Sandro de Azambuja

ESTUDO E IMPLEMENTAÇÃO DA ANÁLISE DE AGRUPAMENTO EM AMBIENTES VIRTUAIS DE APRENDIZAGEM

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática / Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática

Orientadores:

Claudia Lage Rebello da Motta, D.Sc.
Marcos da Fonseca Elia, Ph.D.

Rio de Janeiro, RJ, Brasil
2005

A991 Azambuja, Sandro de.
Estudo e implementação da análise de agrupamento em ambientes virtuais de aprendizagem / Sandro de Azambuja. – Rio de Janeiro, 2005.
xv, 197 f.: il.

Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, Instituto de Matemática, Núcleo de Computação Eletrônica, 2005.

Orientadores: Claudia Lage Rebello da Motta; Marcos da Fonseca Elia

1. Analise de Agrupamento – Teses. 2. Ambientes Virtuais de Aprendizagem – Teses. 3. Informática na Educação – Teses. Claudia Lage Rebello Motta (Orient.). II. Marcos da Fonseca Elia. (Orient.). III. Universidade Federal do Rio de Janeiro. Instituto de Matemática. Núcleo de Computação Eletrônica. IV. Título.

CDD:

Sandro de Azambuja

ESTUDO E IMPLEMENTAÇÃO DA ANÁLISE DE AGRUPAMENTO EM AMBIENTES VIRTUAIS DE APRENDIZAGEM

Rio de Janeiro, 31 de março de 2005

Aprovada por:

Claudia Lage Rebello da Motta, D.Sc.
Universidade Federal do Rio de Janeiro

Marcos da Fonseca Elia, Ph. D.
Universidade Federal do Rio de Janeiro

Flávia Maria Pinto Ferreira Landim, D.Sc.
Universidade Federal do Rio de Janeiro

Sérgio Crespo Coelho da Silva Pinto, D.Sc.
Universidade do Vale do Rio dos Sinos, RS

Para meus amores: Eliana, minha esposa, Pedro e Daniela, nossos filhos. O amor deles me permitiu chegar até aqui.

Agradecimentos

À Eliana, por todo o amor, carinho, apoio, paciência, tolerância, compreensão e, sobretudo, por ter cuidado dos nossos filhos e de mim, seu marido, durante toda essa jornada em busca do título. Amor, as palavras são insuficientes para te agradecer.

Ao meu filho Pedro, por ter me proporcionado momentos de diversão e descontração através de brincadeiras e jogos, das idas ao Maracanã para assistir o Flamengo, das idas ao Leme para jogar gol-a-gol, e por tudo que tenha sofrido devido a minha impaciência, e por ter agüentado os vários momentos em que estive ausente durante esse último ano.

À Daniela, minha garotinha, que, mesmo sem saber, foi muito importante fazendo companhia ao seu pai durante a madrugada, permitindo, quando estava meio sonolenta, que eu desse as mamadeiras nesse período.

À minha mãe, Ilda, pela formação educacional que me proporcionou e por todo o apoio e compreensão ao receber o Pedro nos finais de semana em sua casa, e pelas vindas à nossa para ajudar a cuidar dos seus netos.

Ao meu querido professor Marcos Elia, por toda a paciência e dedicação que devotou a esse esforçado aluno, ao iluminar vários caminhos que foram trilhados durante essa árdua mas produtiva experiência, que teve seu ápice quando o programa finalmente gerou os grupos. Mestre, muito obrigado!

Às professoras Claudia Motta e Sueli Mendes, por terem acredito em meu potencial e em especial à Claudia, por apresentar o possível tema da dissertação desde a entrevista, em 2002, e por ter me orientado nos momentos críticos.

À professora Flavia Landim, por ter indicado, no início de 2004, importantes livros sobre Análise de Agrupamento.

À tia Deise, Sandra, Zezé, Edileuza, Regina e Adriana, por todo o apoio administrativo a esse velho aluno do Fundão, e à Selma Mendes, bibliotecária do NCE, pela atenção dada.

Ao meu pai, por ter sobrevivido à operação cardíaca em 2004. Ao meu primo Felipe, por toda a força e dedicação doadas durante o período da programação VB e ao meu irmão, Xandre, que mesmo longe torceu pelo meu êxito.

