resolva em tempo polinomial o nosso problema, pois se existisse, poderíamos resolver *clique* em tempo polinomial da seguinte forma:

[3]Um grafo completo é aquele em que cada um de seus vértices está conectado a todos os outros

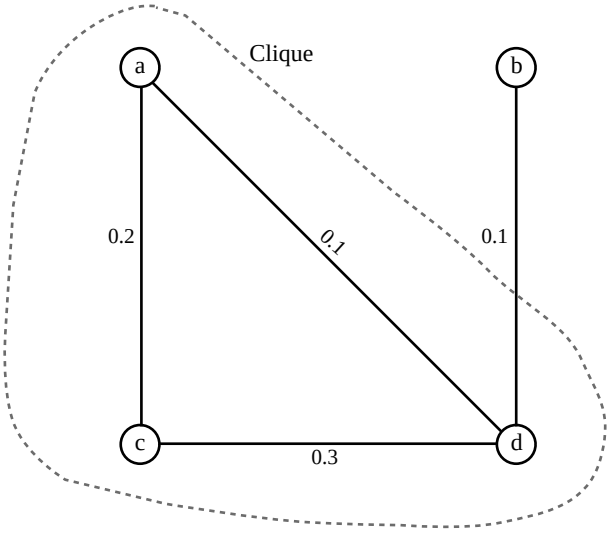

Figura 5.4: Clique de tamanho 3 correspondendo a um grupo homogêneo com limiar = 0.4

- Usar o algoritmo polinomial para encontrar os cliques de tamanho k (algoritmo que resolve nosso problema).
- Caso cliques de tamanho k fossem encontrados, a resposta para o problema *clique* clássico seria afirmativa, pois se existe um ou mais cliques de tamanho k, é verdade que o grafo contém um clique de tamanho $\geq k$.
- Caso não fossem encontrados cliques de tamanho k a resposta para o problema *clique* clássico seria negativa, pois se existissem cliques de tamanho $\geq k$ o algoritmo teria encontrado (pelo menos um) cliques de tamanho $= k$. Isto ocorreria porque os cliques de tamanho $\geq k$ teriam como subgrafos cliques de tamanho $= k$.

Assim, para encontrar grupos usando o critério da dissimilaridade estrita foi desenvolvido um algoritmo com *backtracking*, uma técnica bastante usada para abordar problemas intratáveis [18]. A Figura 5.5 mostra o algoritmo implementado para encontrar os grupos.

Após executar o programa para encontrar os grupos segundo o critério da dissimilaridade estrita, verificou-se que esse critério é muito rigoroso, não sendo possível encontrar grupos de tamanhos maiores, mesmo quando o limiar de dissimilaridade está bem "frouxo". Por exemplo, não foi encontrado nenhum grupo heterogêneo de 24 pessoas mesmo fixando o limiar mínimo de dissimilaridade em média + 0.5 × desvio padrão.

É possível entender intuitivamente porque isso ocorreu. À medida que aumentamos o tamanho do grupo, torna-se cada vez mais difícil que *todos* os membros deste grupo sejam extremamente parecidos (ou extremamente diferentes). Sempre existem alguns membros do grupo mais próximos do que desejamos (quando procuramos grupos heterogêneos) e, analogamente, mais separados do que o desejado na formação dos grupos homogêneos. Deste modo, foram estabelecidos novos critérios para a construção dos grupos, como veremos a seguir.

Critérios Heurísticos de Dissimilaridade

Os algoritmos de *clustering* são usados para organizar um conjunto de objetos em grupos, ou *clusters*, de tal forma que é forte o grau de associação entre os membros do mesmo cluster, e fraco entre os membros de clusters diferentes.

Então, esses algoritmos parecem promissores para encontrar os nossos grupos homogêneos (membros do mesmo cluster), e até mesmo heterogêneos (membros de clusters diferentes). Em vista disso, o seu uso para a formação dos grupos foi investigado.

```
Algoritmo Encontrar_Grupos(tam, min_dissim, max_dissim, ids)
Entrada: tam (tamanho dos grupos desejados),
        min_dissim (limiar de dissimilaridade mínima),
        max_dissim (limiar de dissimilaridade máxima),
        ids (array de usuários).
Saída: grupos (lista de grupos encontrados)

begin
        grupos = lista vazia;
        {grupo_atual é o grupo que está sendo formado pelo procedimento solve}
        grupo_atual = lista vazia;
        { o último parâmetro é a posição do id atual que será considerado por solve }
        solve(tam, min_dissim, max_dissim, grupos, ids, grupo_atual, 0);
end

procedure solve(tam, min_dissim, max_dissim, grupos, ids, grupo_atual, posicao):

begin
        { este bound indica que não adianta continuar, pois não há mais elementos suficientes para formar
        um grupo do tamanho desejado }
        if (ids.size - posicao < tam - grupo_atual.size) then return;
        elemento_atual = ids[posicao];
        if (compatível(elemento_atual, grupo_atual, min_dissim, max_dissim)) then
                grupo_atual.add(elemento_atual);
                if (grupo_atual.size = tam) then
                        grupos.add(grupo_atual);
                else
                        solve(tam, min_dissim, max_dissim, grupos, ids, grupo_atual, posicao + 1);
                grupo_atual.remove(elemento_atual);
        solve(tam, min_dissim, max_dissim, grupos, ids, grupo_atual, posicao + 1);
end

procedure compatível(elemento_atual, grupo_atual, min_dissim, max_dissim):

begin
        for todo elemento em grupo_atual do
                if (dissimilaridade de elemento com elemento_atual < min_dissim ou
                                > max_dissim) then
                        return false;
        return true;
end
```

Figura 5.5: Algoritmo usando *backtracking* para encontrar todos os grupos de tamanho especificado de acordo com o critério da dissimilaridade estrita.

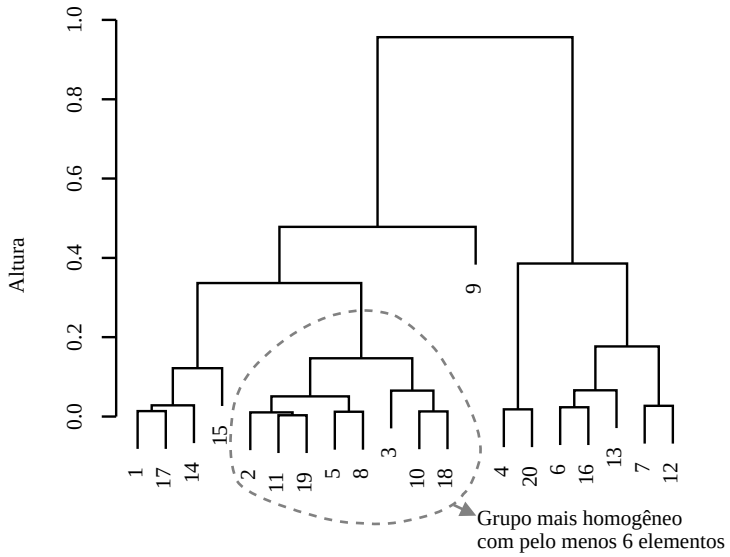

Figura 5.6: Extração de um grupo homogêneo a partir de um dendograma

Geração de Grupos Homogêneos De uma maneira geral, existem dois tipos de métodos de análise de clusters [28]:

- Métodos de particionamento: algoritmos que dividem o conjunto de dados em k clusters, onde o inteiro k precisa ser especificado pelo usuário.
- Métodos hierárquicos: algoritmos que geram uma hierarquia completa de aglomerações do conjunto de dados. Métodos *aglomerativos* iniciam de forma que cada objeto no conjunto de dados forma o seu próprio cluster, e então sucessivamente unem os clusters até que um único grande cluster resta, o conjunto inteiro dos dados (estratégia *bottom up*). Os métodos *divisivos* começam considerando que o conjunto inteiro de dados é um cluster, e então separam os clusters sucessivamente, até que cada objeto é o seu próprio cluster (estratégia *top down*).

