RESUMO

Sistemas de Recomendação baseados Cooperação Indireta podem ser implementados em bibliotecas por meio da aplicação de conceitos e procedimentos de Análise de Redes. Uma medida de Distância Temática, inicialmente desenvolvida para variáveis dicotômicas, foi generalizada e aplicada a matrizes de co-ocorrências, permitindo o aproveitando de toda a informação disponível sobre o comportamento dos usuários com relação aos itens consultados. Como resultado formaram-se Subgrupos Especializados altamente coerentes, para os quais Listas-Base e Listas Personalizadas foram geradas da maneira usual. Aplicativos programáveis capazes de manipularem matrizes, como o software S-plus, foram utilizados para os cálculos (com vantagens sobre o software especializado UCINET 5.0), sendo suficientes para o processamento de Grupos Temáticos de até 10.000 usuários.

PALAVRAS-CHAVE

Análise de Agrupamentos; Análise de Cestas; Análise de Redes; Bibliotecas; Distância Temática; Filtro Colaborativo; Sistemas de Recomendação.

ABSTRACT

Recommendation Systems based on Indirect Cooperation can be implemented in libraries with the use of Network Analysis concepts and procedures. In this project, a Thematic Distance measurement, initially defined for dichotomic variables, was generalized and applied to co-occurrences matrices, allowing all available information about patrons behavior in relation to library items to be used in the identification of Specialized Subgroups. As a result, extremely coherent Base-Lists and Personalized-Lists were generated in the usual way. Programmable applications capable of handling matrices, such as S-plus, were used for calculations (with

advantages over UCINET 5.0, a Network Analysis specialized software), proving adequate for the processing of Thematic Groups of up to 10.000 patrons.

KEY WORDS

Basket Analysis; Cluster Analysis; Cooperative Filtering; Libraries; Network Analysis; Recommendation Systems; Thematic Distance.

SUMÁRIO

ANÁLISE DE REDES EM PROCEDIMENTOS DE COOPERAÇÃO INDIRETA: UTILIZAÇÃO NO SISTEMA DE RECOMENDAÇÕES DA BIBLIOTECA KARL A. BOEDECKER*

Francisco Aranha[1]

I. INTRODUÇÃO: DE PORTEIROS A PORTAIS

Com a perda do monopólio sobre a informação, as bibliotecas tradicionais devem encontrar novas maneiras de agregar valor a seus serviços para continuarem competitivas frente às diversas e sofisticadas fontes alternativas hoje disponíveis aos usuários, particularmente aquelas que podem ser acessadas pela Internet. Somente um ativa reformulação de seu papel e objetivos poderá garantir a continuidade a longo prazo destas instituições (CARSON *et al*., 1997).

Dentre as mudanças necessárias na filosofia na gestão de bibliotecas, uma das mais profundas e indispensáveis será a transferência do foco **no acervo** para o foco **no usuário e seu comportamento**. Os profissionais da informação deverão deixar de ser porteiros ("*gatekeepers*"), meros organizadores e controladores do acesso às estantes, para tornarem-se portais ("*gateways*"), isto é, **mapeadores e auditores** de fontes internas **e externas** de informações relevantes ao usuário (Stephen Abram, segundo MILLER, 1998).

Além disso, as bibliotecas serão compelidas a implantar uma estratégia de relacionamento com usuários baseada nos princípios do marketing um-a-um. Estes

* O NPP agradece aos alunos que participaram da pesquisa que originou o presente relatório como auxiliar de pesquisas, Patrícia Rosa, e como monitor de pesquisas, Sandro Venezuela.
[1] Colaboraram os auxiliares de pesquisa Sandro Venezuela e Patrícia Rosa.

se traduzem na "disposição e na possibilidade da organização mudar de comportamento com relação a usuários individuais, com base no que o usuário diz e no que se sabe sobre ele" (PEPPERS e ROGERS, 1999, p. 151).

1. SISTEMA DE RECOMENDAÇÕES COM BASE EM COOPERAÇÃO INDIRETA

Uma das formas mais imediatas e eficazes pelas quais uma biblioteca pode gerar valor para os usuários é disponibilizar-lhes um Sistema de Recomendações capaz de gerar listas individualizadas de itens, do acervo próprio e do acervo de terceiros, relevantes para suas pesquisas.

Sistemas de Recomendações análogos ao das populares livrarias virtuais podem ser implementados também em bibliotecas, por meio de sistemas de **cooperação indireta**, isto é, por meio de

> "procedimentos que transformam as atividades coletivas de um grupo em fonte de informação personalizada para seus membros, permitindo a cada um beneficiar-se de maneira particular do conhecimento implícito no processo de coleta e consolidação de informação dos demais membros do grupo" (PAYTON, 1998, p.1).

A utilização desta estratégia no contexto de bibliotecas corresponde a utilizar conhecimento sobre

- os itens já manipulados por cada particular consulente e
- a movimentação total de itens do acervo,

para gerar recomendações a subgrupos especializados de usuários.

2. COLABORAÇÃO INDIRETA NA BIBLIOTECA KARL A. BOEDECKER: O PRIMEIRO PROJETO

A Biblioteca Karl A. Boedecker (KAB) foi criada em 1954 como o objetivo de fornecer apoio bibliográfico às atividades de ensino e pesquisa desenvolvidas pela comunidade acadêmica da Escola de Administração de Empresa de São Paulo (EAESP). Seu acervo compõe-se de cerca de 67 mil exemplares de 47 mil títulos de livros; e de 3,2 mil exemplares de 1,8 mil títulos de teses e dissertações (BIBLIOTECA KARL A. BOEDECKER, 1998).

Ao longo de 1999, a KAB apoiou a realização de um projeto piloto de Sistema de Recomendações de títulos a seus usuários, nos moldes descritos acima, isto é, baseado em cooperação indireta. O projeto foi financiado e patrocinado pelo NPP - Núcleo de Pesquisas e Publicações da EAESP/FGV e resultou em um protótipo totalmente operacional, um artigo recentemente publicado (ARANHA, 2000) e num Relatório de Pesquisa atualmente em fase de editoração.

Para o teste piloto foram analisadas 22.500 transações de empréstimo ocorridas entre 01/03 e 19/07/99 e selecionados 410 usuários entre professores e alunos de pós-graduação para a geração das recomendações. SANTOS (1999) desenhou e implementou o necessário *data warehouse* de transações, utilizando infra-estrutura tecnológica cedida pela Informix do Brasil.

O protótipo desenvolvido está organizado em torno de uma funcionalidade crítica (ARANHA, 2000; item III.1) que é a identificação de Grupos Temáticos (GT) e de Subgrupos Especializados (SGE). Este núcleo funcional foi implementado por meio de técnicas de Análise de Agrupamentos, embora o levantamento bibliográfico inicial, realizado por ocasião do desenvolvimento do piloto do Sistema de Recomendações, tivesse indicado claramente vantagens da aplicação de técnicas de Análise de Redes.

Foram três os motivos da escolha desta linha de ação:

- de um lado, os textos sobre Análise de Redes eram demasiado extensos e complexos para serem enfrentados num momento em que havia outros problemas mais urgentes a resolver – como a coleta e exploração dos dados de circulação da biblioteca, a fundamentação dos conceitos de cooperação indireta, e a investigação de conceitos de consolidação e ligação de dados (ARANHA, em editoração);
- de outro lado, a aplicação de Análise de Redes exigia a utilização de software específico, que deveria ainda ser adquirido e estudado;
- as técnicas de Análise de Agrupamento, ao contrário, eram amplamente conhecidas pelo pesquisador e o software necessário estava disponível e compreendido.

A utilização de técnicas de Análise de Rede foi programada para um segundo projeto, que é o objeto deste relatório.

3. ANÁLISE DE REDES NO SISTEMA DE RECOMENDAÇÕES: O SEGUNDO PROJETO

Apesar do Sistema de Recomendações desenvolvido no projeto anterior ter obtido sucesso e se revelado eficaz, restou em aberto a questão da utilização de Análise de Redes como técnica analítica do sistema.

Neste segundo projeto de pesquisa, investigamos:

- aspectos da teoria de Análise de Redes relevantes ao problema da cooperação indireta;

- técnicas e ferramentas de análise de rede adequadas à identificação dos Subgrupos Especializados – etapa crítica de funcionamento do Sistema de Recomendações; e
- vantagens e desvantagens da análise de redes em relação à análise de agrupamentos.

4. SOFTWARE UTILIZADO

O software escolhido para executar a análise dos dados neste projeto é o "UCINET 5.0 for Windows" (BORGATTI, EVERETT e FREEMAN, 1999; http://www.analytictech.com). Esta escolha deve-se ao fato do UCINET 5.0 realizar, entre outras atividades, funções e rotinas eficientes voltadas para análise de rede. Ao final do trabalho, no entanto, acabamos migrando para o software de estatística S-plus, que apresentava uma linguagem de programação conveniente, permitia um maior controle dos procedimentos e oferecia uma interface mais amigável e mais sofisticada.

5. ESTRUTURA DO TRABALHO

A Primeira Parte do trabalho, dividida em cinco capítulos, apresenta o Sistema de Recomendações desenvolvido na Biblioteca Karl A. Boedecker ao longo do ano de 1999 e especifica a questão de pesquisa objeto do presente estudo, relativa à aplicação de conceitos e técnicas de Análise de Redes aos procedimentos de cooperação indireta.

A Segunda Parte analisa o efeito do excesso de informações (Capítulo 1) e discute conceitos de Personalização (Capítulo 2), Filtro por Conteúdo (Capítulo 3), Cooperação Indireta (Capítulo 4) e Análise de Redes (Capítulo 5).

Uma descrição detalhada das estratégias utilizadas na revisitação dos dados do projeto anterior é apresentada na Terceira Parte. O Capítulo 1 relata as etapas de implementação do projeto; o Capítulo 2 descreve o processo de criação de Grupos Temáticos; e o Capítulo 3, o processo de formação de Subgrupos Especializados. A criação das Listas-Base e das Listas Personalizadas é o assunto do Capítulo 4.

Por fim, a Quarta Parte apresenta conclusões e oportunidades de novas pesquisas, respectivamente nos Capítulos 1 e 2.

II. FILTRO COOPERATIVO E ANÁLISE DE REDES

Neste item II, discutem-se tópicos da teoria de Gestão do Conhecimento e Análise de Redes necessários à implementação dos procedimentos utilizados no item III.

1. EFEITO RAUL SEIXAS

Com a informática e as telecomunicações, o volume de informações disponíveis sobre todos os aspectos da atividade humana e em todos os ramos da ciência passou a crescer a taxas vertiginosas: “Nos últimos trinta anos, produziu-se um volume de informações novas maior do que nos cinco mil anos precedentes. Cerca de mil livros são publicados no mundo por dia e o total do conhecimento impresso duplica a cada oito anos” (LARGE, 1984).

A contrapartida analítica desta abundância de dados não é proporcional: o crescimento da capacidade de acesso dos usuários aos dados é muito inferior ao crescimento da oferta; o aumento da capacidade de “digestão” e entendimento dos dados é ainda mais restrita. Uma estimativa eloqüente indica que o leitor contemporâneo de um jornal tradicional recebe mais informação em um único domingo do que um cidadão vivendo na Inglaterra do século XVII receberia em toda

a sua vida (WURMAN, 1991). O resultado final é antes um "estado de choque" intelectivo, ou uma "indigestão" de dados, do que uma expansão da compreensão ou do conhecimento. O "efeito Raul Seixas" ("É tanta coisa no menu, que eu nem sei o que comer...") manifesta-se de maneira generalizada. Alguns exemplos atuais:

- No esforço de organizar o conteúdo da Internet, os portais de acesso acabam tão complexos quanto o conteúdo que queriam organizar, passando "de portais a labirintos" (ANGULO e ALBERTIN, 2000);

- Sites de varejo passam a oferecer tantos e tão variados produtos que a busca de informação sobre os itens torna-se proibitivamente complexa e demorada; não é de se estranhar, portanto, que, segundo Jeff Bezos, criador da Amazon.com, o principal motivo pelo qual clientes voltam a uma particular livraria virtual é justamente o grau de **ajuda ativa** que o *site* oferece na localização de títulos de interesse para o cliente (RAMO, 1999);

- A programação de TVs por assinatura é tão extensa em número de canais e de opções de conteúdo que freqüentemente o assinante sequer toma conhecimento da disponibilidade de programas de seu interesse. Em conseqüência, começam a prosperar serviços como o "Personalized Television" (SMYTH e COTTER, 2000), que, pela Internet, disponibiliza para 20.000 usuários da Inglaterra e da Irlanda sugestões personalizadas de programas de TV.

2. PERSONALIZAÇÃO DE CONTEÚDO

O objetivo da personalização de conteúdo é garantir que a pessoa certa receba a informação certa no momento certo.

Esta confluência de acertos caracteriza a **relevância** da informação. Empresas de marketing têm enfatizado a necessidade do emissor de mensagens ir além da

personalização superficial, cuidando da relevância como indicador da qualidade: “o consumidor está dizendo que não adianta receber uma comunicação personalizada se ela não for relevante para seus interesses e necessidades” (ROSENWALD, 2000). Na mundo digital, entre “mares de *spam*, irritantes assistentes em forma de clipes, e *websites* impossivelmente vastos, não é surpresa que estejamos entrando na Renascença dos Mecanismos de Busca *(Search Engines*)” (DUNGAN, 2000).

O aumento da eficiência na comunicação é do interesse tanto de quem busca informação (pesquisador, usuário de biblioteca, usuário da Internet, consumidor) quanto de quem a fornece (portal, biblioteca, site, empresa).

Há duas estratégias básicas para personalizar o conteúdo de comunicações: o filtro de conteúdo e a cooperação indireta.

3. FILTRO POR CONTEÚDO

O Filtro por Conteúdo é uma das técnicas já tradicionais de recuperação de informações (SMYTH e COTTER, 2000). Seu sucesso apoia-se na habilidade de

- **representar acuradamente cada item** de informação no acervo, com base em um subconjunto de suas características, principalmente seu enquadramento em uma tipologia temática; e de

- **representar o interesse do usuário** através de um perfil baseado no mesmo subconjunto de características extraídas dos itens.