Aos Giambiagi, por toda a compreensão, ajuda e apoio: Myriam, Irene, Marcia, Fabio, Mario (*in memorian*), o qual foi uma das razões da minha opção pelo mestrado, e aos meus sobrinhos, Luciano e Emiliano Tolivia.

Aos amigos que fiz no mestrado, sofredores como eu no dia-a-dia dessa batalha: Ilan, Jorge Fernando, Macário, Adolfo, Carlos França, George, Solange, Teresa, Renata, Rafael di Lego e Patrick. Agradeço especialmente ao Ilan e JF pelo apoio logístico na defesa.

Aos meus amigos particulares: Xanxan, por toda a ajuda e força que me deu; Marcos César e Léo, pelos vários encontros no Burguesão, onde nos divertíamos tentando saber quem torcia pelo pior time; ao JJ, Mauro, Velloso, Valéria, Carlos Nélio, Luiz, França e a todos que acompanharam e torceram por mim durante a construção da dissertação.

Vai passar

Francis Hime - Chico Buarque/1984

Vai passar
Nessa avenida um samba popular
Cada paralelepípedo
Da velha cidade
Essa noite vai
Se arrepiar
Ao lembrar
Que aqui passaram sambas imortais
Que aqui sangraram pelos nossos pés
Que aqui sambaram nossos ancestrais

Num tempo
Página infeliz da nossa história
Passagem desbotada na memória
Das nossas novas gerações
Dormia
A nossa pátria mãe tão distraída
Sem perceber que era subtraída
Em tenebrosas transações

Seus filhos
Erravam cegos pelo continente
Levavam pedras feito penitentes
Erguendo estranhas catedrais

E um dia, afinal
Tinham direito a uma alegria fugaz
Uma ofegante epidemia
Que se chamava carnaval
O carnaval, o carnaval
(Vai passar)

Palmas pra ala dos barões famintos
O bloco dos napoleões retintos
E os pigmeus do bulevar
Meu Deus, vem olhar
Vem ver de perto uma cidade a cantar
A evolução da liberdade
Até o dia clarear

Ai, que vida boa, olerê
Ai, que vida boa, olará
O estandarte do sanatório geral vai passar
Ai, que vida boa, olerê
Ai, que vida boa, olará
O estandarte do sanatório geral
Vai passar

RESUMO

AZAMBUJA, Sandro de. Estudo e implementação da análise de agrupamento em ambientes virtuais de aprendizagem. Rio de Janeiro, 2005. Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, Instituto de Matemática / Núcleo de Computação Eletrônica, 2005.

Este trabalho propõe a aplicação das técnicas estatísticas de Análise de Agrupamento em um conjunto de dados, apurados de um arquivo onde são registradas as informações relativas à participação de alunos em um Ambiente Virtual de Aprendizagem, como método de identificação e geração de grupos homogêneos para desempenhar tarefas em cenários pedagógicos. Cada grupo gerado através destas técnicas, estaria apresentando aqueles alunos mais semelhantes conforme os critérios pertinentes ao cenário pedagógico escolhido. No contexto da Análise de Agrupamento, investiga-se, sob a forma de estudos de caso, os algoritmos e os indicadores de similaridade, referenciados na literatura, mais adequados a este tipo de aplicação, como também, a estrutura e as variáveis necessárias a um arquivo de registro de acessos, o Log, do qual será retirado o conjunto de dados. Para ilustrar a aplicação desses estudos, propõe-se ainda um protótipo de uma interface informatizada do sistema ora proposto para uma plataforma de Ensino a Distância. Espera-se que a posterior utilização desses grupos, ou derivados destes, que atenderam a cenários pedagógicos pré-determinados ou parametrizados, permita o aumento das interações nesses ambientes, proporcionando assim maiores condições para a ocorrência da aprendizagem desejada em ambientes virtuais.

Palavras-chaves: Análise de Agrupamento, EAD, Ambientes Virtuais de Aprendizagem, Informática na Educação, Grupos Naturais.

ABSTRACT

AZAMBUJA, Sandro de. Estudo e implementação da análise de agrupamento em ambientes virtuais de aprendizagem. Rio de Janeiro, 2005. Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, Instituto de Matemática / Núcleo de Computação Eletrônica, 2005.