Alguns algoritmos de ambos os métodos podem receber como entrada uma matriz de dissimilaridades, que é justamente o tipo de dados de que dispomos.

Para formar os grupos homogêneos, foi considerado mais adequado utilizar um método hierárquico, afinal desejamos obter 100 grupos homogêneos de tamanhos pré-fixados (3, 6, 12, 24), e não grupos de tamanhos os mais variados, o que acontece se solicitarmos 100 grupos a um método de particionamento. Todavia, como temos um conjunto de dados muito grande (2551 usuários) o processo de clustering se torna muito *caro* (tanto em tempo quanto em memória). Ainda mais, é muito complexo "navegar" numa única grande hierarquia gerada, para encontrar 100 grupos distintos, os mais homogêneos possíveis.

Então, seguindo a filosofia "dividir para conquistar", a partir dos 2551 usuários foram selecionados aleatoriamente (com reposição) 100 grupos de 200 usuários cada. Para cada um destes grupos foi executado o algoritmo de clustering *divisive analysis* – diana (Seção A.1), resultando em 100 hierarquias diferentes. De cada uma dessas hierarquias foram extraídos o grupo mais homogêneo de 3, 6, 12 e 24 elementos. Desta forma obteve-se todos os grupos homogêneos desejados.

Para exemplificar como esses grupos foram extraídos, suponha que desejamos formar um grupo de 6 elementos a partir da árvore aglomerativa (dendograma) gerada pelo diana. A Figura 5.6 mostra um dendograma para 20 objetos fictícios (note que para a extração dos grupos reais, cada árvore possui 200 objetos). O grupo dos 6 elementos mais homogêneos está contido no "ramo" da árvore de menor altura que tem pelo menos 6 objetos. Na figura, o ramo com esta propriedade está circulado. No entanto, o ramo escolhido pode possuir um número maior de elementos que o desejado, neste caso, queremos um grupo de 6 elementos, mas o "melhor ramo" possui 8 elementos. Como então escolher os 6 melhores entre eles?

Para isso, foi implementado um algoritmo com o seguinte comportamento:

1. As "junções" dos grupos são percorridas da menor para a maior altura, até encontrar uma junção que forme um ramo com pelo menos os n elementos desejados.

2. Se o ramo encontrado tem exatamente n elementos, estes formam o grupo procurado. Caso contrário, todas as combinações de tamanho n dos elementos do ramo são encontradas e calculada a dissimilaridade média em cada uma delas. O grupo procurado é a combinação de tamanho n com menor dissimilaridade média do ramo selecionado.

   No entanto, para os grupos de tamanho 24, não é mais possível gerar e testar todas as combinações (por exemplo, $\text{comb}(40, 24) > 6.28 \times 10^{10}$). Para esses grupos foi usada a seguinte heurística:

   (a) Para cada elemento do ramo selecionado, calcula-se o somatório das dissimilaridades dele para todos os outros do ramo.

   (b) Os n elementos com menor dissimilaridade total são escolhidos como o grupo procurado.

Geração de Grupos Heterogêneos Inicialmente foi tentada a geração dos grupos heterogêneos utilizando o método de clustering *partitioning around medoids* – pam (Seção A.2). Os passos seguidos foram:

- Para cada um dos 100 grupos de 200 usuários foi solicitado 4 particionamentos diferentes, correspondendo aos tamanhos de grupo desejados. Isto é, foi utilizado $k = 3, 6, 12, 24$.

- A partir de cada particionamento gerado para os grupos de 200 usuários, é possível extrair um grupo heterogêneo de tamanho k, selecionando o elemento mais central de cada um dos clusters da partição.

Por exemplo, para encontrar um grupo heterogêneo de 6 pessoas, o método pam é executado com $k = 6$ para um dos grupos de 200 usuários. Para cada um dos 6 grupos gerados, é selecionado o elemento mais central. O elemento mais central foi definido como aquele com *menor* dissimilaridade total (somatório das dissimilaridades dele para todos os outros do mesmo cluster). Os 6 elementos centrais formam o grupo heterogêneo procurado.

No entanto, este método não funcionou como esperado. Muitas vezes os elementos mais centrais de clusters diferentes eram bastante próximos, e o grupo gerado não era necessariamente heterogêneo (comumente a dissimilaridade média do grupo era próxima ou menor do que a global). Provavelmente os grupos gerados pelo pam eram muito próximos, dada a falta de uma estrutura de grupos bem definida dos dados dos usuários.

Um método mais simples funcionou melhor para a formação dos grupos heterogêneos. Neste método, assim como no caso homogêneo, a partir de cada um dos grupos de 200 usuários criados aleatoriamente é extraído um grupo de 3, 6, 12 e 24 elementos. Para encontrar os elementos heterogêneos, é calculada a dissimilaridade total de cada um dos usuários com os outros 199. Os elementos que formarão o grupo heterogêneo de tamanho k são aqueles com as k *maiores* dissimilaridades totais.

Geração de Grupos Normais Os grupos normais são aqueles em que o comportamento da dissimilaridade dentro do grupo é semelhante ao da população em geral.

Para obter um grupo normal de tamanho k, são selecionados aleatoriamente na população do experimento (2551 pessoas) k usuários. Estes formarão um grupo normal. Para evitar supresas da aleatoriedade, é testada se a média da dissimilaridade do grupo obtido realmente não difere estatisticamente da média da população, utilizando o teste de comparação de uma média (no caso a do grupo) com um valor específico (média conhecida da população, expressa na Tabela 5.1), com $\alpha = 0.05$. Este teste estatístico é descrito na Seção B.1 [20].

Desta maneira, todos os grupos normais foram obtidos.

Tabela 5.2: Estratégias fuzzy utilizadas na recomendação para grupos

Estratégia	Quantificador de Agregação	Quantificador de Exploração
1	O maior número possível	O maior número possível
2	O maior número possível	Pelo menos a metade
3	Pelo menos a metade	A maioria
4	A maioria	O maior número possível

5.4 Metodologia Experimental

Para cada um dos 1200 grupos de usuários obtidos (4 tamanhos × 3 graus de homogeneidade × 100 repetições, ver Figura 5.1) foram geradas recomendações utilizando o método fuzzy (Seção 4.4), com diferentes estratégias. A recomendação foi realizada em relação a 50 filmes do conjunto de testes escolhidos para cada grupo. A filtragem colaborativa individual foi realizada como na Seção 2.3.1, limitando-se o número de vizinhos aos 40 mais próximos[4]. A Tabela 5.2 mostra as estratégias fuzzy utilizadas neste experimento.