Neste contexto, a relevância de um item para um usuário é proporcional **à similaridade do perfil do item com o perfil do usuário**; os itens selecionados para serem submetidos à atenção do usuário são os mais parecidos com o seu perfil de interesse.

O problema da utilização de Filtros de Conteúdo é justamente a necessidade de se caracterizarem os itens e o interesse dos usuários: esta atividade é complexa, onerosa, e requer conhecimento especializado sobre cada campo do conhecimento humano. Numa biblioteca, a caracterização dos itens corresponde ao processo de classificação, indexação e catalogação dos novos livros ou periódicos; por outro lado, o próprio usuário se encarrega de caracterizar seu interesse, fornecendo informações aos mecanismos de busca ou pesquisando a base de dados a partir do índice de assuntos. Novos tipos de suporte de informação, como vídeos, CDs, etc, podem exigir uma linguagem própria para esta classificação.

Mesmo concluída com sucesso a etapa de caracterização dos itens, as recomendações por meio de Filtro de Conteúdo sempre terão um escopo limitado, uma vez que esta técnica somente seleciona itens parecidos com o perfil do usuário; como este é geralmente identificado pelo registro de escolhas anteriores, **um subespaço temático tende a ser reforçado**, principalmente no caso de usuários novos para o sistema. Os itens recomendados podem ser relevantes, mas nem de longe correspondem à amplitude total dos interesses do usuário.

4. COOPERAÇÃO INDIRETA

Em contraposição ao Filtro por Conteúdo, a técnica de Cooperação Indireta move-se para além da experiência de cada usuário isolado: procura valer-se da união das experiências de grupos de usuários.

Em vez de identificar a semelhança entre usuários e itens, **busca localizar usuários parecidos entre si**.

Sendo parecidos os membros de um grupo, o que se descobre pelo monitoramento de seu comportamento no sistema, infere-se que itens que interessaram a um membro do grupo interessarão aos demais membros.

A grande vantagem desta técnica é não ser necessário entrar no mérito do conteúdo dos itens. O problema é que itens novos, conhecidos por poucos usuários, demoram a aparecer como sugestões. Isto é, novos itens têm um alto período de latência.

Várias iniciativas de implementação de sistemas de cooperação indireta foram documentadas recentemente. PAYTON (1998) procurou facilitar o contato entre pessoas com interesses comuns, explicitando seu padrão de navegação na Internet. KAUTZ, SELMAN e SHAH (1997a, 1997b) caracterizaram redes de pesquisadores que mantinham vínculos sociais, analisando a co-ocorrência de nomes em documentos públicos na Internet. SCHWARTZ e WOOD (1993) identificaram colaboradores potenciais pela análise do tráfego de emails em pontos selecionados da rede. SWANSON e SMALHEISER (1999) desenvolveram o software Arrowsmith para identificar relações pouco evidentes entre achados científicos na área de biomédicas.

Caminhando em direção oposta, isto é, perseguindo o objetivo de **dificultar** a cooperação, foram publicados vários trabalhos na área de detecção de fraudes contra companhias seguradoras e de saúde (CABENA e OUTROS, 1998), combate à lavagem de dinheiro e combate ao crime organizado em geral (JENSEN, 1997; HANN, 1998).

Dentre estes trabalhos, o de SCHWARTZ e WOOD (1993) é de interesse mais imediato ao Sistema de Recomendações da biblioteca. Entre outros *insights* interessantes, dá pistas quanto à conveniência da utilização das técnicas de Análise de Redes no mecanismo central do Sistema de Recomendações.

5. ANÁLISE DE REDES

As técnicas de Análise de Redes (KNOKE e KUKLINSKY, 1982; WASSERMAN e FAUST, 1992; WASSERMAN e GALASKEWICZ, 1994) caracterizam um objeto

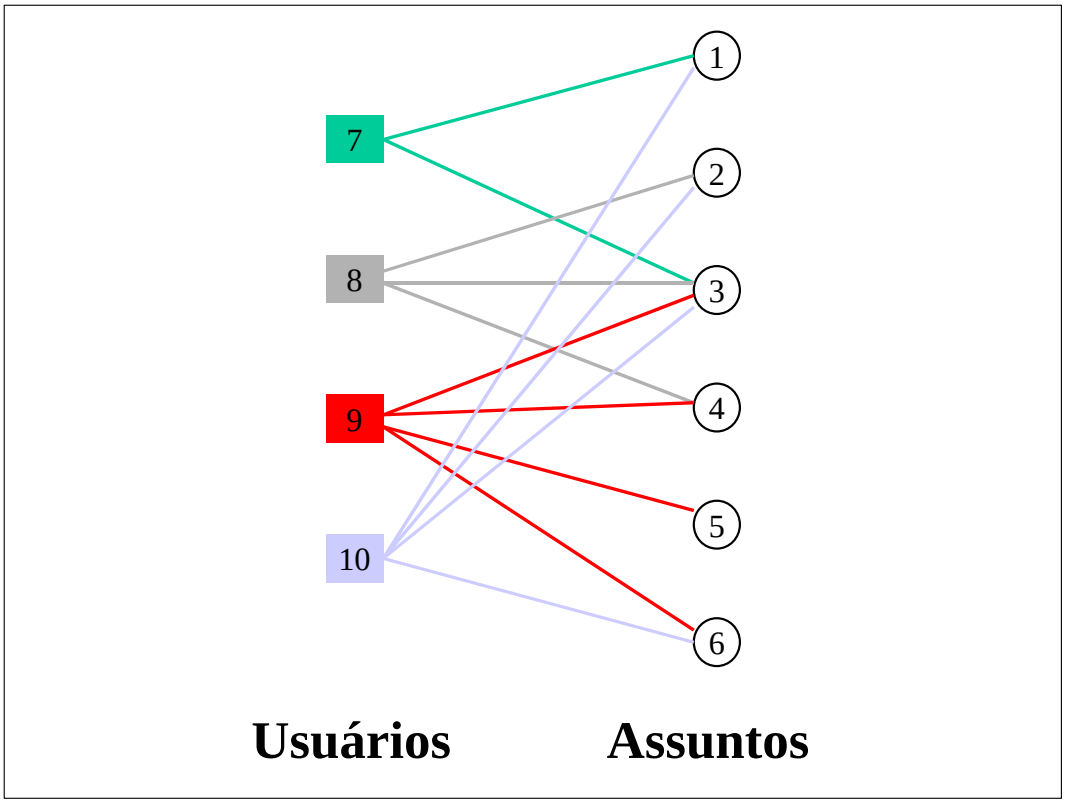
de interesse pelas suas **relações** com os demais objetos em estudo, em vez de os caracterizar por seus atributos individuais. Redes de relacionamentos podem ser especificadas por amizade, dominância, comunicação, etc. (KNOKE e KUKLINSKI, 1982).

Neste projeto de criação de um Sistema de Recomendações, as relações entre Usuários da biblioteca são estabelecidas pela consulta a Assuntos Significativos comuns. Assuntos Significativos (AS) são 42 categorias de assuntos obtidas através do artifício de se considerarem progressivamente menos dígitos da classificação CDU (Classificação Decimal Universal) de cada item, de forma a balancear entre elas o número de itens transacionados (ARANHA, 2000; BARRY e LINOFF 1997; item III.2.1).

A seguir definimos e interpretamos alguns conceitos de Análise de Redes (KNOKE e KUKLINSKY, 1982; WASSERMAN e FAUST, 1992; WASSERMAN e GALASKEWICZ, 1994) que serão aplicados no desenvolvimento das recomendações ao usuários da biblioteca. Na verdade, nos tópicos relevantes para este trabalho, a teoria é simples. A dificuldade enfrentada na elaboração do projeto foi identificar as estruturas e operações matriciais úteis à cooperação indireta e interpretar o seu significado nos termos do problema.

Figura 1

Rede de Usuários e Assuntos Significativos



5.1. Nó e ligação

No contexto de Análise de Redes, um “nó” é um objeto de interesse, seja ele, por exemplo, um livro, um assunto ou uma pessoa. Estes objetos podem estar envolvidos em “relações”, isto é, em ações ou qualidades que só existem se dois ou mais objetos são considerados conjuntamente. Estas “relações” são também chamadas de “ligações” ou “arestas”.

Uma “ligação” não é, portanto, uma característica intrínseca de nenhum dos envolvidos na relação, mas uma “propriedade emergente” (KNOKE e KUKLINSKI, 1982) da conexão ou elo entre os nós.

Na Figura 1 há um exemplo de rede em que Usuários (nós 7 a 10) relacionam-se com Assuntos Significativos (nós 1 a 6). O Usuário 7 conecta-se à rede através das arestas (7,1) e (7,3). O Usuário 8 conecta-se através das arestas (8,2), (8,3) e (8,4). E assim por diante.

5.2. Matriz de vizinhança

Uma matriz $\mathbf{V} = \{v_{ij}\}$ indica as relações de vizinhança entre dois nós, *i* e *j*, assumindo o valor 0, se os nós não são vizinhos, e o valor 1, caso contrário.

Figura 2

Matriz de Vizinhança Correspondente à Rede Representada na Figura 1

$$V = \left[\begin{array}{cccccc:cccc} 0&0&0&0&0&0&1&0&0&1\\ 0&0&0&0&0&0&0&1&0&1\\ 0&0&0&0&0&0&1&1&1&1\\ 0&0&0&0&0&0&0&1&1&0\\ 0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&1&1\\ \hdashline 1&0&1&0&0&0&0&0&0&0\\ 0&1&1&1&0&0&0&0&0&0\\ 0&0&1&1&1&1&0&0&0&0\\ 1&1&1&0&0&1&0&0&0&0 \end{array}\right] = \left[\begin{array}{c:c} AA & AU \\ \hdashline UA & UU \end{array}\right]$$

A Figura 2 representa, na forma de matriz de vizinhança, o exemplo de rede dado na Figura 1. Na matriz, os 6 assuntos e 4 usuários são indicados nas linhas e colunas, isto é, as primeiras seis linhas e colunas representam os assuntos e as demais quatro linhas e colunas, representam os usuários. Com isso, têm-se que as partições nomeadas como AA e UU recebem valor zero (pois não há ligações diretas entre os assuntos e entre os usuários, respectivamente); e as partições AU e UA têm a interpretação dada à matriz de vizinhança, ligando os nós de assuntos com os nós de usuários.

5.3. Potência da matriz de vizinhança

A conexão indireta de um conjunto de nós em uma rede pode ser revelada elevando-se uma matriz de vizinhança **V** a potências sucessivas, isto é, multiplicando-se a matriz de vizinhança por ela mesma, t vezes.

Os elementos da matriz $\mathbf{V}^t$ indicam o número de conexões em t etapas entre os nós *i* e *j*. A matriz $\mathbf{V}^t$ também fornece indicações sobre a estrutura geral da rede, por exemplo o grau de conexão entre os nós (KNOKE e KUKLINSKI, 1982).

A Figura 3 apresenta o quadrado da matriz **V** definida como exemplo na Figura 2. Em $\mathbf{V}^2$, as partições AA2 e UU2 representam as ligações entre assuntos-assuntos e usuários-usuários, repectivamente. Ou seja, o elemento $v_{3,2}$, cujo valor é 2, indica que os assuntos 2 e 3 ocorreram juntos duas vezes; examinando a Figura 1, podemos verificar que as ocorrências conjuntas foram mediadas pelos usuários 2 e 4.

Figura 3

Matriz V^2: Vizinhos em Dois Passos

$$V^2 = \left[\begin{array}{cccccc:cccc} 2 & 1 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 2 & 4 & 2 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 2 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 1 & 1 & 2 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 3 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 4 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 4 \end{array}\right] = \left[\begin{array}{c:c} AA2 & UA2 \\ \hdashline AU2 & UU2 \end{array}\right]$$

5.4. Matriz de co-ocorrências

Uma matriz de Co-ocorrências $\mathbf{C} = \{c_{ij}\}$ indica o **número de vezes** em que dois nós participam de um mesmo evento. Por exemplo, podemos citar o número de vezes em que dois produtos aparecem juntos na compra de um cliente; ou o número de vezes que dois usuários retiram livros do mesmo assunto.

A Figura 4 apresenta a matriz de Co-ocorrências entre os usuários do exemplo proposto na Figura 1. Com exceção da diagonal, que indica quantos assuntos cada usuário retirou, a matriz mostra o número de assuntos em comum retirados por dois usuários (confronte a matriz com o grafo da Figura 1).

Figura 4

Exemplo de Rede de Usuários e
Correspondente Matriz de Co-ocorrências

$$UU2 = \begin{bmatrix} 2 & 1 & 1 & 2 \\ 1 & 3 & 2 & 2 \\ 1 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{bmatrix}$$

5.5. Matriz de distâncias

Uma matriz de distâncias $\mathbf{D} = \{d_{ij}\}$ indica o "afastamento" entre dois nós, *i* e *j*, assumindo um valor de uma escala contínua entre 0 e infinito. O valor 0 indica que dois nós ocupam a mesma posição.

Embora existam várias definições de medidas de distância (HAIR *et. al*, 1995), e entre elas a mais comumente utilizada seja a distância Euclidiana, neste trabalho utilizaremos o conceito de "distância temática", nas formas "dicotômica" e "generalizada".

5.6. Distância Temática Dicotômica

Uma maneira de caracterizar a intensidade de ligação entre dois nós, *i* e *j*, levando em conta **todos os seus relacionamentos com os demais nós da rede**, foi definida por SCHWARTZ e WOOD (1993) como a proporção dos seus vizinhos não compartilhados (em relação ao total de vizinhos distintos dos dois nós).

Esta medida foi chamada por Schwartz e Wood de “Distância de Interesses” (nome que fazia mais sentido no contexto do problema que estes dois autores estavam estudando). Assim:

$$\text{DistânciaDeInteresses}(i, j) = \begin{cases} \dfrac{\left|[C(n_i) \cup C(n_j)] - [C(n_i) \cap C(n_j)]\right|}{[C(n_i) \cup C(n_j)]}, & C(n_i) \cup C(n_j) \neq 0 \\ 1, & C(n_i) \cup C(n_j) = 0 \end{cases}$$

onde

$C(n_i)$ = número de nós conectados ao nó n_i.