This work considers Cluster Analysis as a tool for identifying groups of pupils belonging to a Virtual Learning Environment. Our main aim is to identify groups which are similar with respect to some criteria. Each obtained group represents the students which are more similar according to some criteria and will be used for some pedagogical activity. In the Cluster Analysis context, we investigate different algorithms and different similarities indexes, as well as, different structures and variables of a file with the registered students of the Environment. We also propose a prototype of an interface for a platform of distance learning. We hope that the use of these different techniques, which take into account different pedagogical scenarios, will result in more interaction among the students in these environments.

Keywords: Distance Learning, Cluster Analysis, Natural Groups.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

AVA	Ambientes Virtuais de Aprendizagem
CLF	*Common Log File*
CSCL	*Computer Supported Collaborative Learning*
CSCW	*Computer Supported Collaborative Work*
EAD	Educação a Distância
ECLF	*Extended Common Log File*
HTML	*HyperText Markup Language*
HTTP	*HyperText Transfer Protocol*
IDAC	Instituto de Ação Cultural
IDH	Índice de Desenvolvimento Humano
IIS	*Internet Service Provider*
IM	Instituto de Matemática
IPEA	Instituto de Pesquisa Econômica Aplicada
NCE	Núcleo de Computação Eletrônica
NCSA	*National Center for Supercomputing Applications*
ODBC	*Open Database Connectivity*
PIB	Produto Interno Bruto
PUC	Pontifícia Universidade Católica
UFRJ	Universidade Federal do Rio de Janeiro
UNICAMP	Universidade Estadual de Campinas
URL	*Unified Resource Locator*
W3C	*World Wild Web Consortium*
WIMPE	*Web Interface for Managing Programs Electronically*
WWW	*World Wide Web*
ZDP	Zona de Desenvolvimento Proximal

SUMÁRIO

1 INTRODUÇÃO

A utilização de Ambientes Virtuais de Aprendizagem (AVA), especialmente as plataformas de Educação a Distância, para apoiar ou até mesmo substituir o ambiente tradicional de ensino, a sala de aula, cresceu bastante nos últimos anos devido, principalmente, à democratização do acesso à rede mundial, a Internet. Atualmente, parte dos alunos tem a facilidade de acessar a *web* nos laboratórios das escolas que freqüentam, o que lhes permite visitar salas de aulas virtuais, bibliotecas *on line*, participar de grupos de trabalho e discussão, compartilhar idéias e recursos, interagir com professores, especialistas e outros alunos, etc. Aumentando assim suas opções de aprendizagem.

Quando um aluno participa de um curso através de uma plataforma de educação à distância, as possíveis interações com os outros participantes (colegas e professores) nesse ambiente são distintas daquelas comuns à sala de aula. Para Souto (2003), a assistência pedagógica individualizada que é fornecida pelo professor ao aluno, no ambiente clássico de ensino, é consolidada através das observações e avaliações que o professor constrói, ao longo das aulas, a partir das interações, principalmente as face-a-face, que ocorrem entre todos os participantes. Esse conhecimento (tácito) que um professor tenha adquirido, proporciona para o mesmo uma maior facilidade ao desempenhar a tarefa de montar grupos de alunos, necessários a alguma demanda da disciplina. O clima existente nesses grupos, ao desempenharem tarefas colaborativas e cooperativas, é, na maioria das vezes, ideal para o compartilhamento do conhecimento entre os seus integrantes, o que irá contribuir para o processo de aprendizagem.

Nas plataformas EAD, as possíveis interações ocorrem através das funcionalidades das ferramentas existentes. A comunicação e a interação social, tão comuns ao ambiente tradicional de ensino, passam a ser intermediadas e ficam restritas às comunicações que ocorrem nos *chats*, *fóruns*, *videoconferência* e outros.

Portanto, em ambientes de Educação a Distância (EAD) a tarefa de observar e avaliar o aluno é dificultada pela perda do contato face-a-face entre o mesmo e o professor (SILVA *et al*, 2001) e entre os participantes em geral, diminuindo assim as interações "produtivas" que poderiam vir a ocorrer nesse ambiente, o que dificulta sobremaneira a tarefa da construção de grupos de alunos (homogêneos ou heterogêneos) pelo professor.