Para avaliar o comportamento das estratégias fuzzy em relação aos diferentes tamanhos de grupo e graus de homogeneidade, é necessário a utilização de uma métrica que possa representar este comportamento. Como vimos na Seção 4.4, a partir de um conjunto de rankings que representa as ordens de preferências individuais dos membros de um grupo, obtemos com o método fuzzy um ranking que representa a ordem preferência do grupo, de acordo com os critérios de maioria fuzzy adotados. Desta forma, uma métrica adequada deve ser capaz de *medir o grau de relação* entre rankings individuais da entrada e o ranking final.

A métrica adotada no experimento foi o coeficiente de correlação de rankings de Kendall [12], representado pela letra grega τ (tau). Dados dois rankings, este coeficiente (descrito na Seção B.2) fornece um valor entre -1 e $+1$, que indicam respectivamente discordância máxima (um ranking é o inverso do outro) e concordância máxima (os rankings são idênticos). Valores entre os extremos indicam concordâncias intermediárias.

Em cada recomendação realizada, foi calculado τ entre o ranking final gerado para o grupo e os rankings individuais dos usuários (calculados pelo processo de filtragem colaborativa). Então foi calculado o $\overline{\tau}$ (tau médio). Mais formalmente:

- seja $\{O^1, O^2, \ldots, O^n\}$ o conjunto de preferências individuais (rankings) dos n membros do grupo, e O o ranking final gerado. $\overline{\tau}$ é definido como,

$$\overline{\tau} = \frac{1}{n} \sum_{i=1}^{n} \tau(O^i, O).$$

O experimento teve o objetivo de avaliar se $\overline{\tau}$ é afetado pela variação do tamanho e grau de homogeneidade do grupo. Isto é, verificar se houve diferença significativa nos $\overline{\tau}$ para os diversos níveis de tamanho ou grau de homogeneidade dos grupos. O procedimento estatístico apropriado para testar a igualdade de diversas médias é a *análise de variância* [20].

5.4.1 Influência da Variação do Grau de Homogeneidade

Para determinar a influência da variação do grau de homogeneidade no valor de $\overline{\tau}$ foi realizada uma análise de variância de um único fator (*one-way anova*), fixando-se a estratégia e o tamanho do grupo. Assim foram realizadas 16 análises de variância (4 estratégias × 4 tamanhos de grupo). A Tabela 5.3 mostra a organização dos dados de entrada para uma análise de variância deste tipo.

[4]Segundo dados experimentais em [8] a qualidade da recomendação cai sensivelmente quando se usa vizinhanças de tamanho maior.

Tabela 5.3: Dados usados numa análise de variância para determinar a influência do grau de homogeneidade, com estratégia e tamanho do grupo fixos

	Grupos (repetições)			
Níveis	1	2	$\cdots$	100
homogêneo	$\overline{\tau}_{1,hom}$	$\overline{\tau}_{2,hom}$	$\cdots$	$\overline{\tau}_{100,hom}$
normal	$\overline{\tau}_{1,norm}$	$\overline{\tau}_{2,norm}$	$\cdots$	$\overline{\tau}_{100,norm}$
heterogêneo	$\overline{\tau}_{1,het}$	$\overline{\tau}_{2,het}$	$\cdots$	$\overline{\tau}_{100,het}$

Tabela 5.4: Dados usados numa análise de variância para determinar a influência do tamanho do grupo, com estratégia e grau de homogeneidade fixos

	Grupos (repetições)			
Níveis	1	2	$\cdots$	100
3	$\overline{\tau}_{1,3}$	$\overline{\tau}_{2,3}$	$\cdots$	$\overline{\tau}_{100,3}$
6	$\overline{\tau}_{1,6}$	$\overline{\tau}_{2,6}$	$\cdots$	$\overline{\tau}_{100,6}$
12	$\overline{\tau}_{1,12}$	$\overline{\tau}_{2,12}$	$\cdots$	$\overline{\tau}_{100,12}$
24	$\overline{\tau}_{1,24}$	$\overline{\tau}_{2,24}$	$\cdots$	$\overline{\tau}_{100,24}$

5.4.2 Influência da Variação do Tamanho do Grupo

Para determinar a influência da variação do tamanho do grupo no valor de $\overline{\tau}$ foi realizada uma análise de variância de um único fator (*one-way anova*), desta vez tendo os tamanhos dos grupos como níveis e fixando-se a estratégia e o grau de homogeneidade. Assim foram realizadas 12 análises de variância (4 estratégias $\times$ 3 graus de homogeneidade). A Tabela 5.4 mostra a organização dos dados de entrada para uma análise de variância deste tipo.

5.5 Resultados do Experimento

As análises de variância realizadas demonstraram que o grau de homogeneidade do grupo é de extrema importância no comportamento das estratégias de recomendação (Tabelas C.1 – C.16). Em todas as estratégias utilizadas e para todos os tamanhos de grupos, foi observada uma diferença extremamente significativa nos valores de $\overline{\tau}$ entre os níveis de homogeneidade. O *valor p*, que indica a probabilidade de se cometer um erro ao se rejeitar a hipótese nula (não existência de diferença significativa) teve sempre valor inferior ao mínimo detectável pelo *software* de análise estatística utilizado[5], 2.2×10^{-16}. Também foram comparadas as médias dos níveis par-a-par, utilizando o método da menor diferença significativa (*least significant difference* – LSD) [20]. O método LSD calcula o menor valor que a diferença entre duas médias deve ter para ser considerado significativo. A diferença entre duas médias $\overline{y}_i$ e $\overline{y}_j$ é significativa (ao nível α usado para o cálculo do LSD) se $|\overline{y}_i - \overline{y}_j| >$ LSD. Nas análises de variância considerando-se a variação do grau de homogeneidade todas as médias dos níveis foram significativamente diferentes (par-a-par), com nível de probabilidade $< 1\%$. E, como esperado, em todos os casos

$$\text{média homogêneo} > \text{média normal} > \text{média heterogêneo.}$$

Outro dado importante observado nestas análises de variância é o valor "R-SQUARE". O R^2 pode ser interpretado (de forma não rigorosa) como a proporção da variabilidade dos dados que pode ser "explicada" pelo modelo de análise de variância utilizado [20]. R^2 pode assumir valores entre 0 e 1, sendo os valores maiores mais desejáveis. O valor de R^2 nas análises de variância para determinação da influência da homogeneidade do grupo foi sempre alto, indicando que o grau de homogeneidade é realmente determinante no comportamento de todas as estratégias utilizadas, para todos os tamanhos de grupo.

[5]Foi utilizado o software para computação estatística "R", disponível livremente em http://www.gnu.org/software/r/R.html

As análises de variância da influência do tamanho do grupo (Tabelas D.1 – D.12) permitiram verificar que este fator também influencia o comportamento da recomendação. No entanto, ao contrário do grau de homogeneidade, o tamanho do grupo *não* se mostrou com o alto grau de influência no valor das médias independentemente das outras variáveis fixadas. As médias só foram todas significativamente diferentes para os grupos homogêneos. Já os grupos normais não apresentaram diferenças estatisticamente significativas entre os grupos de 12 e 24 pessoas. Nos grupos heterogêneos, o valor P foi o menor de todos, chegando a ser maior que 1% (mais ainda significativo a nível de 5%) em duas estratégias. A maioria das médias dos grupos heterogêneos não apresentou diferenças significativas quando testadas par-a-par pelo método LSD. Ainda é importante notar que R^2 alcançou valores bem mais baixos do que nas análises da influência do grau de homogeneidade. Nos grupos heterogêneos, o valor de R^2 foi especialmente baixo (sempre < 0.04) o que indica que o tamanho do grupo é bem menos importante que o grau de homogeneidade na definição da resposta dos métodos de recomendação.