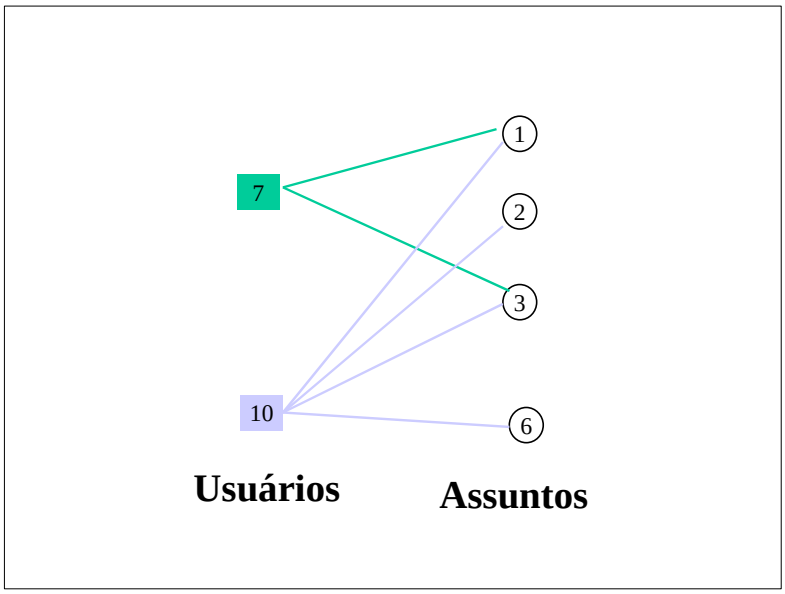
Esta medida varia de zero (valor que ocorre quando dois nós têm todos os vizinhos em comum) até 1 (valor que ocorre quando dois nós não têm nenhum vizinho em comum), indicando uma vizinhança progressivamente menos compartilhada.

Observe, no entanto, que, assim definida, a medida não leva em consideração a **intensidade** das ligações entre os nós, mas apenas a **existência** da ligação. Neste sentido, não utiliza toda a informação disponível sobre as relações de vizinhança entre os nós.

Calculada sobre a matriz que indica a vizinhança entre usuários, a “distância temática dicotômica” indicará a intensidade da “semelhança entre os usuários” no sentido de participarem de um mesmo grupo de relacionamentos. Como os relacionamentos são definidos pela ocorrência de consulta a livros de mesmo tema, o grupo se caracteriza por compartilhar interesse em um tema. Por esse motivo, e adequando o nome da medida ao contexto do presente projeto, vamos “rebatizá-la” como Distância Temática Dicotômica.

Figura 5

Exemplos de Distâncias Temáticas em Rede de Usuários



A Figura 5 reproduz uma parte da Figura 1, retratando os relacionamentos dos usuários 7 e 10. O número total de vizinhos distintos destes dois usuários é 4 (nós 1, 2, 3 e 6). O número de vizinhos comuns entre os usuários 7 e 10 é dois (nós 1 e 3). Assim, a Distância Temática Dicotômica entre eles é 0,5 (2/4).

5.7. Distância Temática Generalizada

No decorrer deste trabalho, verificamos que o conceito de Matriz de Vizinhança **podia ser estendido** para refletir **a intensidade** das relações entre os vizinhos. Assim, em vez de conter apenas valores zero ou um, a matriz **V** poderia conter valores entre 0 e infinito, de tal maneira que números grandes indicariam uma

"vizinhança forte" (e, inversamente, os valores pequenos indicariam uma "vizinhança fraca").

Elevando-se esta matriz **V** ao quadrado, obtém-se uma matriz **C** de Co-ocorrências "Generalizada", em que um valor na diagonal da matriz representa o **nível de atividade total** de um nó, e os valores fora da diagonal representam o **nível de atividade conjunta** entre dois nós.

Por sua vez, a Distância Temática, quando calculada sobre uma matriz de Co-ocorrências Generalizadas, continua variando entre zero e um, e mantém o seu sentido de indicar o grau de vizinhança compartilhada.

Neste caso, pode-se definir a matriz **T** de Distância Temática Generalizada como

$$\mathbf{T} = \frac{\mathbf{1}.\mathbf{diag}^t(\mathbf{C}^2) + \mathbf{diag}(\mathbf{C}^2).\mathbf{1^t} - 2(\mathbf{C}^2)}{\mathbf{1}.\mathbf{diag}^t(\mathbf{C}^2) + \mathbf{diag}(\mathbf{C}^2).\mathbf{1^t} - (\mathbf{C}^2)}$$

onde

1	é um vetor coluna, de dimensão conveniente, formado pelo valor 1 em todas as células;
diag[ ]	é um vetor contendo a diagonal de uma matriz quadrada que entra como argumento da função;
t sobrescrito	indica transposição do vetor ou matriz;
($\mathbf{C}^2$)	é o quadrado da matriz de Co-ocorrências **C**;
.	indica a multiplicação de matrizes.

III. TERCEIRA PARTE: IMPLEMENTAÇÃO DA ANÁLISE

1. ETAPAS DE IMPLEMENTAÇÃO DO SISTEMA DE RECOMENDAÇÕES

Ambos os projetos, o anterior, realizado por análise de agrupamento convencional, e este, que utiliza análise de redes, foram realizados em cinco etapas (Tabela 1).

- **Etapa 1: Coleta e Registro dos Dados.** Nesta etapa os dados foram coletados na biblioteca; foram criados os mecanismos de armazenamento e transmissão dos dados para um data warehouse; a hierarquia de assuntos foi equilibrada em categorias com número aproximado de ocorrências (Assuntos Significativos). O banco de dados formado foi utilizado nos dois projetos.

- **Etapa 2: Pré-processamento.** No primeiro projeto esta etapa consistiu apenas em fazer uma consulta ao banco de dados gerando uma tabela **UA** em que os registros representavam usuários e as colunas os assuntos significativos. Neste segundo projeto, esta etapa consistiu na montagem de uma matriz de dados **D** e no cálculo do seu quadrado **D2** (veja item 2.2).

- **Terceira Etapa: Análise.** No primeiro projeto, realizaram-se duas Análises de Agrupamento sobre a tabela **UA,** uma para definir os Grupos Temáticos e outra para formar os SubGrupos Especializados. No segundo projeto, primeiro calcularam-se as matrizes de co-ocorrência de assuntos (**AA2**) e de co-ocorrência de usuários (**UU2**); em seguida, sobre as matrizes de co-ocorrência calcularam-se matrizes de distância temática; e sobre estas, aplicou Análise de Agrupamentos para formar, respectivamente, os GTs e os SBEs.

- **Quarta Etapa: Formação das Listas-Base**. Em ambos os projetos, consolidaram-se as transações dos usuários classificados em um mesmo SGE, criando uma lista de títulos de interesse potencial para todos os membros.

- **Quinta Etapa: Criação das Listas de Recomendação.** Também em ambos os projetos, para cada usuário distinto, a lista-base é expurgada das transações que o próprio usuário realizou, sendo as demais mantidas como recomendações.

Tabela 1

Etapas de Implementação dos Dois Projetos de Sistemas de Recomendações

	Primeiro Projeto: Análise de Agrupamentos		**Segundo Projeto: Análise de Redes**	
Etapa 1 Coleta e Registro	Coleta de dados da biblioteca, montagem do banco de dados e formação dos Assuntos Significativos			
Etapa 2 Pré-processamento	Definição de Consultas ao Banco de Dados: Tabela **UA**		Leitura dos Dados na Matriz **D**, Cálculo da Matriz **D2**	
	Assuntos Significativos	**Usuários**	**Assuntos Significativos**	**Usuários**
Etapa 3 Análise	**Formação de Grupos Temáticos (GT)** por meio de Análise de Agrupamentos de Variáveis (Algoritmo Hierárquico Aglomerativo – AHA, Correlação como Similaridade) sobre **UA**	**Formação de Subgrupos Especializados (SGE)** por meio de Análise de Agrupamentos de Objetos (AHA, Distância Eucidiana, critério Ward) sobre **UA**	Cálculo da Distância Temática **T(AA2)**	Cálculo da Distância Temática **T(UU2)**
			Formação dos Grupos Temáticos por meio de AA (AHA, distância Euclidiana, critério Ward) sobre **T(AA2)**	**Formação de Subgrupos Especializados** por meio de AA de Objetos (AHA, Distância Eucidiana, critério Ward) sobre **T(UU2)**
Etapa 4 Formação das Listas-Base		Consolidação das Transações por SGE		Consolidação das Transações por SGE
Etapa 5 Criação das Listas de Recomendação		Eliminação das Transações Conhecidas pelo Usuário		Eliminação das Transações Conhecidas pelo Usuário

2. ANÁLISE DOS ASSUNTOS: CRIAÇÃO DOS GRUPOS TEMÁTICOS

2.1. Assuntos Significativos (AS) e Grupos Temáticos (GT)

Conforme o projeto anterior,

> Grupos Temáticos são características dos itens da biblioteca. Sua definição procura refletir o campo de conhecimento abrangido pelo item. A identificação dos GTs **parte** da Classificação Decimal Universal (IBICT, 1987; SILVA, 1994), mas é conduzida de forma a refletir a maneira como, **na prática, os pesquisadores agrupam diversos assuntos em grandes áreas de pesquisa** [...]. Segundo esta abordagem, não é o bibliotecário quem define o Grupo Temático de um livro, mas os próprios pesquisadores, pelo uso que fazem do item.
>
> Assim, por exemplo, se itens relativos à ciência política são utilizados por pesquisadores de Administração Pública, estes item são filiados ao GT de Administração Pública e não a um GT de Ciências Sociais, como aconteceria se fosse adotado integralmente o procedimento normativo indicado pela CDU. (ARANHA, em editoração).

Há necessidade de se definirem os GTs por dois motivos complementares e convergentes:

- De um lado, está a necessidade de redução da diversidade dos itens, que são muito numerosos (cerca de 60.000), principalmente quando comparados ao número de transações analisadas (cerca de 22.500).

Cada transação caracteriza uma ligação de um usuário a um item. A ligação entre dois usuários decorre do fato deles terem retirado o mesmo item; analogamente a ligação entre dois itens decorre do fato de terem sido retirados pelo mesmo usuário. Se cada item for considerado um nó distinto, haverá raros itens ligados a mais de um usuário, e, em conseqüência a matriz que representa a rede será demasiado esparsa para a formação de grupos.

Consolidando-se os "escaninhos temáticos" do CDU até que os ramos contenham quantidades aproximadamente iguais de itens (veja ARANHA, 2000) criam-se categorias de assuntos mais abrangentes, que contém um maior número de itens. Por meio deste artifício, pode-se enxergar a rede em um nível mais agregado, com apenas 42 nós relativos aos objetos sendo transacionados.

Observada a partir desta perspectiva mais resumida, a matriz que representa a rede, em geral, e a partição que representará a associação entre os assuntos, em particular, tornam-se mais densas. Com assuntos e usuários mais interligados, a formação de grupos de usuários torna-se mais significativa.

A criação dos GTs leva esta estratégia ainda mais longe, reduzindo os 42 Assuntos Significativos iniciais a 12 Grupos Temáticos.

- De outro lado, estes grandes grupos de itens servem para identificar as áreas gerais de pesquisa, e permitem separar os usuários por seus interesses principais, facilitando a interpretação do foco inferido e possibilitando uma redução de dimensionalidade do problema quando se passa a agrupar os usuários (veja item III.3).

Como refletem a forma pela qual os usuários combinam itens ao realizar suas pesquisas, a configuração dos GTs altera-se ao longo do tempo, e deve ser periodicamente atualizada.

2.2. Organização do banco de dados na matriz D

Neste projeto retomamos o banco de dados utilizado no projeto anterior (veja item II.2), contendo informações sobre 22.500 transações de empréstimos ("TRANSAÇÕES") realizadas entre 01/03 e 19/07/1999; e sobre 410 usuários selecionados entre professores e alunos de pós-graduação ("USUÁRIOS"). Cada transação teve identificado o Assunto Significativo ("AS" ou "ASSUNTO") ao qual pertencia.

A Tabela 2, a seguir, lista os Assuntos Significativos (ASs) tal como foram apresentados em ARANHA (2000) e tal como recodificados para uso com o Software UCINET (BORGATTI, EVERETT e FREEMAN, 1999; item 2.2.c).

Tabela 2

Assuntos Significativos dos Itens Estudados

Código Ucinet	Código Artigo	Descrição do Assunto
1	0	Generalidades. Ciência e Conhecimento. Organização
2	1	Filosofia, Psicologia
3	2	Religião, Teologia
4	3	Ciências Sociais, Direito, Administração, etc.
5	30	Metodologia
6	31	Demografia, Sociologia
7	32	Política
8	33	Economia
9	330.1	Teoria Econômica, Conceitos de Economia
10	330.3	Dinâmica Econômica
11	331	Trabalho, Emprego, Economia do Trabalho
12	334	Cooperativismo
13	336	Finanças Públicas
14	339	Comércio Internacional
15	34	Direito, Jurisprudência
16	35	Administração Pública, Governo, Assuntos Militares
17	36	Assistência, Previdência e Seguridade Social
18	37	Educação, Ensino
19	39	Antropologia
20	5	Matemática, Ciências Naturais
21	6	Ciências Aplicadas, Medicina, Tecnologia
22	61	Ciências Médicas
23	62	Engenharia, Tecnologia em Geral
24	63	Agricultura
25	65	Organização e Administração
26	651	Escritório
27	654	Telecomunicações e Telecontrole
28	657	Contabilidade
29	658	Administração de Empresas, Organização Comercial
30	658.0	Administração
31	658.1	Finanças
32	658.3	Pessoal, Fator Humano, RH
33	658.5	Administração da Produção
34	658.6	Comércio
35	658.7	Administração de Materiais
36	658.8	Marketing, Vendas e Distribuição
37	659	Publicidade e Propaganda, Relações Públicas
38	68	Indústria
39	7	Arte, Esportes
40	8	Língua, Literatura
41	9	Geografia, Biologia, História
42	R	Referência

A Tabela 3 apresenta as freqüências dos doze primeiros ASs (dentre os 42) e dos doze primeiros USUÁRIOS (dentre os 410) que mais aparecem no banco de dados.