Segundo TAROUCO *et al* (2000), a existência das interações entre os participantes em um ambiente EAD, juntamente com a troca de informações que ocorrem dentro de um grupo, são essenciais para que ocorra a aprendizagem.

As plataformas de EAD disponibilizam, ao professor, uma série de recursos para acompanhar as interações do aluno com os outros participantes e com os recursos da plataforma. Esses recursos acessam os registros das atividades do aluno que se encontram armazenados em um arquivo físico, em formato texto ou banco de dados, situado em algum diretório do servidor que suporta a plataforma EAD. Esse arquivo é conhecido por Log[1] e, normalmente, apresenta grande quantidade de informações sobre os acessos dos participantes à plataforma.

A presença desses recursos, divididos em Acompanhamento e Avaliação, em ambientes virtuais de aprendizagem pesquisados por Silva e Vieira (2001), encontram-se na Figura 1.1.

---

[1] No campo da informática é comum discriminar por **Log** o arquivo físico que contém informações referentes ao acesso de usuários a um banco de dados, a uma página *web*, a um sistema, etc. Em adicional, expressões como "Log de eventos", "Reter Log", "Log de Sistemas", "Registros de Eventos do Log", "Log da Rede", "Log da Plataforma", originam-se desse termo.

	Ambiente / Mecanismo	AulaNet	Blackboard	Carnegie	ClassNet	CyberQ	Docent	E-college	EduSystem	Embanet	FirstClass	IntraLearn	LearnLinc	LearnSpace	Serf	TopClass	Virtual-U	Web Course In A Box	WebCT
Acompanhamento	Rastreamento		×	×		×	×	×	×	×	×	×		×	×	×		×	×
	Redirecionamento por teste								×	×		×				×			×
	Registros de *chats*				×														×
	Registros de listas	×			×														×
Avaliação	Análise de texto					×													
	Auto-avaliação		×					×				×				×			×
	Reuso de questões	×	×																
	Testes temporizados		×						×			×				×			×
	Testes personalizados		×	×															
	Testes via Web	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×		×	×
	Testes adaptáveis		×			×						×				×			×
	Trabalhos via Web	×		×	×				×		×				×		×		×

Figura 1.1 – Mecanismos de Avaliação e Acompanhamento em Ferramentas EAD

Se na sala de aula, a ocorrência das interações produz condições para que o professor organize grupos e estes, ao desempenharem tarefas, produzirão outras interações que novamente serão apuradas pelo professor, os grupos podem ser considerados causadores de interações. Então, se houvesse um meio de se gerar grupos para um ambiente virtual de aprendizagem, onde grande parte das interações apresentam-se pouco satisfatórias para que o professor forme grupos como faria na sala de aula, esses grupos produzidos, ao desempenharem tarefas nesse ambiente, possivelmente estarão contribuindo para o aumento das interações, as quais poderão ser observadas pelo professor.

A solução para a demanda da identificação de grupos de alunos (homogêneos) em ambientes virtuais de aprendizagem é representada por um conjunto de técnicas conhecidas

por Análise de Agrupamento. Essas técnicas são provenientes de uma das áreas da Estatística conhecida por Análise Multivariada.

A aplicação e incorporação da Análise de Agrupamento para a geração de grupos de alunos nesses ambientes representam o principal tema da presente dissertação. Essa solução requer a utilização dos dados apurados de um Log mínimo e suficiente que contenha informações sobre os participantes da plataforma onde ocorrerá a geração dos grupos.

1.1 Justificativas

Análise de Agrupamento é um conjunto de técnicas que, utilizada sob determinadas condições nos dados existentes, contribui para a formação de grupos naturais ou homogêneos (os integrantes de cada grupo formado são os mais parecidos entre si) de alunos, proporcionando aos professores dos ambientes virtuais um melhor embasamento para a tomada de decisões quando da composição de outras demandas, como por exemplo outros grupos que venham a ser necessário a algum trabalho a ser executado (tais como seminários, discussão síncrona ou assíncrona, análise de artigos e outros) nesses ambientes.

Pretendemos preencher uma lacuna detectada nas plataformas EAD, pois não foram encontrados indícios de ambientes que possuam uma ferramenta para fornecer grupos de alunos, conforme as interações destes com o ambiente, através de Análise de Agrupamento ou de algum outro método.