Em ambos os tipos de análise, as estratégias fuzzy utilizadas parecem responder da mesma forma à variação dos fatores tamanho do grupo e grau de homogeneidade. No entanto, não foram realizadas comparações entre as estratégias. Conforme visto no Capítulo 3, não existe um critério absoluto para decidir que "uma estratégia A é melhor que uma estratégia B". A conveniência da aplicação de uma estratégia de decisão para grupos é dependente do tipo de problema que se pretende resolver e das características do grupo envolvido.

Capítulo 6

Conclusões e Trabalhos Futuros

A área de recomendação para grupos é ainda muito inexplorada, porém bastante promissora. Como podemos verificar com este trabalho, trata-se de uma matéria interdisciplinar, envolvendo a matemática, psicologia, economia e ciência da computação. Novas tecnologias que serão adotadas nos próximos anos, como a TV Interativa, deverão gerar uma demanda muito forte para personalização, incluindo a recomendação para grupos.

Neste trabalho foi possível ter uma visão das dificuldades envolvidas na recomendação para grupos, como a inexistência de um método "perfeito" de recomendação. Dada a flexibilidade e "consistência humana" de um método fuzzy utilizado para classificação de alternativas segundo opiniões diversas de especialistas e o sucesso dos sistemas de recomendação baseados em filtragem colaborativa, a combinação dessas duas técnicas foi proposta como um caminho viável para a implementação de um sistema de recomendação para grupos.

Essa abordagem foi implementada e teve o seu comportamento avaliado experimentalmente, o que nos levou a enxergar a extrema importância do grau de homogeneidade do grupo de pessoas no resultado de uma recomendação para um grupo. Também notou-se a influência do tamanho do grupo no comportamento da recomendação, porém em grau menor do que a homogeneidade.

Dada a vastidão da área e escassez de pesquisas em recomendação para grupos muitas oportunidades estão abertas. Podemos enxergar como extensões claras:

- O método fuzzy já permite que as pessoas sejam tratadas de forma heterogênea (fornecer diferentes graus de importância para as opiniões de cada pessoa), uma forma de usar essa característica na geração de recomendações seria determinar dinamicamente a importância da opinião de cada indivíduo, atribuindo pesos automaticamente baseado em esquemas de decisão da psicologia, por exemplo, o esquema de decisão influenciado pela distância que a opinião de cada indivíduo tem da média do grupo (Seção 3.3.1).

- Explicação das sugestões para o grupo, podendo inclusive usar técnicas de inteligência artificial para explicar as sugestões em linguagem natural. Por exemplo: "O filme *A* deve agradar a João e Maria, mas José provavelmente gostará mais do filme *B*". Caso fossem utilizadas técnicas baseadas no conteúdo, a explicação poderia envolver o motivo pelo qual o grupo gostará de determinado item: *Este filme tem Julia Roberts e é parecido com o filme X que vocês gostaram.*

- Identificação automática de "facções". Um sistema de recomendação para grupos poderia sugerir a subdivisão de grupos muito heterogêneos, mas que possuísse "facções" (subgrupos mais homogêneos), a fim de melhorar satisfação das pessoas, *é melhor que vocês 3 assistam ao filme A e vocês 4 ao filme B.*

Como se percebe, muitas oportunidades de pesquisa estão abertas na área de recomendação para grupos, uma área multidisciplinar de grande potencial, para a qual este trabalho espera ser uma pequena contribuição.

Bibliografia

[1] Charu Aggarwal, Joel Wolf, Kun-Lung Wu, and Philip Yu. Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In Surajit Chaudhuri and David Madigan, editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 201–212, N.Y., August 15–18 1999. ACM Press.

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216. ACM Press, 1993.

[3] Kenneth J. Arrow. *Social Choice and Individual Values.* Wiley, New York, 2nd edition, 1963.

[4] Marko Balabanovic and Yoav Shoham. Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[5] Daniel Billsus and Michael J. Pazzani. Learning collaborative information filters. In *Proceedings of the 15th International Conference on Machine Learning*, pages 46–54. Morgan Kaufmann, San Francisco, CA, 1998.

[6] Compaq Systems Research Center. Eachmovie collaborative filtering data set. http://www.research.compaq.com/SRC/eachmovie/, 2001.

[7] F. Chiclana, F. Herrera, E. Herrera-Viedma, and M.C. Poyatos. A classification method of alternatives for multiple preference ordering criteria based on fuzzy majority. *Journal of Fuzzy Mathematics*, 4(4):801–813, December 1996.

[8] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Theoretical Models, pages 230–237, 1999.

[9] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. Recommending and evaluating choices in a virtual community of use. In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, volume 1 of *Papers: Using the Information of Others*, pages 194–201, 1995.

[10] Verlin B. Hinsz. Group decision making with responses of a quantitative nature: The theory of social decision schemes for quantities. *Organizational Behavior and Human Decision Processes*, 80(1):28–49, October 1999.

[11] Tatsuya Kameda and R. Scott Tindale. Past, present, and future directions in group decision-making. In S. L. Schneider and J. Shanteau, editors, *Emerging perspectives on judgment and decision research.* Cambridge University Press, Cambridge, In press.

[12] Maurice Kendall. *Rank Correlation Methods.* Charles Griffin & Company, London, 4th edition, 1975.

[13] John M. Levine. Transforming individuals into groups: Some hallmarks of the SDS approach to small group research. *Organizational Behavior and Human Decision Processes*, 80(1):21–27, October 1999.

[14] Henry Lieberman, Neil W. Van Dyke, and Adrian S. Vivacqua. Let's browse: A collaborative web browsing agent. In *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, Collaborative Filtering and Collaborative Interfaces, pages 65–68, 1999.

[15] Weiyang Lin, Sergio A. Alvarez, and Carolina Ruiz. Collaborative recommendation via adaptive association rule mining. In *Proceedings of the Web Mining for E-Commerce Workshop (WebKDD'2000)*, Boston, MA, August 2000.

[16] Pattie Maes. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, July 1994.

[17] Thomas W. Malone, Kenneth R. Grant, Franklyn A. Turbak, Stephen A. Brobst, and Michael D. Cohen. Intelligent information-sharing systems. *Communications of the ACM*, 30(5):390–402, May 1987.

[18] Udi Manber. *Introduction to Algorithms: A Creative Approach*. Addison-Wesley, 1989.

[19] Susan Mohammed and Erika Ringseis. Cognitive diversity and concensus in group decision making: The role of inputs, processes, and outcomes. *Organizational Behavior and Human Decision Processes*, 85(2):310–335, July 2001.

[20] Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, New York, 4th edition, 1997.

[21] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, Sharing Information and Creating Meaning, pages 175–186, 1994.

[22] Donald G. Saari and Fabrice Valognes. Geometry: Voting, and paradoxes. *Mathematics Magazine*, 78:243–259, October 1998.