Tabela 3

Ocorrências de Assuntos Significativos e Usuários mais Freqüentes

Assuntos	Freqüência
65	110
658.8	100
33	93
5	76
0	71
336	65
330.1 e 658	56
1 e 658.3	49
31 e 339	47

Usuários	Freqüência
98	15
154	14
167, 198 e 297	12
77, 181, 193, 276, 286, 316 e 438	10

2.2.a. Definição da matriz **D**

A partir dos dados originais, foi montada uma matriz de 452 linhas e colunas: 42 linhas (colunas) representando os ASSUNTOS e 410 linhas (colunas) representando os USUÁRIOS. No corpo desta matriz foi registrado **o número de vezes que um determinado USUÁRIO tomou emprestado um determinado item de um AS**. A partir de agora, esta matriz será chamada de matriz dos dados ou matriz **D**. Trata-se de uma matriz de co-ocorrências (veja item II.5.4).

Figura 6

Matriz D de Relacionamentos entre USUÁRIOS e ASSUNTOS

$$\mathbf{D} = \left[\begin{array}{ccc:ccc} d_{1,1} & \ldots & d_{1,42} & d_{1,43} & \ldots & d_{1,452} \\ \ldots & & \ldots & \ldots & & \ldots \\ d_{42,1} & \ldots & d_{42,42} & d_{42,43} & \ldots & d_{42,452} \\ \hdashline d_{43,1} & \ldots & d_{43,42} & d_{43,43} & \ldots & d_{43,452} \\ \ldots & & \ldots & \ldots & & \ldots \\ d_{452,1} & \ldots & d_{452,42} & d_{452,43} & \ldots & d_{452,452} \end{array}\right] = \left[\begin{array}{c:c} \mathbf{AA} & \mathbf{AU} \\ \hdashline \mathbf{UA} & \mathbf{UU} \end{array}\right]$$

2.2.b. Partições da matriz D

A matriz **D** será particionada em quatro:

- **AA, representando o cruzamento entre linhas e colunas dos 42 assuntos**; esta partição é formada exclusivamente por zeros, pois, representaria as relações entre assuntos e, como dissemos, no corpo da matriz **D** apenas está representado o número de vezes que um USUÁRIO transacionou um AS;
- **UU, representando o cruzamento entre as linhas e colunas dos 410 usuários**; esta partição representaria a relação entre USUÁRIOS; é exclusivamente formada por zeros, pelo mesmo motivo que a partição **AA** também o é;
- **UA representando o cruzamento das linhas dos usuários com as colunas dos assuntos** e **AU** representando as linhas dos assuntos com as colunas dos usuários; $\mathbf{UA = (AU)^t}$.

2.2.c. Carga da matriz D

Existem 4 tipos de arquivos que podem servir de entrada de dados para o UCINET 5.0:

- ***Raw*** : arquivos compostos apenas por números, correspondem a uma matriz que guarda variáveis numericamente codificadas;

- **DL**: arquivos do tipo "*data language*", que contém dados do tipo *raw* mais uma série de informações que descrevem os dados, tais como número de linhas e colunas da matriz, nomes das variáveis, etc; a sintaxe e o vocabulário da "*data language"* estão descritos brevemente no manual do UCINET;

- **UCINET 3.0**: arquivos gerados em versão anterior do UCINET; são similares aos arquivos do tipo DL, com a desvantagem de possuírem recursos mais limitados;

- **EXCEL**: planilhas de arquivos de dados padrão do Microsoft Excel.

No projeto da Biblioteca KAB, em função do formato original do banco de dados, o tipo de arquivo mais adequado é o DL no formato de "lista de ligações" denominado **edgelist1**. Este formato especifica dados que correspondem a uma matriz de ligações individuais e a magnitude de cada ligação.

Para carregar o arquivo do tipo DL no software UCINET 5.0 foi utilizado o seguinte procedimento:

```
Data > Import > DL > Input file: nomedoarquivo.txt
```

A figura a seguir mostra a estrutura do arquivo de dados de transações da biblioteca, codificado para *Data Language* e pronto para ser lido no UCINET:

Tabela 4

Dados da Biblioteca Preparados para Leitura no UCINET

```
dl n=452 format=edgelist1
labels embedded
data:
001     045    21
001     046    22
 ...     ...    ...
041     438    3
042     075    1
042     077    2
```

Como se vê, o arquivo acima é composto por um cabeçalho, que contém informações sobre os dados, e por uma listagem formada por 3 colunas numéricas.

No cabeçalho observa-se a seguinte estrutura:

- **dl**: identifica que o arquivo é do tipo Data Language;
- **n=452**: indica a dimensão da matriz (número de assuntos mais número de usuários);
- **format edgelist1**: identifica o formato de lista de ligações do arquivo DL;
- **labels embedded**: informa que os rótulos para as variáveis estão embutidos no próprio conjunto de dados, isto é, correspondem ao próprio valor das realizações das duas primeiras colunas;
- **Data**: indica que deste ponto em diante o arquivo contém o conjunto de dados propriamente dito.

No conjunto de dados são observadas 3 colunas.

- A 1ª coluna: contém valores de 1 a 42, representando os 42 diferentes Assuntos Significativos dos livros;
- A 2ª coluna: contém valores de 43 a 452, representando os usuários da biblioteca; e
- A 3ª coluna: indica a número de itens do assunto informado na 1ª coluna retirados pelo usuário caracterizado na 2ª coluna.

Observe, por exemplo, a 4ª linha do arquivo da Tabela 4: "001 045 21". Esta linha informa que o usuário identificado pelo número 45 retirou da biblioteca 21 livros que tratam do assunto identificado pelo número 1.

Ao carregar o arquivo de dados, o software UCINET 5.0 armazena todas as informações lidas em uma única matriz de dados. A estrutura desta matriz de dados (**D**), com alguns dados já carregados, está representada a seguir:

Tabela 5

Dados Lidos na Matriz D

		1	2	3		40	41	42	43	44	45	...	450	451	452
		001	002	003	...	040	041	042	043	044	045	...	450	451	452
1	001	0	0	0	...	0	0	0	0	0	4	...	0	0	0
2	002	0	0	0	...	0	0	0	0	6	0	...	0	3	0
3	003	0	0	0	...	0	0	0	3	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
40	040	0	0	0	...	2	0	1	0	0	0	...	23	0	0
41	041	0	0	0	...	0	7	5	0	0	0	...	0	12	0
42	042	0	0	0	...	0	0	0	5	0	0	...	0	0	1
43	043	0	0	0	...	0	0	0	0	0	0	...	0	0	0
44	044	0	0	0	...	0	0	0	0	0	0	...	0	0	0
45	045	0	0	0	...	0	0	0	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
450	450	0	0	0	...	0	0	0	0	0	0	...	0	0	0
451	451	0	0	0	...	0	0	0	0	0	0	...	0	0	0
452	452	0	0	0	...	0	0	0	0	0	0	...	0	0	0

A matriz acima é quadrada triangular superior e contém 452 linhas e colunas. Nos rótulos atribuídos às linhas e colunas são observados os números de 1 a 42 identificando os assuntos (1ª coluna do arquivo DL) seguidos pelos números de 43 a 452 correspondentes aos usuários (2ª coluna do arquivo DL). O número de transações entre livros e pessoas (3ª coluna do arquivo DL) encontra-se no corpo da matriz: cada célula corresponde a uma ligação que pode ser do tipo assunto-assunto, assunto-usuário, usuário-assunto e usuário-usuário.

Note que esta matriz deveria ser simétrica pois se, por exemplo, o usuário 45 retirou 4 livros referentes ao assunto 1, 4 livros do assunto 1 foram retirados pelo usuário 45. Como isso não acontece automaticamente após a leitura dos dados, isto é, se alimentadas as transações entre usuários e assuntos apenas, a relação inversa não é anotada também. Por isso, é necessário simetrizar a matriz. Isto pode ser feito pela aplicação do procedimento:

```
Transform > Symmetrize > Input > Dataset: nomedoarquivo.##h
```

A matriz de dados simetrizada pode então ser particionada nas quatro submatrizes **AA**, **UU**, **AU** e **UA**, definidas anteriormente.

2.2.d. Particionamento da matriz **D**

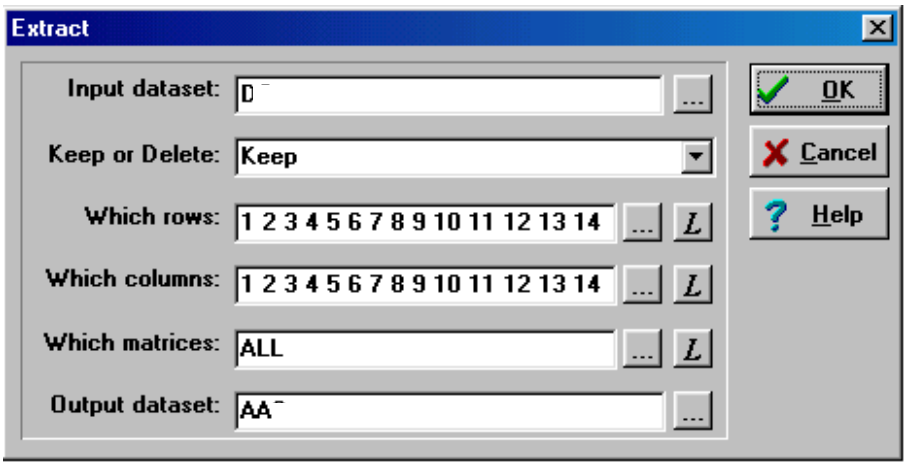
Para selecionar apenas uma partição da matriz **D**, digamos a partição **AA**, deve ser utilizado o seguinte procedimento do UCINET 5.0:

```
Data > Extract
```

Esta seqüência de comandos invocará a janela representada na Figura 7 a seguir:

Figura 7

Janela de Diálogo de Extração de Partições



que deve ser preenchida da seguinte maneira:

- **Input dataset**: Neste campo deve ser informado o nome da matriz da qual se pretende extrair uma partição, no caso, a matriz **D**. Essa matriz pode ser procurada clicando-se no botão  , que está logo em frente ao campo;

- **Keep or Delete**: Neste campo informamos se os dados selecionados devem ser mantidos ou apagados; no caso da obtenção da partição, selecione *Keep*;

- **Which rows**: Neste item devem ser registradas as linhas a incluir na partição desejada; no caso são as linhas 1 a 42 (referentes aos assuntos). Para selecioná-las de maneira prática, clique no botão  e na nova janela selecione as linhas de interesse clicando em OK para finalizar a tarefa;

- **Wich columns**: Neste campo, informam-se as colunas de interesse, adotando-se o mesmo procedimento empregado no item anterior;

- **Which matrices**: O "default" deste item (*all*) deve ser mantido; e

- **Output dataset**: Neste campo nomeia-se a partição a ser extraída.

O resultado da aplicação conveniente do procedimento acima, uma vez para cada submatriz desejada, são as 4 partições definidas anteriormente:

- A matriz **AA** (42 x 42) corresponde à quantidade de transações entre assunto e assunto. Como já vimos, essa matriz é composta somente de zeros;

Tabela 6

Matriz AA Lida no UCINET

AA =

		1	2	3		40	41	42
		001	002	003	...	040	041	042
1	001	0	0	0	...	0	0	0
2	002	0	0	0	...	0	0	0
3	003	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...
40	040	0	0	0	...	0	0	0
41	041	0	0	0	...	0	0	0
42	042	0	0	0	...	0	0	0

- A matriz **AU**(42 x 410), em que as 42 linhas indicam os assuntos dos livros e as 410 colunas os usuários da biblioteca; corresponde ao número de transações entre a variável assunto e a variável usuário;

Tabela 7

Matriz AU Lida no UCINET

AU =

		43	44	45	...	450	451	452
1	001	0	0	4	...	0	0	0
2	002	0	6	0	...	0	3	0
...	...	...	...	...	...	...	...	...
40	040	0	0	0	...	23	0	0
41	041	0	0	0	...	0	12	0
42	042	5	0	0	...	0	0	1

- A matriz **UA**(410 x 42) correspondente a matriz transposta da matriz **AU**; nesta matriz, os assuntos dos livros encontram-se nas colunas e os usuários correspondem às linhas, ou seja, a matriz contém transações entre usuários e assuntos;

Tabela 8

Matriz UA Lida no UCINET

		1	2	3		40	41	42
		001	002	003	...	040	041	042
43	043	0	0	0	...	0	0	0
44	044	0	0	0	...	0	7	0
...	...	...	...	...	...	...	...	...
450	450	0	0	0	...	23	0	0
451	451	0	0	0	...	0	12	0
452	452	0	0	0	...	0	0	1

UA =

- A matriz **UU**(410 x 410) corresponde à quantidade de transações entre usuários e usuários, ou seja, como já vimos, também esta matriz contém apenas zeros.