1.2 Objetivos e Organização

A atual dissertação tem como objetivo principal pesquisar e implementar uma ferramenta de Análise de Agrupamento para plataformas de Educação a Distância. A

identificação de cenários de agrupamento mais utilizados em plataformas de EAD; a definição de um Log mínimo que possa ser utilizado para aplicação de Análise de Agrupamento; e a execução de uma metodologia composta por dois Estudos de Caso são também objetivos a serem trabalhados. Para alcançar os objetivos delineados, o presente trabalho foi organizado em quatro partes resumidas a seguir.

1. Revisão da Literatura – composta por três partes: 1) O referencial educacional que sustenta a proposta; 2) Uma descrição suficiente sobre Análise de Agrupamento contendo exemplos de aplicações extraídos da literatura consultada e dados das plataformas existentes no NCE; 3) Informações pesquisadas sobre a formação e composição dos arquivos Log, buscando a identificação de um Log mínimo e suficiente para execução de técnicas de análise de agrupamento
2. Estudos de caso, total de dois, onde estarão sendo aplicadas as técnicas adequadas de Análise de Agrupamento, já identificadas na Revisão, no Log coletado nos ambientes virtuais de aprendizagem existentes no programa de Mestrado em Informática da UFRJ, objetivando a formação de grupos de alunos. A formulação da proposta, após a revisão da literatura e os estudos de caso, estará sendo formalmente apresentada ao final desse capítulo.
3. Análise, construção e implementação do protótipo de uma interface para uma plataforma EAD do NCE. Essa interface teve sua implementação concluída na presente dissertação. Nesse capítulo estarão sendo fornecidos os cenários mais comuns de agrupamento, identificados na plataforma utilizada.
4. Conclusões e Trabalhos Futuros – onde serão mostrados uma avaliação dos objetivos da dissertação e possíveis trabalhos futuros.

2 REVISÃO DA LITERATURA

2.1 Referencial Educacional

Nessa seção da revisão da literatura pretendeu-se realizar um extrato, a partir de pesquisas das teorias educacionais associadas a Vygotsky e Paulo Freire, com o objetivo de justificar e embasar a proposta central do presente trabalho. Também foram utilizadas na seção, como apoio a esse referencial, as idéias de Pedro Demo que poderão repercutir através da implementação da nossa proposta.

2.1.1 Pesquisa na Teoria Educacional de Vygotsky

A presente revisão tem o objetivo de identificar os aspectos considerados relevantes nas teorias de Vygotsky, acerca do processo educativo, que contribuam para o embasamento necessário à proposta do presente trabalho. Inicialmente apresentamos um pequeno resumo biográfico de Vygotsky e, em seguida, os aspectos relevantes identificados.

2.1.1.1 Aprendizagem como Fator de Desenvolvimento

O bielo-russo Lev Semenovich Vygotsky (1896-1934) iniciou seus trabalhos no campo da psicologia logo após a Revolução Russa de 1917. Durante sua curta vida gerou uma vasta produção intelectual, da literatura à neurologia, passando pela educação de adultos, deficientes e especiais; pela psicologia onde atuou fortemente; e lecionando em diversos institutos e universidades de Moscou. Essa variedade de áreas do conhecimento deveu-se a necessidade que possuía de reunir informações de áreas distintas, com o objetivo de compreender os diferentes aspectos da conduta do ser humano.

Vygotsky é considerado o primeiro psicólogo moderno a sugerir mecanismos pelos quais a cultura torna-se parte da natureza do ser humano. Defendeu também a associação da psicologia cognitiva experimental com a neurologia e a fisiologia.

Na área da pedagogia, criticou profundamente a concepção existente no início do século XX que associava a teoria de aprendizagem a situações estímulo-resposta. Também criticou as idéias de que as propriedades das funções intelectuais de um adulto, estariam pré-formadas na criança esperando apenas a maturação desta, o que o colocava também como um dos pioneiros do Construtivismo. Durante seus estudos educacionais, introduziu o termo "podologia" que possui um significado próximo ao que hoje conhecemos por psicologia educacional (VYGOTSKY, 1991).

Entre 1925 e 1934, quando morreu devido à tuberculose, formou um grupo de jovens cientistas desenvolvendo importantes trabalhos nas áreas da educação, psicologia e no estudo das anormalidades físicas e mentais.