[23] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM Conference on Electronic Commerce (EC-00)*, pages 158–167, N.Y., October 17–20 2000. ACM.

[24] Badrul M. Sarwar, Joseph A. Konstan, Al Borchers, Jon Herlocker, Brad Miller, and John Riedl. Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system. In *Proceedings of the ACM 1998 conference on Computer supported cooperative work*, pages 345–354. ACM Press, 1998.

[25] Upendra Shardanand and Patti Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, volume 1 of *Papers: Using the Information of Others*, pages 210–217, 1995.

[26] Barry Smyth and Paul Cotter. A personalized television listings service. *Communications of the ACM*, 43(8):107–111, 2000.

[27] Jim Stroud. TV personalization: A key component of interactive tv. Technical report, The Carmel Group, 2001. Available at http://www.carmelgroup.com/tv/TCG-TVPersonalization.pdf.

[28] Anja Struyf, Mia Hubert, and Peter J. Rousseeuw. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1, 1996. http://www.jstatsoft.org/.

[29] Loren Terveen and Will Hill. Human-computer collaboration in recommended systems. In John M. Carroll, editor, *Human-Computer Interaction in the New Millennium*, chapter 22. Addison Wesley, 2002.

Apêndice A

Algoritmos de Clustering

A.1 Algoritmo *Divisive Analysis* – `diana`

O algoritmo `diana` é um método de clustering hierárquico divisivo. O cluster inicial (no passo 0) consiste de todos os n objetos. Em cada passo subseqüente, o maior cluster disponível é dividido em dois clusters menores, até que finalmente todos os clusters contém apenas um objeto.

O algoritmo completo consiste de $n-1$ divisões. Em cada passo, é selecionado o cluster C com o maior diâmetro, onde

$$\text{diam}(C) = \max_{i,j \in C} d(i,j).$$

Assumindo que $\text{diam}(C) > 0$, C é dividido em dois clusters A e B, de acordo com a seguinte estratégia (inicialmente $A = C$ e $B = \emptyset$):

1. Movendo um objeto de A para B.

   Para cada objeto $i \in A$ calcular $a(i)$, a dissimilaridade média para todos os outros objetos de A. O objeto m de A para o qual $a(m)$ é máximo, é movido para B:

   $$A = A \setminus \{m\}, B = m.$$

2. Movendo outros objetos de A para B, que é chamado de *splinter group*.

   Se $|A| = 1$, pare. Caso contrário, calcular $a(i)$ para todo $i \in A$, e a dissimilaridade média de i para todos os objetos de B, simbolizada por $d(i,B)$. Selecionar o objeto $h \in A$ para o qual $a(h) - d(h,B) = \max_{i \in A}(a(i) - d(i,B))$.

   Se $a(h) - d(h,B) > 0 \Rightarrow$ mover h de A para B, e voltar para o passo 2.

   Se $a(h) - d(h,B) \leq 0 \Rightarrow$ pare. A e B ficam como estão.

A.2 Algoritmo *Partitioning Around Medoids* – `pam`

O algoritmo `pam` é um método de clustering de particionamento. Ele se baseia na procura de k *objetos representativos*, chamados *medoids*, entre os objetos do conjunto de dados. Estes medoids são computados de tal maneira que a dissimilaridade de todos os objetos para o medoid mais próximo é mínima, istó é, o objetivo é encontrar um subconjunto $\{m_1, \ldots, m_k\} \subset \{1, \ldots, n\}$ que minimize a função objetivo

$$\sum_{i=1}^{n} \min_{t=1,\ldots,k} d(i, m_t). \tag{A.1}$$

Cada objeto é então alocado ao cluster correspondente ao medoid mais próximo. Isto é, o objeto i é colocado no cluster v_i quando o medoid m_{v_i} está mais próximo de i do que qualquer outro medoid m_w, ou seja

$$d(i, m_{v_i}) \leq d(i, m_w) \text{ para todo } w = 1, \ldots, k.$$

O algoritmo funciona em dois passos:

1. Passo CONSTRUIR

   Construir os medoids iniciais

   - m_1 é o objeto com menor $\sum_{i=1}^{n} d(i, m_1)$
   - m_2 minimiza o objetivo (A.1) o máximo possível

     $\vdots$
   - m_k minimiza o objetivo (A.1) o máximo possível

2. Passo TROCAR

   Repetir até a convergência:

   Considere todos os pares de objetos (i, j) com

   $$i \in \{m_1, \ldots, m_k\} \text{ e } j \notin \{m_1, \ldots, m_k\}$$

   troque os $i \leftrightarrow j$ que diminuírem mais o objetivo (se existir)

Apêndice B

Preliminares Estatísticos

B.1 Comparação de uma Média com um Valor Específico

Em algumas situações desejamos comparar a média μ de uma população com um valor específico, digamos μ_0. As hipóteses são

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Se a população é normal com variância conhecida, ou se a população é não-normal mas o tamanho da amostra em grande o suficiente para que o teorema central do limite se aplique, a hipótese pode ser testada usando uma aplicação direta da distribuição normal. A estatística de teste é

$$Z_0 = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}} \tag{B.1}$$

Se $H_0 : \mu = \mu_0$ é verdadeiro, então a distribuição de Z_0 é $N(0, 1)$. Então, a regra de decisão para $H_0 : \mu = \mu_0$ é rejeitar a hipótese nula se $|Z_0| > Z_{\alpha/2}$. O valor da média μ_0 especificado na hipótese nula é normalmente obtido por evidências passadas, conhecimento teórico, ou experimental.

B.2 O Coeficiente de Correlação de Rankings de Kendall

O coeficiente de correlação de rankings de Kendall, denotado pela letra grega τ (tau), mede o grau de correspondência entre dois rankings. Ele apresenta as seguintes propriedades:

1. se a concordância entre os rankings é perfeita, isto é, todo objeto tem o mesmo rank em ambos, τ assume o valor $+1$, indicando correlação positiva perfeita;

2. se a discordância é perfeita, isto é, um ranking é o inverso do outro, τ assume o valor -1, indicando correlação negativa perfeita;

3. para outras configurações, τ está situado entre esses dois valores; e o aumento do seu valor de -1 até 1 significa um aumento da concordância entre os ranks.

Para calcular o valor do coeficiente τ entre dois rankings, procedemos da seguinte forma:

- Seja n o número de objetos (tamanho) do rank. O número de pares AB de objetos é então $\begin{pmatrix} n \\ 2 \end{pmatrix} = \frac{n(n-1)}{2}$
- Seja P a quantidade de concordâncias entre os rankings, inicialmente $P = 0$.
- Seja Q a quantidade de discordâncias entre os rankings, inicialmente $Q = 0$.