Tabela 9

Matriz UU Lida no UCINET

		43	44	45		450	451	452
43	043	0	0	0	...	0	0	0
44	044	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...
450	450	0	0	0	...	0	0	0
451	451	0	0	0	...	0	0	0
452	452	0	0	0	...	0	0	0

UU =

2.3. Matriz D2

Definimos ainda a matriz **D2** como sendo o quadrado da matriz **D.**

2.3.a. Partições de D2

Particionada da mesma forma que a matriz **D**, **D2** produz quatro submatrizes: **AA2**, **UU2**, **UA2** e **AU2** (Figura 8)

Figura 8

Matriz D2 e suas Partições

$$\mathbf{D}2 = \left[\begin{array}{ccc:ccc} d_{1,1} & \ldots & d_{1,42} & d_{1,43} & \ldots & d_{1,452} \\ \ldots & & \ldots & \ldots & & \ldots \\ d_{42,1} & \ldots & d_{42,42} & d_{42,43} & \ldots & d_{42,452} \\ \hdashline d_{43,1} & \ldots & d_{43,42} & d_{43,43} & \ldots & d_{43,452} \\ \ldots & & \ldots & \ldots & & \ldots \\ d_{452,1} & \ldots & d_{452,42} & d_{452,43} & \ldots & d_{452,452} \end{array}\right] = \left[\begin{array}{c:c} \mathbf{AA}2 & \mathbf{AU}2 \\ \hdashline \mathbf{UA}2 & \mathbf{UU}2 \end{array}\right]$$

Note que:

- **AA2 não é** o quadrado de **AA** (o quadrado de **AA** seria **0**)**,** mas uma partição de **D2**; da mesma forma que **UU2 não é** o quadrado de **UU** (o quadrado de **UU** também seria **0**);
- Na matriz **D2**, as partições **UA** e **AU** são formadas exclusivamente por zeros, indicando que agora não há relação entre USUÁRIOS e ASSUNTOS;
- **AA2** indica os assuntos que se relacionam por terem sido tomados pelos mesmos usuários, bem como a intensidade desta relação; sua diagonal representa uma medida de **atividade total de cada assunto**, pois corresponde à soma de

quadrados das linhas da matriz **D**; os valores fora da diagonal representam o grau de **atividade conjunta de dois assuntos distintos** e corresponde à soma de seus produtos cruzados;

- **UU2** indica os USUÁRIOS que se relacionam por terem transacionado os mesmos ASs, bem como a intensidade desta relação; sua diagonal representa uma medida de atividade total de cada usuário, pois corresponde à soma dos quadrados das linhas da matriz **D**; os valores fora da diagonal representam uma medida de atividade conjunta de dois usuários distintos e corresponde a uma soma de produtos cruzados.

2.3.b. Cálculo de **D2** a partir de **D**

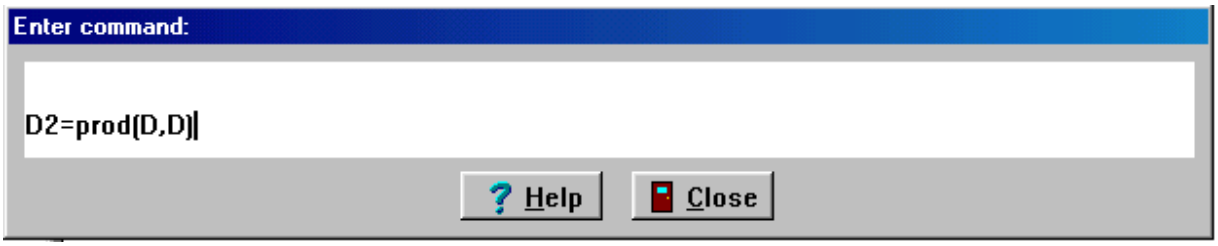
Com auxílio do UCINET, **D2** é calculada a partir da matriz D utilizando-se o procedimento:

```
Tools > Matrix Algebra
```

que invoca a janela de diálogo da Figura 9 a seguir; esta deve ser preenchida como indicado.

Figura 9

Janela de Comandos para Álgebra Matricial no UCINET



Pressionando-se, em seguida, a tecla <enter> obtém-se a matriz D2 desejada. Para fechar a janela de Álgebra Matricial deve-se utilizar o botão 

.

2.3.c. Particionamento de D2

O particionamento de **D2** faz-se de forma análoga ao procedimento usado para particionar **D** (veja item 2.2.d).

2.4. Matrizes de Distâncias Temáticas Dicotômicas

Para a identificação dos GTs por meio de Análise de Rede, inicialmente aplica-se a definição de Distância Temática Dicotômica a cada par de assuntos da partição **AA2**. Os resultados do cálculo da Distância Temática Dicotômica entre os pares de Assuntos Significativos da matriz **AA2** são organizados na matriz **T** definida como:

$$\mathbf{T} = t_{ij} = \begin{cases} \dfrac{\left|[C(AS_i) \cup C(AS_j)] - [C(AS_i) \cap C(AS_j)]\right|}{[C(AS_i) \cup C(AS_j)]}, & C(AS_i) \cup C(AS_j) \neq 0 \\ 1, & C(AS_i) \cup C(AS_j) = 0 \end{cases}$$

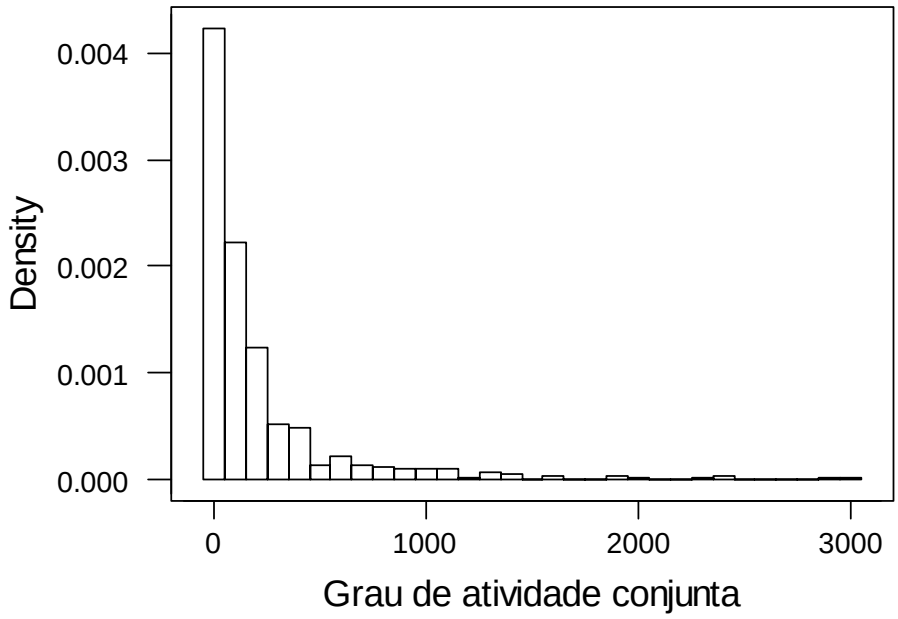
2.4.a. Matriz filtrada **AA2(p)**

Note que tanto relacionamentos “fortes” quanto relacionamentos “fracos” entre Assuntos Significativos serão considerados indicadores de vizinhança por ocasião do cálculo da matriz **T,** já que apenas o **número** de conexões entra na fórmula da Distância Temática Dicotômica (e não sua intensidade).

Note ainda, examinando o histograma da matriz subdiagonal de **AA2** (Figura 10), que há uma grande dispersão na intensidade das ligações entre assuntos, predominando as ligações fracas.

Figura 10

Histograma das Ligações de AA2



Assim, faz sentido considerar um procedimento anterior à dicotomização, que elimine as ligações fracas, possivelmente produzindo um resultado final mais nítido. Eliminar as ligações fracas seria, nesta caso, análogo a, num aparelho de TV, selecionar um conveniente nível de contraste para a imagem exibida na tela. Se "p" é a proporção das ligações mais fracas de **AA2** que desejamos "filtrar" (isto é, transformar em 0), podemos definir **AA2** filtrada em (100*p) % como sendo

$$\mathbf{AA2(p)} = \begin{cases} aa2_{ij}, & \text{se } aa2_{ij} > aa2_{(p)} \\ 0, & \text{se } aa2_{ij} \leq aa2_{(p)} \end{cases}$$

onde

$aa2_{ij}$ é o elemento da linha i, coluna j, de **AA2**;

$aa2_{(p)}$ é o percentil 100*p das ligações da matriz subdiagonal de **AA2**.

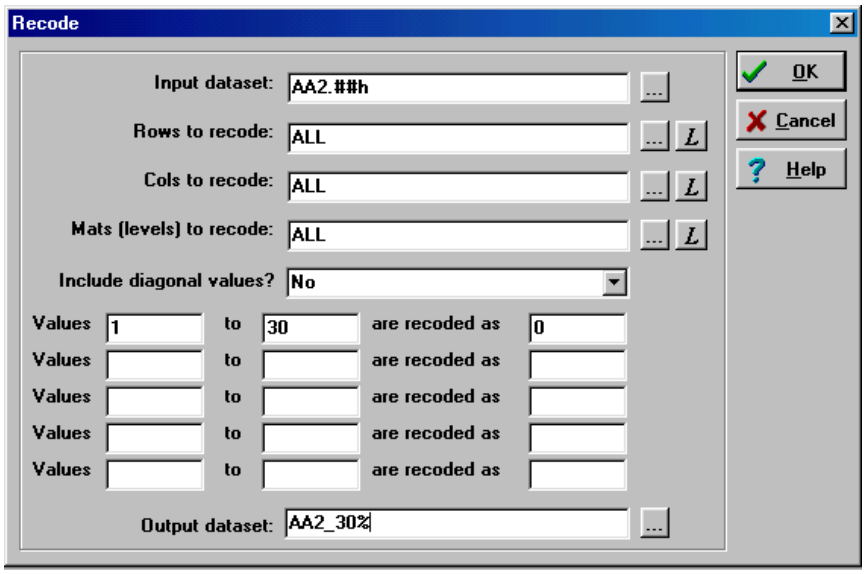
As matrizes filtradas **AA2(p)** podem ser obtidas no UCINET através do procedimento a seguir. O comando

```
Transform > Recode ou pelo comando Ctrl+Alt+R
```

invoca a janela de diálogo representada na Figura 11.

Figura 11

Janela de Dialógo para Recodificação no UCINET



Esta janela de diálogo deve ser preenchida com segue.

- **Input dataset**: Neste campo deve ser informado o nome da matriz cujos valores serão recodificados, no caso ***AA2***;
- **Include diagonal values?**: Deve ser selecionada a opção *No* para que não seja aplicada a nota de corte na diagonal;
- **Values ... to ...** : Nestes campos devem ser informados os valores mínimos e máximos (nota de corte) a serem recodificados, no exemplo são, respectivamente, *1* e *30*;
- **Are recoded as**: Neste campo deve ser informado o valor que substituirá as observações contidas na faixa especificada na mesma linha, no caso *zero*; e

- **Output dataset**: Neste campo deve ser informado o nome da matriz a receber o resultado da recodificação, no exemplo, ***AA2_30%***.

Por uma questão de unificação de notação, considera-se que a matriz original **AA2** equivale à matriz **AA2(0)**, isto é, à matriz **AA2** em que 0% das ligações mais fracas foram filtradas.

Os percentis das ligações da matriz subdiagonal de **AA2** estão representados na Tabela 10. O histograma da Figura 12 representa a matriz filtrada **AA2(0,30)**. Comparando-o com o histograma da Figura 10, observa-se a redução da densidade de ligações fracas.

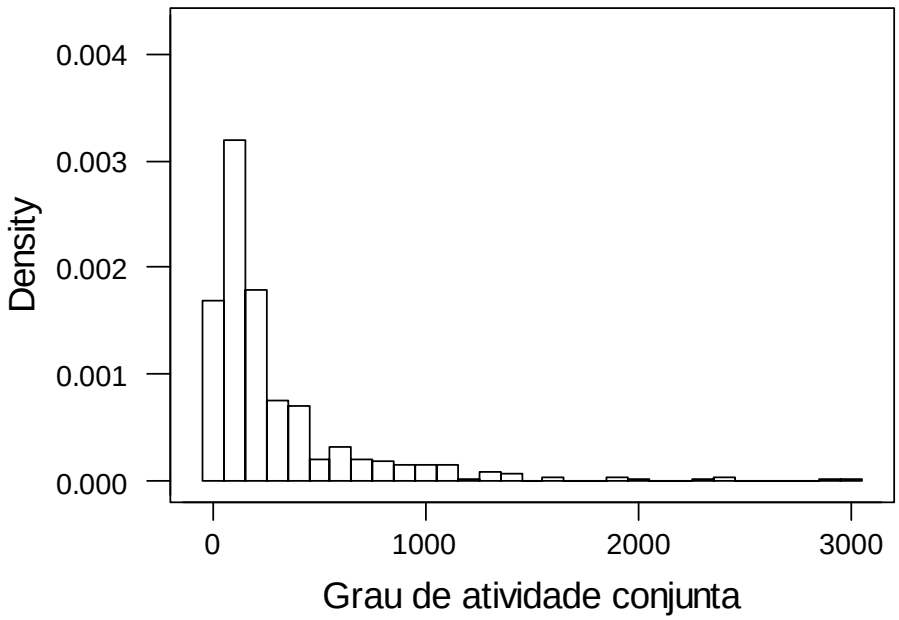
Tabela 10

Percentis do Grau de Atividade Conjunta da Matriz Subdiagonal de AA2

Percentil	10%	20%	30%	40%	50%	60%	70%	80%	90%
Grau	6	18	30	44	70	114	180	291	585

Figura 12

Histograma das Ligações da Matriz Subdiagonal de AA2(0,30)



2.4.b. Matriz dicotomizada **AA2(p)***

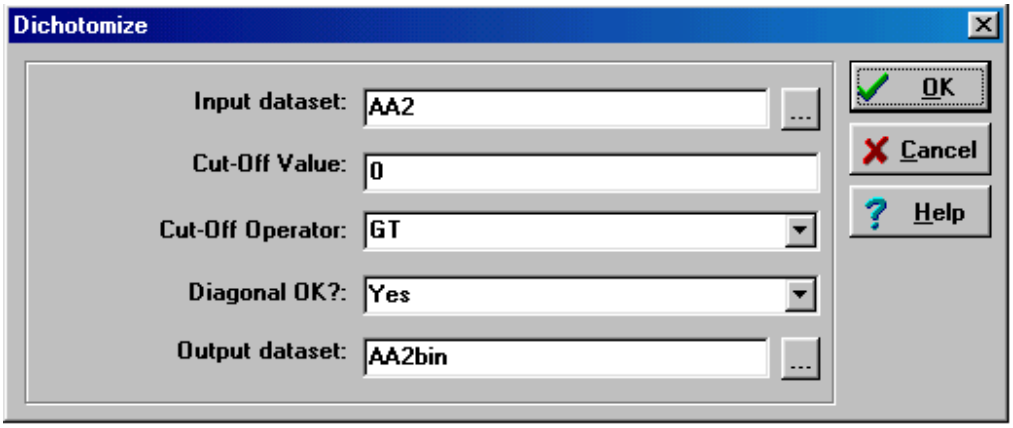
Como a Distância Temática Dicotômica não leva em consideração a intensidade do relacionamento entre dois Assuntos Significativos, é conveniente dicotomizar a matriz **AA2** para simplificar os cálculos de **T**. Chamaremos de **AA2(p)*** ao resultado da dicotomização da matriz **AA2(p)**.