2.1.1.2 As Contribuições de Vygotsky para a Educação

Vygotsky considerava o aprendizado um aspecto necessário e fundamental no desenvolvimento das funções psicológicas superiores, havendo para ele uma forte ligação entre o desenvolvimento pleno do ser humano e o que ele aprende num determinado grupo cultural, a partir da interação com outros indivíduos, o que veio a ser posteriormente denominado construtivismo social. Para ele, o aprendizado nas crianças pode ser entendido como uma internalização dos sistemas de signos que provoca transformações comportamentais, estabelecendo um elo de ligação entre as normas iniciais e tardias do desenvolvimento do indivíduo. Vygotsky define internalização da seguinte forma: "... é a

reconstrução interna de uma operação externa, onde uma série de transformações se processa" (VYGOTSKY, 1991).

A internalização transforma ou modifica a situação estimuladora, mudando sua estrutura e funções, caracterizando-se como uma aquisição social onde se processou opções escolhidas de acordo com as vivências e possibilidades de troca e interação do indivíduo.

As relações entre desenvolvimento e aprendizagem são destaques na obra de Vygotsky e ele analisa essa questão a partir de duas perspectivas: a que se refere à compreensão da relação entre o aprendizado e o desenvolvimento, que ocorre desde o primeiro dia de vida do indivíduo; e a que se refere aos aspectos dessa relação durante o período escolar (OLIVEIRA, 1997).

Em relação ao período escolar, Martins (1997) apresenta uma síntese dos objetivos e finalidades da ação educativa dentro do espectro de reflexões da psicologia sócio-histórica, na linha da tríade fundamental da ação educacional centrada no "Ser Feliz", "Ser-Cidadão" e "Ser competente para o exercício de uma profissão":

> [...] a ação educativa [...] cria condições para que os alunos se tornem cidadãos que pensem e atuem por si mesmos. Acima de tudo, espera-se que eles sejam pessoas livres de manipulações e conduções externas e que consigam ter a capacidade de pensar e examinar criticamente as idéias que lhes são apresentadas e a realidade social que partilham. (MARTINS, 1997, p. 1).

O indivíduo, como aluno, está sujeito a diversas interações durante o processo formal de ensino-aprendizagem, e também fora desse ambiente. Algumas dessas relações podem ser pensadas como situações onde interagem aluno e professor; aluno e aluno; aluno e pais; etc.

Durante essas interações, é comum o aluno ser auxiliado na construção da solução de algum problema que não esteja conseguindo resolver. Para essa situação, Vygotsky apresenta um conceito de grande importância: Zona de Desenvolvimento Proximal.

A Zona de Desenvolvimento Proximal (ZDP) é um dos principais pilares da teoria vygotskiana, utilizada pelos educadores em geral para reforçar o papel do professor como mediador ou articulador da educação a ser internalizada pelo educando.

A ZDP, que surge em decorrência do aprendizado, pode ser definida classicamente como:

> [...] a distância entre o nível de desenvolvimento real, que se costuma determinar através da solução independente de problemas, e o nível de desenvolvimento potencial, determinado através da solução de problemas sob a orientação de um adulto ou em colaboração com companheiros mais capazes. (VYGOTSKY, 1991)

E ainda:

> ZDP é onde se encontram aquelas funções que ainda não amadureceram, mas que estão em processo de maturação, funções que amadurecerão, mas que estão presentemente em estado embrionário. Essas funções poderiam ser chamadas de *brotos* ou *flores* do desenvolvimento, ao invés de *frutos* do desenvolvimento. (VYGOTSKY, 1991).

Quando um aluno somente consegue resolver determinada tarefa ou problema auxiliado por um outro sujeito, este aluno está revelando o seu nível de desenvolvimento proximal, composto por suas noções e conceitos acerca da tarefa em questão. O professor, ao conhecer essa zona proximal do aluno, poderá intervir junto a ele com o intuito de provocar, estimular e apoiar funções de aprendizagem que ainda não estejam totalmente consolidadas, pois percebeu a dinâmica interna do desenvolvimento do aluno. Para Vygotsky, o nível da ZDP de uma criança é bem mais indicativo do seu desenvolvimento mental do que aquilo que ela consegue fazer sozinha, representado pelo desenvolvimento real. Assim é possível verificar não somente os ciclos completados pela criança, como também os que ainda estão se formando, permitindo o delineamento da competência associada ao educando e quais, possivelmente, serão suas futuras aquisições (VYGOTSKY, 1991 *apud* MULTIEDUCAÇÃO, 2005). A Figura 2.1, a seguir, exibe um ambiente Sócio-Interacionista que pode ser apurado da teoria vygotskiana.