- Para cada par AB, faça:
  - se ambos os rankings concordarem com a ordem relativa de AB (isto é, se em ambos os rankings A tem um rank maior do que B ou vice-versa), adicionar 1 a P.
  - caso contrário, os rankings discordam da ordem relativa de AB, então adicione 1 a Q.
- O score S entre os dois rankings é definido como $S = P - Q$.
- τ é definido como

$$\frac{\text{Score obtido}}{\text{Score Máximo Possível}}$$

O Score Máximo ocorre quando os rankings concordam em relação a todos os pares, isto é, $P = n(n-1)/2, Q = 0 \Rightarrow S_{\max} = n(n-1)/2$. Portanto, temos

$$\tau = \frac{S}{n(n-1)/2} \tag{B.2}$$

Como $P + Q = n(n-1)/2$ também podemos expressar τ como:

$$\begin{aligned} \tau &= \frac{P-Q}{n(n-1)/2} \\ &= 1 - \frac{2Q}{n(n-1)/2} \qquad \text{(B.3)} \\ &= \frac{2P}{n(n-1)/2} - 1 \qquad \text{(B.4)} \end{aligned}$$

Apêndice C

Análises de Variância: Influência do Grau de Homogeneidade

C.1 Estratégia "o maior número possível" + "o maior número possível"

Tabela C.1: Resultado da análise de variância. Estratégia "o maior número possível" + "o maior número possível". Grupos de tamanho 3

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,34556	2	0,67278	279,72	2,2e-16	0,65321	6,16034
Resíduos	0,71435	297	0,00241				
Total	2,05991	299					

Níveis	Média	Desvio padrão
homogêneo	0,87251	0,03690
normal	0,80849	0,04649
heterogêneo	0,70970	0,06077
Total	0,79690	0,08300

Comparação de médias	5%	1%
LSD	0,01366	0,01800

Diferenças de Médias		
homogêneo - normal	0,06402	**
homogêneo - heterogêneo	0,16281	**
normal - heterogêneo	0,09879	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.2: Resultado da análise de variância. Estratégia "o maior número possível" + "o maior número possível". Grupos de tamanho 6

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,44848	2	0,72424	522,21	2,2e-16	0,77859	4,79323
Resíduos	0,41190	297	0,00139				
Total	1,86038	299					

Níveis	Média	Desvio padrão
homogêneo	0,85927	0,01943
normal	0,78471	0,03763
heterogêneo	0,68948	0,04865
Total	0,77782	0,07888

Comparação de médias	5%	1%
LSD	0,01038	0,01367

Diferenças de Médias		
homogêneo - normal	0,07455	**
homogêneo - heterogêneo	0,16979	**
normal - heterogêneo	0,09523	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.3: Resultado da análise de variância. Estratégia "o maior número possível" + "o maior número possível". Grupos de tamanho 12

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,97671	2	0,48836	724,71	2,2e-16	0,82994	3,36938
Resíduos	0,20014	297	0,00067				
Total	1,17685	299					

Níveis	Média	Desvio padrão
homogêneo	0,84000	0,01518
normal	0,76427	0,02744
heterogêneo	0,70040	0,03222
Total	0,76822	0,06274

Comparação de médias	5%	1%
LSD	0,00720	0,00949

Diferenças de Médias		
homogêneo - normal	0,07572	**
homogêneo - heterogêneo	0,13960	**
normal - heterogêneo	0,06387	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.4: Resultado da análise de variância. Estratégia "o maior número possível" + "o maior número possível". Grupos de tamanho 24

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,71407	2	0,35704	809,12	2,2e-16	0,84492	2,72861
Resíduos	0,13106	297	0,00044				
Total	0,84513	299					

Níveis	Média	Desvio padrão
homogêneo	0,83020	0,01491
normal	0,76519	0,02227
heterogêneo	0,71086	0,02461
Total	0,76875	0,05317

Comparação de médias	5%	1%
LSD	0,00584	0,00769

Diferenças de Médias		
homogêneo - normal	0,06501	**
homogêneo - heterogêneo	0,11935	**
normal - heterogêneo	0,05434	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

C.2 Estratégia "o maior número possível" + "pelo menos a metade"

Tabela C.5: Resultado da análise de variância. Estratégia "o maior número possível" + "pelo menos a metade". Grupos de tamanho 3

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,37807	2	0,68904	255,64	2,2e-16	0,63256	6,55081
Resíduos	0,80050	297	0,00270				
Total	2,17857	299					

Níveis	Média	Desvio padrão
homogêneo	0,87085	0,03877
normal	0,80306	0,04934
heterogêneo	0,70571	0,06441
Total	0,79321	0,08536

Comparação de médias	5%	1%
LSD	0,01446	0,01905

Diferenças de Médias		
homogêneo - normal	0,06779	**
homogêneo - heterogêneo	0,16514	**
normal - heterogêneo	0,09734	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.6: Resultado da análise de variância. Estratégia "o maior número possível" + "pelo menos a metade". Grupos de tamanho 6

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,41128	2	0,70564	470,38	2,2e-16	0,76005	4,99081
Resíduos	0,44555	297	0,00150				
Total	1,85683	299					

Níveis	Média	Desvio padrão
homogêneo	0,85781	0,01983
normal	0,78028	0,0412770
heterogêneo	0,689970	0,04902
Total	0,77602	0,07880

Comparação de médias	5%	1%
LSD	0,01078	0,01420

Diferenças de Médias		
homogêneo - normal	0,07752	**
homogêneo - heterogêneo	0,16784	**
normal - heterogêneo	0,09032	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.7: Resultado da análise de variância. Estratégia "o maior número possível" + "pelo menos a metade". Grupos de tamanho 12

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,96335	2	0,48167	646,94	2,2e-16	0,81331	3,55362
Resíduos	0,22113	297	0,00074				
Total	1,18448	299					

Níveis	Média	Desvio padrão
homogêneo	0,83882	0,01620
normal	0,75685	0,03004
heterogêneo	0,70083	0,03269
Total	0,76550	0,06294

Comparação de médias	5%	1%
LSD	0,00757	0,00997

Diferenças de Médias		
homogêneo - normal	0,08197	**
homogêneo - heterogêneo	0,13799	**
normal - heterogêneo	0,05602	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.8: Resultado da análise de variância. Estratégia "o maior número possível" + "pelo menos a metade". Grupos de tamanho 24

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,73210	2	0,36605	761,06	2,2e-16	0,83673	2,86653
Resíduos	0,14285	297	0,00048				
Total	0,87495	299					

Níveis	Média	Desvio padrão
homogêneo	0,82927	0,01560
normal	0,75406	0,02460
heterogêneo	0,70957	0,02438
Total	0,76430	0,05409

Comparação de médias	5%	1%
LSD	0,00610	0,00803

Diferenças de Médias		
homogêneo - normal	0,07521	**
homogêneo - heterogêneo	0,11970	**
normal - heterogêneo	0,04449	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

C.3 Estratégia "pelo menos a metade" + "a maioria"

Tabela C.9: Resultado da análise de variância. Estratégia "pelo menos a metade" + "A Maioria". Grupos de tamanho 3

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,33124	2	0,66562	278,73	2,2e-16	0,65241	6,11367
Resíduos	0,70925	297	0,00239				
Total	2,04049	299					

Níveis	Média	Desvio padrão
homogêneo	0,87463	0,03654
normal	0,81153	0,04677
heterogêneo	0,71277	0,06035
Total	0,79964	0,08261

Comparação de médias	5%	1%
LSD	0,01361	0,01792

Diferenças de Médias		
homogêneo - normal	0,06310	**
homogêneo - heterogêneo	0,16187	**
normal - heterogêneo	0,09876	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.10: Resultado da análise de variância. Estratégia "pelo menos a metade" + "A Maioria". Grupos de tamanho 6