A dicotomização de uma matriz pode ser feita no UCINET invocando-se a janela de diálogo apresentada na Figura 13, a seguir, através do comando:

```
Transform > Dichotomize
```

Figura 13

Janela de Diálogo para Dicotomização no UCINET



A janela de diálogo deve ser preenchida como segue.

- **Input dataset**: Neste campo deve ser informado o nome da matriz que se pretende transformar em binária, no caso **AA2**;
- **Cut-Off Value:** Neste campo deve ser informado o valor **zero**;
- **Cut-Off Operator**: Deve ser escolhida a função desejada, no caso **GT (Greater Than)**; a combinação desta informação com a do campo anterior resulta em que valores maiores que zero serão recodificados para 1; e, finalmente, em
- **Output dataset**: Deve ser informado o nome da matriz de saída do resultado, no exemplo, **AA2bin**.

2.4.c. Matriz **T(p)**

Seja a matriz **T(p)** a matriz T de Distância Temática calculada sobre **AA2(p)***. Usando-se álgebra matricial, a matriz **T(p)** pode ser calculada facilmente através da expressão

$$\mathbf{T(p)} = \frac{\mathbf{1}.\mathbf{diag}^t[(\mathbf{AA2(p)}^*)^2] + \mathbf{diag}[(\mathbf{AA2(p)}^*)^2].\mathbf{1}^t - 2.(\mathbf{AA2(p)}^*)^2}{\mathbf{1}.\mathbf{diag}^t[(\mathbf{AA2(p)}^*)^2] + \mathbf{diag}[(\mathbf{AA2(p)}^*)^2].\mathbf{1}^t - (\mathbf{AA2(p)}^*)^2}$$

onde

1	é um vetor, de dimensão conveniente, formado pelo valor 1 em todas as células;
diag[.]	é um vetor contendo a diagonal de uma matriz quadrada que entra como argumento da função;
t sobrescrito	indica transposição do vetor ou matriz;
$(AA2(p)^*)^2$	é o quadrado da matriz AA2 filtrada em 100p % das ligações mais fracas;
.	indica a multiplicação de matrizes.

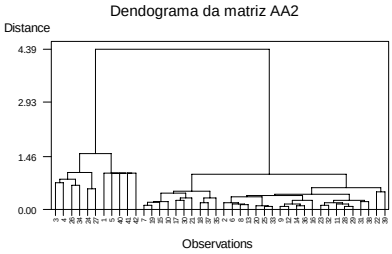
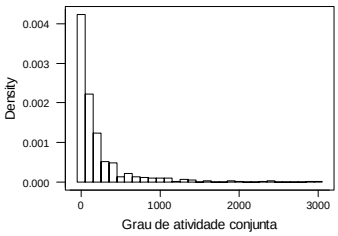
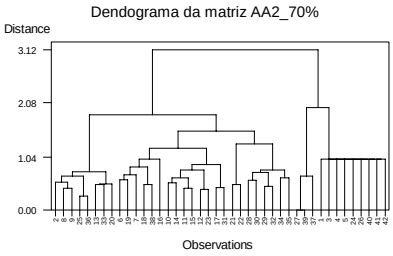
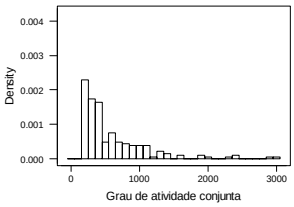
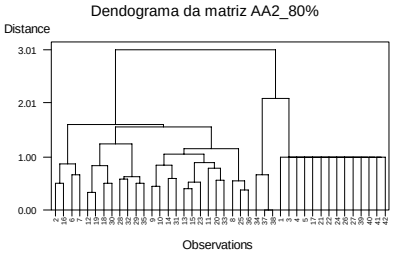
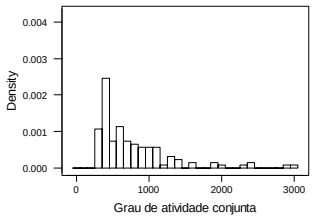
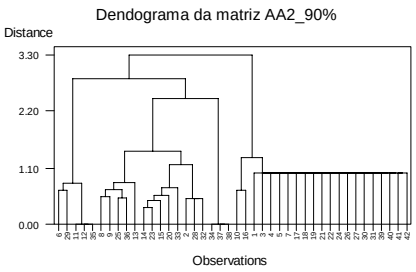
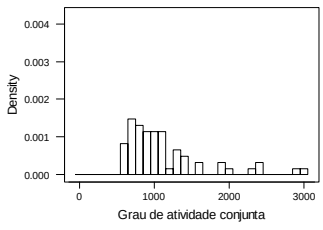
2.5. Reunião de Assuntos Significativos em Grupos Temáticos

Sobre a matriz de Distância Temática Dicotômica **T(p)**, aplica-se um algoritmo hierárquico aglomerativo por distância euclidiana e critério Ward (veja, por exemplo, HAIR, ANDERSON, TATHAM e BLACK, 1995).

Na análise a seguir, fizemos p assumir os valores 0; 0,60; 0,70; 0,80 e 0,90. Os resultados dos dendrogramas e histogramas estão apresentados na Figura 14.

Figura 14

Resultado da Análise de Agrupamentos sobre T(p) em Vários Níveis de p

Matriz	Dendrograma	Histograma
AA2(0)		
AA2(0,7)		
AA2(0,8)		
AA2(0,9)		

A análise dos dendrogramas para as matrizes T(p) levou à formação de Grupos Temáticos diferentes, de um nível de filtragem para outro, tanto em número de grupos como no que diz respeito aos Assuntos Significativos atribuídos a cada grupo.

2.6. Comparação dos GT[T(p)]

A Tabela 11, a seguir, indica, em cada nível de filtragem, o número *n* de GTs formados e a Distância Temática média, geral e de cada grupo, estas ordenadas crescentemente.

A diminuição da Distância Temática geral à medida em que o nível de filtragem aumenta é uma conseqüência de um maior número de zeros na matriz AA2. O que é digno de nota, no entanto, é o fato desta medida diminuir pouco, isto é, de mostrar-se bastante robusta ao nível de filtragem, sugerindo que os grupos formados mantém-se parecidos, de um nível de filtragem para outro. Esta estabilidade já se refletira na semelhança dos dendrogramas da Figura 14.

Tabela 11

Coerência Interna dos Grupos Temáticos Formados em Cada Nível de Filtragem

	n	médias										
GT(0%)	10	Geral 0.86	GT5 0.74	GT7 0.77	GT9 0.78	GT6 0.81	GT8 0.86	GT3 0.90	GT4 0.90	GT10 0.93	GT1 0.94	GT2 0.94
GT(60%)	10	Geral 0.79	GT1 0.59	GT4 0.63	GT7 0.73	GT10 0.78	GT3 0.82	GT9 0.82	GT6 0.84	GT8 0.84	GT5 0.92	GT2 0.93
GT(70%)	10	Geral 0.79	GT5 0.64	GT10 0.64	GT6 0.68	GT3 0.78	GT7 0.78	GT1 0.81	GT4 0.84	GT2 0.86	GT8 0.93	GT9 0.95
GT(80%)	8	Geral 0.75	GT7 0.48	GT3 0.61	GT1 0.69	GT4 0.77	GT6 0.81	GT2 0.82	GT5 0.87	GT8 0.92		
GT(90%)	8	Geral 0.74	GT6 0.48	GT5 0.58	GT1 0.75	GT7 0.76	GT2 0.77	GT3 0.82	GT4 0.85	GT8 0.92		

Uma outra forma de investigar a semelhança dos GTs formados em diversos níveis de filtragem é comparar os resultados de cada nível com um mesmo padrão; este foi definido, arbitrariamente, como sendo o resultado da análise temática realizada no projeto anterior, que passa a ser chamado de GT(A).

Por exemplo, na Tabela 12, o GT(60%) é comparado com o GT(A). A semelhança entre os dois agrupamentos seria máxima se cada linha da matriz tivesse todas as ocorrências em um única casela; e se o mesmo ocorresse em cada coluna. Isto indicaria que cada Grupo Temático de GT(60%) corresponderia exatamente a um único Grupo Temático de GT(A).

Tabela 12

Comparação de GT(60%) e GT(A)

GT (A)	GT (60%)										Total Global
	1	2	3	4	5	6	7	8	9	10	
1				1		1					2
2		2	1		5			1		2	11
3						1	2		1		4
4	3				1						4
5				2						1	3
6		1	1	1							3
7					1	1		2	1		5
8			2		2						4
9			1			1			1		3
10						1	1			1	3
Total Global	3	3	5	4	9	5	3	3	3	4	42

Uma medida resumo da similaridade entre as classificações pode ser definida como:

$$Sim = \sqrt{\frac{\sum_{i=1}^{l} máx}{n} \frac{\sum_{i=1}^{c} máx}{n}}$$

onde l = número de linhas da matriz, c = número de colunas da matriz e n = número total de assuntos significativos (42). Essa medida pode assumir valores entre zero e um; quanto mais próxima de 1 ela estiver, maior a similaridade entre os agrupamentos.

Na Tabela 13, temos a medida de similaridade (Sim) entre o agrupamento do artigo e os encontrados neste trabalho.

Tabela 13

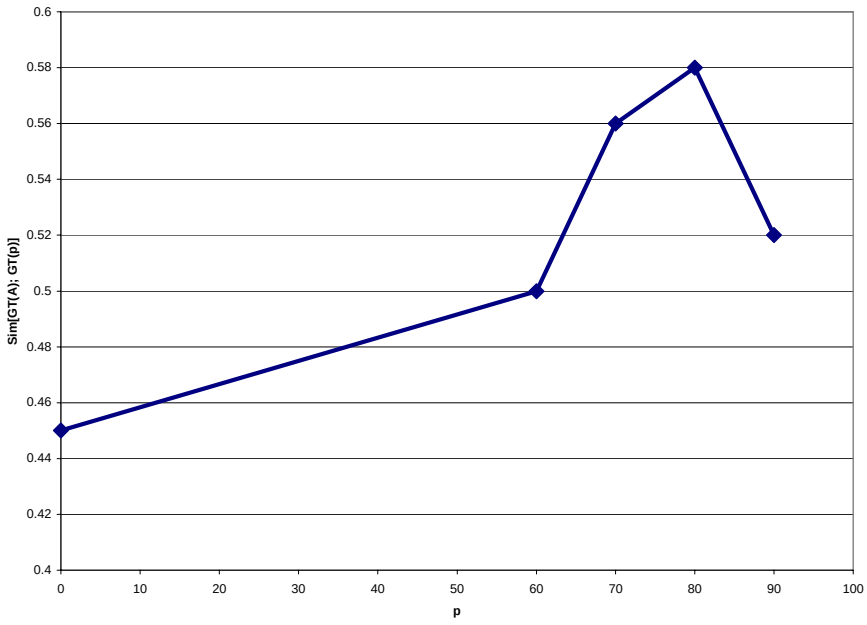
Similaridade entre GT(p) e GT(A)

GT(A) versus	**Sim**
GT(0%)	0.45
GT(60%)	0.50
GT(70%)	0.56
GT(80%)	0.58
GT(90%)	0.52

Estes valores estão representados no gráfico a seguir (Figura 15).

Figura 15

Semelhança dos GT(p)s com os GT(A)



Nota-se o nível de filtragem p = 80% é o que produz resultados mais semelhantes aos obtidos no projeto anterior. Observa-se também que a medida de semelhança fica na faixa de 0,45 a 0,68, um intervalo razoavelmente restrito, de forma que se confirma a interpretação de que os resultados são todos semelhantes entre si.

2.7. Solução com Distância Temática Generalizada

Conforme definição em II.5.7, ocorreu-nos que o conceito de Distância Temática proposto por SCHWARTZ e WOOD (1993), poderia ser adaptado e aplicado diretamente sobre a matriz **AA2**, sem que houvesse prejuízo à sua interpretabilidade. Ao contrário, evitando-se a dicotomização, aproveitar-se-ia melhor a informação existente sobre os relacionamentos entre os assuntos.

Chamamos de GT(T) aos Grupos Temáticos formados a partir do cálculo da Distância Temática diretamente sobre AA2. Os resultados de GT(T) são comparados na Tabela 14, a seguir, com os resultados do projeto anterior, GT(A) (ARANHA, 2000).

Tabela 14

Comparação dos Resultados de GT(A) e GT(T)

Código Ucinet	GT (A)	Descrição do GT (A)	GT (T)	Descrição do GT (T)
1	8	Educação/Antropologia	2	Administração Geral
2	7	RH	12	RH
3	2	Outros	9	Outros A
4	2	Outros	9	Outros A
5	4	Metod./Sociol./Política/Adm. Publica	8	Metodologia
6	4	Metod./Sociol./Política/Adm. Publica	8	Metodologia
7	4	Metod./Sociol./Política/Adm. Publica	8	Metodologia
8	3	Economia	5	Economia e Marketing
9	3	Economia	1	Administração de Produção
10	3	Economia	8	Metodologia
11	7	RH	12	RH
12	9	Cooperativa/Agricultura	4	Contabilidade
13	10	Finanças	6	Finanças
14	3	Economia	6	Finanças
15	2	Outros	5	Economia e Marketing
16	4	Metod./Sociol./Política/Adm. Publica	8	Metodologia
17	2	Outros	9	Outros A
18	8	Educação/Antropologia	9	Outros A
19	8	Educação/Antropologia	9	Outros A
20	10	Finanças	2	Administração Geral
21	6	Adm. Hospitalar	3	Administração Hospitalar
22	6	Adm. Hospitalar	10	Outros B
23	2	Outros	1	Administração de Produção
24	9	Cooperativa/Agricultura	11	Outros C
25	7	RH	2	Administração Geral
26	2	Outros	10	Outros B
27	2	Outros	9	Outros A
28	5	Contabilidade/Materiais	4	Contabilidade
29	7	RH	4	Contabilidade
30	6	Adm. Hospitalar	3	Administração Hospitalar
31	10	Finanças	6	Finanças
32	7	RH	12	RH
33	2	Outros	1	Administração de Produção
34	5	Contabilidade/Materiais	7	Marketing
35	5	Contabilidade/Materiais	4	Contabilidade
36	1	Marketing	5	Economia e Marketing
37	1	Marketing	7	Marketing
38	9	Cooperativa/Agricultura	7	Marketing
39	2	Outros	9	Outros A
40	2	Outros	10	Outros B
41	2	Outros	9	Outros A
42	8	Educação/Antropologia	11	Outros C

O nível de similaridade entre GT(A) e GT(T) foi de 0,61, portanto superior ao resultado obtido quando o processo de agrupamento envolveu filtragem (compare com a Tabela 11). Esta maior similaridade entre GT(A) e GT(T) parece-nos razoável, uma vez que nestes dois procedimentos não houve filtragem de **AA2**.