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,46434	2	0,73217	512,03	2,2e-16	0,77518	4,84225
Resíduos	0,42469	297	0,00143				
Total	1,88903	299					

Níveis	Média	Desvio padrão
homogêneo	0,86285	0,01840
normal	0,78785	0,03882
heterogêneo	0,69214	0,04944
Total	0,78095	0,07948

Comparação de médias	5%	1%
LSD	0,01052	0,01386

Diferenças de Médias		
homogêneo - normal	0,07500	**
homogêneo - heterogêneo	0,17072	**
normal - heterogêneo	0,09571	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.11: Resultado da análise de variância. Estratégia "pelo menos a metade" + "A Maioria". Grupos de tamanho 12

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,00954	2	0,50477	719,49	2,2e-16	0,82891	3,43562
Resíduos	0,20837	297	0,00070				
Total	1,21791	299					

Níveis	Média	Desvio padrão
homogêneo	0,84271	0,01589
normal	0,76685	0,02831
heterogêneo	0,70072	0,03242
Total	0,77009	0,06382

Comparação de médias	5%	1%
LSD	0,00736	0,00970

Diferenças de Médias		
homogêneo - normal	0,07585	**
homogêneo - heterogêneo	0,14198	**
normal - heterogêneo	0,06613	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.12: Resultado da análise de variância. Estratégia "pelo menos a metade" + "A Maioria". Grupos de tamanho 24

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,77310	2	0,38655	829,10	2,2e-16	0,84810	2,81761
Resíduos	0,13847	297	0,00047				
Total	0,91157	299					

Níveis	Média	Desvio padrão
homogêneo	0,83330	0,01655
normal	0,76587	0,02243
heterogêneo	0,70911	0,02493
Total	0,76943	0,05522

Comparação de médias	5%	1%
LSD	0,00603	0,00795

Diferenças de Médias		
homogêneo - normal	0,06743	**
homogêneo - heterogêneo	0,12419	**
normal - heterogêneo	0,05677	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

C.4 Estratégia "a maioria" + "o maior número possível"

Tabela C.13: Resultado da análise de variância. Estratégia "a maioria" + "o maior número possível". Grupos de tamanho 3

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,34562	2	0,67281	277,85	2,2e-16	0,65170	6,15508
Resíduos	0,71917	297	0,00242				
Total	2,06479	299					

Níveis	Média	Desvio padrão
homogêneo	0,87452	0,03684
normal	0,81136	0,04652
heterogêneo	0,71182	0,06118
Total	0,79923	0,08310

Comparação de médias	5%	1%
LSD	0,01369	0,01804

Diferenças de Médias		
homogêneo - normal	0,06316	**
homogêneo - heterogêneo	0,16270	**
normal - heterogêneo	0,09954	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.14: Resultado da análise de variância. Estratégia "a maioria" + "o maior número possível". Grupos de tamanho 6

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	1,40878	2	0,70439	514,76	2,2e-16	0,77611	4,74852
Resíduos	0,40641	297	0,00137				
Total	1,81519	299					

Níveis	Média	Desvio padrão
homogêneo	0,85996	0,01944
normal	0,78598	0,03715
heterogêneo	0,69248	0,04845
Total	0,77947	0,07792

Comparação de médias	5%	1%
LSD	0,01030	0,01357

Diferenças de Médias		
homogêneo - normal	0,07399	**
homogêneo - heterogêneo	0,16748	**
normal - heterogêneo	0,09349	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.15: Resultado da análise de variância. Estratégia "a maioria" + "o maior número possível". Grupos de tamanho 12

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,95547	2	0,47773	720,15	2,2e-16	0,82904	3,34871
Resíduos	0,19703	297	0,00066				
Total	1,15250	299					

Níveis	Média	Desvio padrão
homogêneo	0,83877	0,01664
normal	0,76193	0,02764
heterogêneo	0,70083	0,03082
Total	0,76717	0,06208

Comparação de médias	5%	1%
LSD	0,00715	0,00942

Diferenças de Médias		
homogêneo - normal	0,07684	**
homogêneo - heterogêneo	0,13794	**
normal - heterogêneo	0,06110	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela C.16: Resultado da análise de variância. Estratégia "a maioria" + "o maior número possível". Grupos de tamanho 24

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,72441	2	0,36221	736,53	2,2e-16	0,83221	2,89540
Resíduos	0,14606	297	0,00049				
Total	0,87047	299					

Níveis	Média	Desvio padrão
homogêneo	0,82699	0,01667
normal	0,75965	0,02332
heterogêneo	0,70692	0,02557
Total	0,76452	0,05396

Comparação de médias	5%	1%
LSD	0,00616	0,00812

Diferenças de Médias		
homogêneo - normal	0,06734	**
homogêneo - heterogêneo	0,12007	**
normal - heterogêneo	0,05273	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Apêndice D

Análises de Variância: Influência do Tamanho do Grupo

D.1 Estratégia "o maior número possível" + "o maior número possível"

Tabela D.1: Resultado da análise de variância. Estratégia "o maior número possível" + "o maior número possível". Grupos homogêneos

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,10836	3	0,03612	65,93	2,2e-16	0,33308	2,75245
Resíduos	0,21697	396	0,00055				
Total	0,32534	399					

Níveis	Média	Desvio padrão
03	0,87251	0,03690
06	0,85927	0,01943
12	0,84000	0,01518
24	0,83020	0,01491
Total	0,85049	0,02855

Comparação de médias	5%	1%
LSD	0,00651	0,00857

Diferenças de Médias		
03 - 06	0,01325	**
03 - 12	0,03251	**
03 - 24	0,04231	**
06 - 12	0,01927	**
06 - 24	0,02906	**
12 - 24	0,00979	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela D.2: Resultado da análise de variância. Estratégia "o maior número possível" + "o maior número possível". Grupos normais

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,12987	3	0,04329	35,87	2,2e-16	0,21370	4,45581
Resíduos	0,47786	396	0,00121				
Total	0,60773	399					

Níveis	Média	Desvio padrão
03	0,80849	0,04649
06	0,78471	0,03763
12	0,76427	0,02744
24	0,76519	0,02227
Total	0,78067	0,03903

Comparação de médias	5%	1%
LSD	0,00967	0,01273

Diferenças de Médias		
03 - 06	0,02378	**
03 - 12	0,04422	**
03 - 24	0,04330	**
06 - 12	0,02044	**
06 - 24	0,01952	**
12 - 24	−0,00092	

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela D.3: Resultado da análise de variância. Estratégia "o maior número possível" + "o maior número possível". Grupos heterogêneos

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,02955	3	0,00985	5,12	0,00175	0,03730	6,25266
Resíduos	0,76261	396	0,00193				
Total	0,79216	399					

Níveis	Média	Desvio padrão
03	0,70970	0,06077
06	0,68948	0,04865
12	0,70040	0,03222
24	0,71086	0,02461
Total	0,70261	0,04456

Comparação de médias	5%	1%
LSD	0,01221	0,01608

Diferenças de Médias		
03 - 06	0,02022	**
03 - 12	0,00930	
03 - 24	−0,00116	
06 - 12	−0,01092	
06 - 24	−0,02138	**
12 - 24	−0,01046	

* significativo ao nível de 0,05

** significativo ao nível de 0,01

D.2 Estratégia "o maior número possível" + "pelo menos a metade"

Tabela D.4: Resultado da análise de variância. Estratégia "o maior número possível" + "pelo menos a metade". Grupos homogêneos.

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,10479	3	0,03493	58,17	2,2e-16	0,30589	2,88451
Resíduos	0,23778	396	0,00060				
Total	0,34257	399					

Níveis	Média	Desvio padrão
03	0,87085	0,03877
06	0,85781	0,01983
12	0,83882	0,01620
24	0,82927	0,01560
Total	0,84919	0,02930

Comparação de médias	5%	1%
LSD	0,00681	0,00897

Diferenças de Médias		
03 - 06	0,01304	**
03 - 12	0,03203	**
03 - 24	0,04158	**
06 - 12	0,01899	**
06 - 24	0,02854	**
12 - 24	0,00955	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela D.5: Resultado da análise de variância. Estratégia "o maior número possível" + "pelo menos a metade". Grupos normais

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,15748	3	0,05249	37,19	2,2e-16	0,21981	4,85414
Resíduos	0,55897	396	0,00141				
Total	0,71645	399					

Níveis	Média	Desvio padrão
03	0,80306	0,04934
06	0,78029	0,04128
12	0,75685	0,03004
24	0,75406	0,02460
Total	0,77357	0,04237

Comparação de médias	5%	1%
LSD	0,01044	0,01374

Diferenças de Médias		
03 - 06	0,02277	**
03 - 12	0,04621	**
03 - 24	0,04900	**
06 - 12	0,02344	**
06 - 24	0,02623	**
12 - 24	0,00279	

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela D.6: Resultado da análise de variância. Estratégia "o maior número possível" + "pelo menos a metade". Grupos heterogêneos

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,02163	3	0,00721	3,51	0,01540	0,02591	6,45411
Resíduos	0,81328	396	0,00205				
Total	0,83491	399					

Níveis	Média	Desvio padrão
03	0,70571	0,06441
06	0,68997	0,04902
12	0,70083	0,03269
24	0,70957	0,02438
Total	0,70152	0,04574

Comparação de médias	5%	1%
LSD	0,01259	0,01657

Diferenças de Médias		
03 - 06	0,01574	*
03 - 12	0,00489	
03 - 24	−0,00386	
06 - 12	−0,01086	
06 - 24	−0,01960	**
12 - 24	−0,00874	

* significativo ao nível de 0,05

** significativo ao nível de 0,01

D.3 Estratégia "pelo menos a metade" + "a maioria"

Tabela D.7: Resultado da análise de variância. Estratégia "pelo menos a metade" + "a maioria". Grupos homogêneos

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,10584	3	0,03528	64,13	2,2e-16	0,32698	2,74816
Resíduos	0,21785	396	0,00055				
Total	0,32369	399					

Níveis	Média	Desvio padrão
03	0,87463	0,03654
06	0,86285	0,01840
12	0,84271	0,01589
24	0,83330	0,01655
Total	0,85337	0,02848

Comparação de médias	5%	1%
LSD	0,00652	0,00858

Diferenças de Médias		
03 - 06	0,01178	**
03 - 12	0,03193	**
03 - 24	0,04133	**
06 - 12	0,02014	**
06 - 24	0,02955	**
12 - 24	0,00940	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela D.8: Resultado da análise de variância. Estratégia "pelo menos a metade" + "a maioria". Grupos normais

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,13916	3	0,04639	37,12	2,2e-16	0,21949	4,51521
Resíduos	0,49485	396	0,00125				
Total	0,63401	399					

Níveis	Média	Desvio padrão
03	0,81153	0,04677
06	0,78785	0,03882
12	0,76685	0,02831
24	0,76587	0,02243
Total	0,78303	0,03986

Comparação de médias	5%	1%
LSD	0,00983	0,01294

Diferenças de Médias		
03 - 06	0,02368	**
03 - 12	0,04468	**
03 - 24	0,04566	**
06 - 12	0,02099	**
06 - 24	0,02197	**
12 - 24	0,00098	

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela D.9: Resultado da análise de variância. Estratégia "pelo menos a metade" + "a maioria". Grupos heterogêneos

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,02541	3	0,00847	4,37	0,00485	0,03202	6,25926
Resíduos	0,76808	396	0,00194				
Total	0,79349	399					

Níveis	Média	Desvio padrão
03	0,71277	0,06035
06	0,69214	0,04944
12	0,70072	0,03242
24	0,70911	0,02493
Total	0,70368	0,04459

Comparação de médias	5%	1%
LSD	0,01225	0,01612

Diferenças de Médias		
03 - 06	0,02063	**
03 - 12	0,01204	
03 - 24	0,00366	
06 - 12	−0,00859	
06 - 24	−0,01697	**
12 - 24	−0,00839	

* significativo ao nível de 0,05

** significativo ao nível de 0,01

D.4 Estratégia "a maioria" + "o maior número possível"

Tabela D.10: Resultado da análise de variância. Estratégia "a maioria" + "o maior número possível". Grupos homogêneos

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,13560	3	0,04520	78,97	2,2e-16	0,37431	2,81351
Resíduos	0,22667	396	0,00057				
Total	0,36227	399					

Níveis	Média	Desvio padrão
03	0,87452	0,03684
06	0,85996	0,01944
12	0,83877	0,01664
24	0,82699	0,01667
Total	0,85006	0,03013

Comparação de médias	5%	1%
LSD	0,00665	0,00875

Diferenças de Médias		
03 - 06	0,01456	**
03 - 12	0,03575	**
03 - 24	0,04753	**
06 - 12	0,02119	**
06 - 24	0,03297	**
12 - 24	0,01178	**

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela D.11: Resultado da análise de variância. Estratégia "a maioria" + "o maior número possível". Grupos normais

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,17598	3	0,05866	48,36	2,2e-16	0,26813	4,46117
Resíduos	0,48034	396	0,00121				
Total	0,65632	399					

Níveis	Média	Desvio padrão
03	0,81136	0,04652
06	0,78598	0,03715
12	0,76193	0,02764
24	0,75965	0,02332
Total	0,77973	0,04056

Comparação de médias	5%	1%
LSD	0,00967	0,01273

Diferenças de Médias		
03 - 06	0,02539	**
03 - 12	0,04944	**
03 - 24	0,05171	**
06 - 12	0,02405	**
06 - 24	0,02632	**
12 - 24	0,00227	

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Tabela D.12: Resultado da análise de variância. Estratégia "a maioria" + "o maior número possível". Grupos heterogêneos

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado médio	Valor F	Valor P	R-SQUARE	C.V
Tau médio	0,02085	3	0,00695	3,61	0,01343	0,02665	6,23285
Resíduos	0,76165	396	0,00192				
Total	0,78250	399					

Níveis	Média	Desvio padrão
03	0,71182	0,06118
06	0,69248	0,04845
12	0,70083	0,03082
24	0,70692	0,02557
Total	0,70301	0,04428

Comparação de médias	5%	1%
LSD	0,01218	0,01604

Diferenças de Médias		
03 - 06	0,01934	**
03 - 12	0,01099	
03 - 24	0,00490	
06 - 12	−0,00835	
06 - 24	−0,01444	*
12 - 24	−0,00609	

* significativo ao nível de 0,05

** significativo ao nível de 0,01

Índice