O nível de coerência interna dos resultados de GT(T) medido pela média da distância temática entre os ASs foi de 0,70, portanto inferior ao de todos aos agrupamentos GT(p). Assim como na distância temática, quanto menor o valor da medida, mais coerente é o grupo.

Diante da maior simplicidade de cálculo e da maior coerência dos grupos formados, decidimos adotar nas fases seguintes de análise os resultados de GT(T), apresentados na Tabela 14.

3. CARACTERIZAÇÃO DOS USUÁRIOS: FORMAÇÃO DOS SUBGRUPOS ESPECIALIZADOS

Após agrupar os 42 assuntos significativos em 12 grupos temáticos, passa-se à identificação de **grupos de usuários** que retiram da biblioteca livros tematicamente semelhantes.

Os procedimentos para segmentação dos usuários são, em tudo, análogos aos utilizados para segmentação dos assuntos.

3.1. SubGrupos Especializados (SGE)

Os SubGrupos Especializados são usuários de um mesmo Grupo Temático, com um perfil de uso dos Assuntos Significativos muito semelhante, o que permite inferir um interesse comum entre os membros do SGE.

3.2. Matriz de dados redefinida

A análise realizada para identificação de SubGrupos Especializados é realizada em etapas, tomando-se os usuários de um Grupo Temático de cada vez. Para cada GT o procedimento a seguir descrito é repetido. Na descrição que se segue consideraremos apenas o GT de Marketing. Evidentemente, no entanto, para o projeto piloto foram computados os SGEs de todos os Grupos Temáticos.

Uma vez definido o GT a ser analisado (neste exemplo, o GT de Marketing), constrói-se uma nova matriz de dados, **D,** contendo todos ASs e somente os usuários que transacionaram no GT em foco.

Para carregar no UCINET 5.0 o arquivo correspondente à nova matriz **D**, utiliza-se o mesmo procedimento empregado na carga da matriz **D** completa original (veja item 2.2.c). Nos rótulos atribuídos às linhas e colunas são observados os números de 1 a 42, identificando os grupos temáticos, seguidos pelos números identificadores de usuários. O número de transações entre livros e usuários encontra-se, como antes, no corpo da matriz.

Esta matriz, a exemplo da antiga matriz **D**, deve ser simétrica. Para simetrizá-la utiliza-se o procedimento descrito em 2.2.c.

3.3. A matriz D2

O quadrado da nova matriz de dados, **D2**, é obtido através do procedimento

```
Tools > Matriz Algebra
```

utilizado no item 2.3, para a manipulação da **AA2**.

3.3.a. Particionando a matriz D2

Da nova matriz **D2,** convenientemente particionada, obtêm-se as seguintes sub-matrizes:

- **AA2** representa o cruzamento entre linhas e colunas dos 42 grupos temáticos. Nesta partição, os valores encontrados na diagonal representam uma medida de atividade total de cada grupo temático, pois cada um desses valores corresponde à soma dos quadrados das linhas da matriz **D;** fora da diagonal, os valores representam uma medida de atividade conjunta entre dois grupos temáticos distintos.

  Assim como foi ressaltado com relação à antiga matriz **AA2**, é importante lembrar que a nova partição **AA2** é diferente do quadrado da partição **AA**; o mesmo vale para as próximas partições.

- **UU2** representa o cruzamento entre linhas e colunas dos usuários ativos no GT em foco. Nesta partição, os valores encontrados na diagonal representam uma medida de atividade total de cada usuário, pois cada um desses valores corresponde à soma dos quadrados das linhas da matriz **D**; fora da diagonal, os valores representam uma medida de atividade conjunta entre dois usuários distintos.

- **AU2** representa o cruzamento das linhas referentes aos grupos temáticos com as colunas referentes aos usuários; essas partições contém valores zero em toda extensão; **UA2** é a transposta de **AU2**.

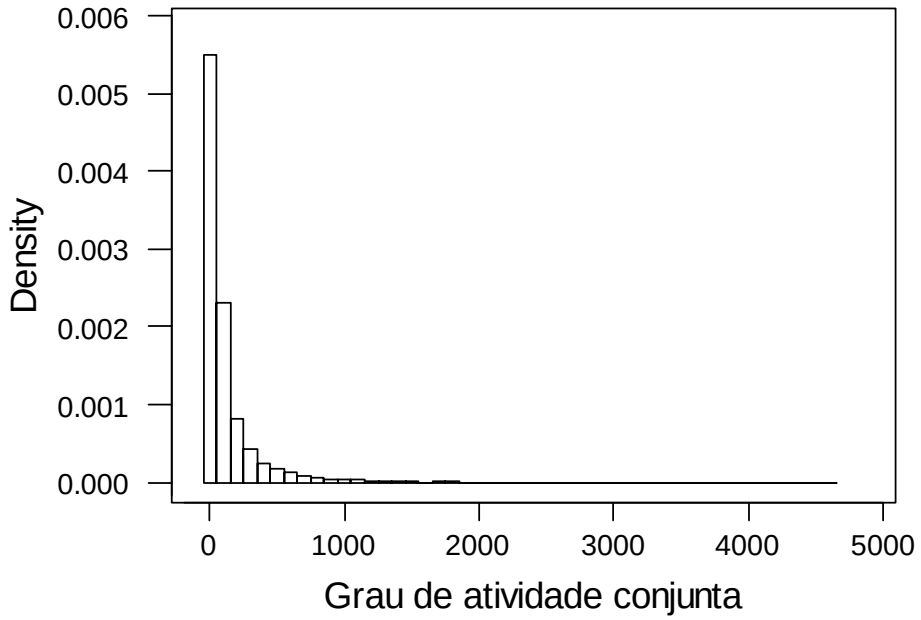
3.3.b. Partição **UU2**

Para o agrupamento dos usuários interessa usar somente a matriz **UU**.

A Figura 16 traz o histograma dos valores encontrados na matriz triangular inferior da partição **UU2** e mostra que, no conjunto, os usuários são pouco relacionados com outros usuários (note a assimetria positiva da distribuição).

Figura 16

Ligações entre Usuários



3.4. Matriz de Distâncias Temáticas

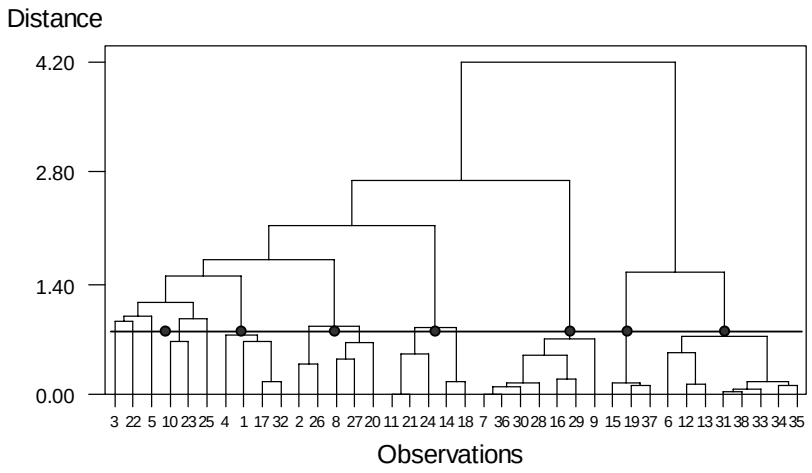
Sobre a matriz **UU2** calcula-se a matriz **T** de Distâncias Temáticas Generalizadas, anteriormente definida (veja item II.5.7).

3.5. Reunião dos usuários em Subgrupos Especializados

Aplicando-se um algoritmo hierárquico aglomerativo, distância euclidiana e critério Ward, sobre a matriz T, obtém-se os agrupamentos de usuários representados na Figura 17.

Figura 17

Dendrograma da Matriz T do GT7 (Marketing)



Com base na análise descritiva de possíveis pontos de corte convenientes, formaram-se 7 SGEs, cujos perfis são apresentados no item 3.6.

3.6. Caracterização dos SGEs

O nível de atividade em cada Assunto Significativo, medido pelo freqüência com que itens daquele assunto foram transacionados pelos membros do Grupo Temático, caracteriza o comportamento do Grupo.

No caso do Grupo Temático de Marketing, como um todo, nota-se no gráfico superior esquerdo da Figura 18 um pico de atividade no AS 36 (Economia e Maketing) e nos ASs 34, 37 e 38 (Marketing). Os gráficos seguintes, na mesma figura, mostram o perfil de cada um dos 7 Subgrupos Especializados em que o Grupo Temático foi subdividido.

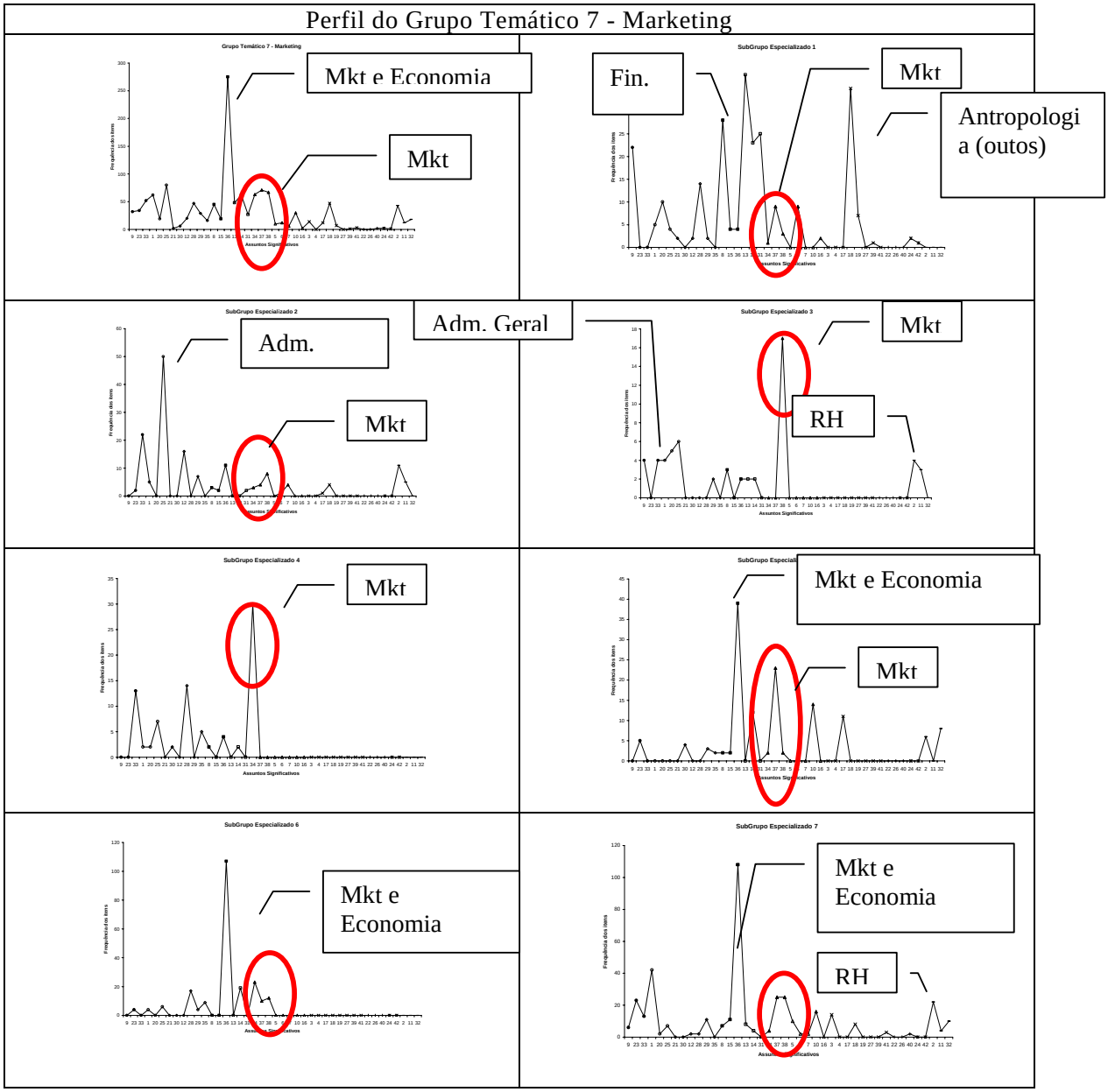
Para o Subgrupo Especializado 1, por exemplo, Marketing é assunto secundário, predominando Finanças e Antropologia na movimentação de itens; já no Subgrupo Especializado 3, Marketing é o assunto principal, com picos secundários bem menores em Administração Geral (AS 25) e Recursos Humanos (ASs 2 e 11).

O que se postula ao adotar o Sistema de Recomendações por Cooperação Indireta é que o tipo de interesse, no caso do exemplo interesse por marketing, é diferente de um subgrupo especializado para o outro. Para o SGE 1 podemos sugerir títulos que tratem de Markting e Finanças; para o SGE 3 podemos sugerir títulos mais especializados.

A forma como os títulos a serem sugeridos são selecionados está descrita sucintamente nos itens a seguir.

Figura 18

Perfil Temático dos Subgrupos Especializados do Grupo Temático de Marketing



4. CRIAÇÃO DAS LISTAS DE RECOMENDAÇÃO

Uma vez criados os SGEs, as recomendações são geradas da mesma forma que no projeto anterior, já que, a partir deste ponto, a metodologia é a mesma para recomendações por Análise de Agrupamentos e por Análise de Redes. Uma explicação detalhada e exemplos destes procedimentos podem ser encontrados em ARANHA (2.000; e em editoração). No entanto, fazemos a seguir um breve resumo destes procedimentos, organizados em duas etapas: a criação das Listas-Base e a Seleção de Itens para Listas Personalizadas.

4.1. Criação de Listas-Base para o SGE

Na penúltima etapa do processo de geração de recomendações por meio de cooperação indireta cria-se uma Lista-Base de recomendações para cada SGE. Esta lista é formada através da seleção e consolidação dos itens que estejam associados aos Assuntos Significativos do Grupo Temático em foco e que tenham sido transacionados pelos membros do Subgrupo Especializado. Os itens são ordenados decrescentemente por número de ocorrências, associando-se esta ordem a um nível também decrescente de relevância (*"ranking"*) que corresponde ao número de usuários que transacionaram o item.

Tabela 15

Exemplo de Formação de Lista-Base

Item			Patron Key			
CDU	Título	Ranking	410	3152	3194	5953
658.62	Gerenciamento por categorias: melhores praticas	4	2	2	2	13
658.818	ECR Brasil : visao geral e potencial de redução de custos	3	2	2		7
658.788	Padronização	3	2	2		9
658.62	Reposição continua de mercadorias	3	2	2		10
657.47	Custeio baseado em atividades	3	2	2		8
657.422.2	Processos financeiros	3	2	2		9
65.012.45	EDI aplicado a cadeia de abastecimento	3	2	2		6
659.126.1	Retail power plays: from trading to brand leadership : s...	2			2	5
658.8	Trade marketing strategies: the partnership between manu...	2			2	9
659.126.1	Marcas de supermercado	1				3
658.86/.87	Retail distribution management: a strategic guide to dev...	1				9
658.86/.87	Marketing channels: a management view	1				5
658.8	Marketing de relacionamento: estratégias bem sucedidas	1			2	
658.8	Database marketing estrategico	1			2	
658.8	Tecnologia de informação e comunicão	1				2
658.8	Principios de marketing	1			1	
658.7	Retail buying	1			2	
658.7	Politicas de suprimento, tecnologia de producao e tecnologia	1		1		
339.37	Contemporary retailing	1			2	
339.37	Retail saturation :examining the evidence	1				5
339.37	Retail management	1			2	
339.37	Retailing management	1			2	
339.37	Modern retailing: theory and practice	1			2	
339.37	Retailing: new perspectives	1			2	

Assim, por exemplo, observa-se na Tabela 15 um Subgrupo Especializado de Marketing formado pelos usuários de número 410, 3152, 3194 e 5953. As linhas da tabela representam os livros consultados por estes usuários, e consolidados na Lista Base. Os usuários 410 e 3152 transacionaram duas vezes cada um dos 7 primeiros títulos da tabela. O usuários 3152 também tomou emprestado uma vez o item "Políticas de Suprimento, tecnologia de produção...". Já os usuários 3194 e 5953 tomaram emprestados itens bastante diversificados. O item mais importante do conjunto é o primeiro, que foi retirado por todos os membros do SGE; em seguida, empatados, vêm 6 títulos com posto 3 (foram retirados por 3 dos quatro membros do

grupo); depois, mais dois títulos empatados com posto 2; e, finalmente, 15 títulos empatados com posto 1.

O pressuposto é de que, por serem os usuários tematicamente parecidos, o item que interessou a um membro do grupo muito provavelmente interessará aos demais.

4.2. Seleção de itens para a Lista Personalizada

Uma vez formada a Lista-Base, pode-se passar à criação da lista personalizada para cada usuário. Esta será simplesmente a Lista-Base da qual se excluem os itens que o usuário em foco já conhece, isto é, já transacionou. Os itens remanescentes serão apresentados como sugestão, na ordem decrescente de prioridade em que foram ordenados na Lista-Base.

Assim, por exemplo, o usuário 410 receberia como recomendações todos os 17 itens da Lista-Base da Tabela 15 que não transacionou. Já o usuário 5953 receberia apenas 10 sugestões: aquelas correspondentes aos itens que permanecem em branco na coluna correspondente a este usuário.

IV. CONCLUSÕES E BIBLIOGRAFIA

1. CONCLUSÕES

As técnicas de Análise de Redes oferecem maior eficiência na identificação dos Grupos Temáticos e dos Subgrupos Especializados, por permitirem o aproveitamento de toda a informação disponível sobre o comportamento dos usuários; permitem também o uso de várias medidas quantitativas das características dos

relacionamentos indiretos entre os usuários da biblioteca, não disponíveis quando se utiliza simplesmente a Análise de Agrupamentos, como no projeto anterior. As recomendações geradas, no entanto, são bastante semelhantes segundo as duas estratégias. As vantagens da Análise de Rede são mais computacionais do que de conteúdo. A forma matricial de organização dos dados facilita sobremaneira a realização dos cálculos necessários à Análise de Redes (AR); ao contrário do que supúnhamos inicialmente, o uso de AR facilita a automatização dos procedimentos.

A generalização do conceito de Distância Temática, inicialmente proposta por SCHWARTZ e WOOD (1993) para variáveis dicotômicas, foi um dos resultados mais importantes deste projeto de pesquisa; sua aplicação a uma matriz de Co-ocorrências propicia uma melhor caracterização da proximidade temática dos usuários.

O software UCINET 5.0 foi útil na etapa de aprendizado e familiarização com as técnicas de Análise de Rede. No entanto, outros aplicativos com capacidade de manipulação de matrizes e funcionalidades estatísticas e de programação, como o S-plus, mostraram-se mais práticos para a automação dos processamentos e para a análise de grandes conjuntos de informações.

Os testes realizados com mapas baseados em escalonamento multidimencional não se mostraram úteis na interpretação dos grupos, e por este motivo continuamos aplicando a caracterização por meio de perfis como os da Figura 18.

2. DESDOBRAMENTOS E OPORTUNIDADES PARA NOVAS PESQUISAS

Embora a Análise de Rede tenha se mostrado mais eficaz e eficiente para o problema analisado neste projeto, pudemos perceber que as exigências, para o processamento das matrizes envolvidas, de memória RAM e de espaço de disco para

swop de memória crescem exponencialmente com o número de usuários a serem analisados; da mesma forma, o tempo necessário à realização dos cálculos cresce exponencialmente.

De fato, enquanto estávamos na fase final de processamento dos dados da biblioteca, fomos convidados por uma empresa de e-commerce a gerar recomendações de produtos para vendas cruzadas a seus clientes. A aplicação da metodologia delineada neste relatório produziu Grupos Temáticos com número de clientes na ordem das centenas de milhares. O processamento destes grupos para a criação de SGEs por meio da estratégia estudada não se mostrou escalável para grupos com mais de 10.000 usuários.

Em conseqüência, a sugestão de novas estratégias de processamento, provavelmente baseadas na decomposição do problema em partes menores, será fundamental para viabilização dos cálculos da matriz de Distância Temática. Estas novas abordagens representam uma interessante área de pesquisa para projetos futuros.

Da mesma forma, será necessário encontrar alternativas computacionalmente convenientes aos algoritmos hierárquicos aglomerativos; consideram-se alternativas promissoras combinações de procedimentos de amostragem com alguma variação dos algoritmos do tipo k-médias, ajustada para a situação em que apenas as distâncias (temáticas) entre os indivíduos são conhecidas, mas não sua posição.

Finalmente, a aplicação de redes neurais de Cohonen talvez ofereça solução às duas categorias de problemas encontradas com relação ao custo computacional; a aplicação deste tipo de modelo também fica como uma possibilidade para novas pesquisas.

3. BIBLIOGRAFIA

ANGULO, Marcelo Junqueira e ALBERTIN, Alberto. *Portais ou Labirintos?* **Resumos dos Trabalhos ENAMPAD 2000,** Rio de Janeiro: ANPAD, 2000, inclui CD-ROM.

AAKER, D. A. and DAY, G. S. **Marketing Research**, 4th ed., New York: John Wiley and Sons, 1990.

ARANHA, Francisco. **Perfil de Usuários da Biblioteca Karl A. Boedecker***:* Geração de Valor para Pesquisadores por Meio de Cooperação Indireta. São Paulo: NPP/EAESP/FGV, relatório em editoração com edição prevista para o segundo semestre de 2000.

ARANHA, Francisco. *"E-Service em Bibliotecas: Geração de Valor para Pesquisadores por Meio de Cooperação Indireta"*. **RAE – Revista de Administração de Empresas** São Paulo: EAESP/FGV, v. 40, n. 4, Out/Dez 2000, pp. 84-93.

BEANE, T. P and ENNIS, D. M. *"Marketing Segmentation, a Review"*. **European Journal of Marketing**, 21, 5, pp. 20-42, 1987.

BERRY, M. J. A. and LINOFF, G. **Mastering Data Mining**. New York: John Wiley and Sons, 2000, 494 pp.

BERRY, M. J. A. and LINOFF, G. **Data Mining Techniques: For Marketing, Sales and Customer Support,** New York: John Wiley, 1997.

BORGATTI, S. P; EVERETT, M. G; e FREEMAN, L. C. **UCINET 5.0 Version 1.0**. Natick: Analytic Technologies, 1999.

CABENA, P.; HADJINIAN, P; et al. **Discovering Data Mining**, Upper Saddle River: Prentice Hall, 1997.

CARSON, Kerry D; CARSON, Paula P e PHILIPS, Joyce S. **The ABCs of Collaborative Change:** The Manager's Guide to Library Renewal. Chicago: ALA Editions, 1997, 272 pp.

DONNARD, Heloisa e BOREGAS, Edmilson. **A experiência da biblioteca no processo de informatização**. Documento apresentado no 1º Encontro Nacional de Usuários do VTLS, realizado na UERJ, Universidade Estadual do Rio de Janeiro, RJ, 16 de Setembro de 1998.

DUGAN, Sean M. *"In Search of Relevance: When Can We Expect the Information Age to Arrive?"*. **InfoWorld,** Framingham: InfoWorld, v. 22, n. 30, Jul 24, 2000, pp. 88.

GOLDBERG, Henry e SENATOR, Ted. *"Restructuring Databases for Knowledge Discovery by Consolidation and Link Formation"*. **Proceedings of the First Intenational Conference on Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press, 1995.

GREEN, P. E. and KRIEGER, A. M. *"Segmenting markets with conjoint analysis"*. In **Journal of Marketing**, 55, October, pp. 20-31, 1991.

HAIR, Joseph; ANDERSON, Rolph; TATHAM, R; e BLACK, Willian. **Multivariate Data Analyis**, Englewood Cliffs: Prentice Hall, 1.995, 745 pp.

HOLMES, C. *"AID comes to the aid of marketing management"*. In **European Journal of Marketing**, 14, pp. 409-13, 1981.

KASS, G. *"An exploratory technique for investigating large quantities of categorical data"*. In **Applied Statistics,** 29, pp. 127-9, 1980.

KNOKE, David e KUKLINKSI, James H. **Network Analysis**, Newbury Park: Sage Publications, Sage University Paper 28, 1982, 96 pp.

KOTLER, Philip, "*Personalização em Massa*", **HSM**, novembro-dezembro 1997, p.136.

LARGE, Peter. **The Micro Revolution Revisted.** New Jersy: Rowman & Allanheld, 1984.

MACKLACHLAND, D. L. and JOHANNSON, J. K. *"Market segmentation with multivariate AID"*. In **Journal of Marketing**, 45, pp. 74-84, 1981.

PEPPERS, Don e ROGERS, Martha. "*Is your company ready for one-to-one marketing?*" in **HBR**, January-February 19999, p. 151.

PAYTON, David. *"Discovering Collaborators by Analyzing Trails Through an Information Space"*, **Artificial Intelligence and Link Analysis Papers from the 1998 Fall Symposium,** October 23-25, Orlando, Florida.

PEPPERS, Don e ROGERS, Martha. *"Is your company ready for one-to-one marketing?"*, **HBR**, January-February 1999, p. 151-163.

RIQUIER, C; LUXTON, S.; and SHARP, B. *"Probabilistic Segmentation Model"*. In **Journal of the Market Research Society**, v. 9, n4, pp. 571-587, October 1997.

ROSENWALD, Peter. "Quatro Anos de Datalistas: Em Busca da Relevância". **Netpeople**, São Paulo: Editora Abril, ano 5, n. 19, Set 2000.

SANTOS, Érico R. **Implantação de Tecnologia de Data Warehouse em Bibliotecas com Uso de Tecnologia Adequada.** São Paulo: EAESP/FGV, Relatório Parcial de PIBIC – Programa de Iniciação Científica, 1999, 53 pp.

SCHWARTZ, Michael F. e WOOD, David C. M. "Discovering Shared Interests Using Graph Analysis". **Communications of the ACM,** v. 36, n.8, August 1993, pp. 78-89.

SHEPARD, David. **Database marketing: o novo marketing direto**, R. Janeiro: Makron Books, 1.993, 347 pp.

SMYTH, Barry e COTTER, Paul. *"A Personalized Television Listings Service"*. **Communications of the ACM.** New York: Association for Computing Machinery, v. 43, n.8, Aug 2000, pp. 107-111.

SWANSON, Don e SMALHEISER, David. *"Link Analysis of Medline Titles as an Aid to Cientific Discovery"*. http://kiwi.uchicago.edu/libtrends.html, link válido em 04.01.00 às 11h41.

VTLS INC. **VTLS Release 1994, Statistics by Command Subsystem**.. Documento de uso interno.

WASSERMAN, Stanley e GALASKIEWCZ, Joseph (eds). **Advances in Social Network Analysis,** Thousand Oaks: Sage Publications, 1994, 299 pp.

WASSERMAN, Stanley e FAUST, Katherine, **Social Network Analysis,** Cambridge: Cambridge University Press, Structural Analysis in The Social Sciences 8, 1994, 825 pp.

WEISS, S. M. and INDURKHYA, N. **Predictive Data Mining: A Practical Guide,** San Francisco: Morgan Daufmann Publishers Inc, 1.998.

WURMAN, Richard Saul. **Ansiedade de Informação**. São Paulo: Cultura, 1991, 380 pp.