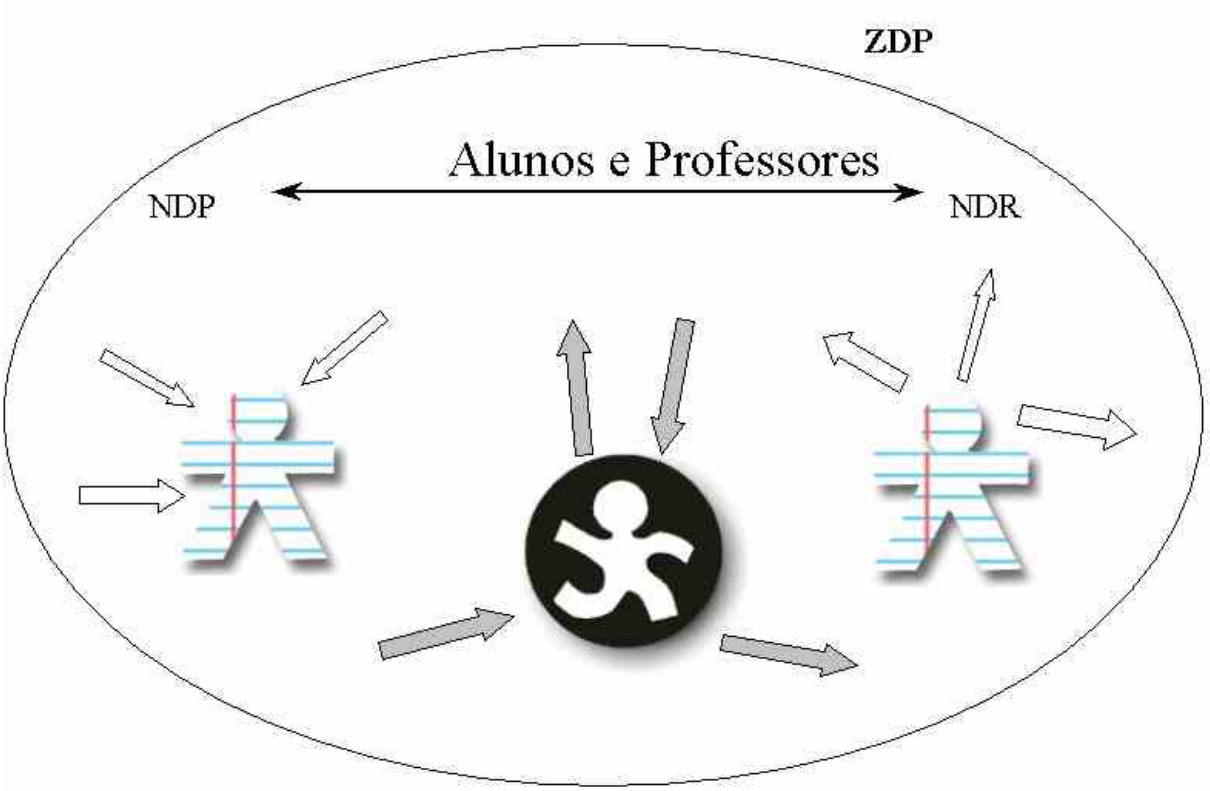


Figura 2.1 – Ambiente Sócio-Interacionista derivado das teorias de Vygotsky[2]

Vygotsky sempre apresentou suas teorias educacionais voltadas para o desenvolvimento da criança, porém seus trabalhos são utilizados por educadores principalmente para a compreensão do papel da educação no desenvolvimento das potencialidades do indivíduo, não importando a faixa etária do mesmo.

Para que essa compreensão seja alcançada, é necessário que ocorra a internalização das funções mentais que garantem ao indivíduo a possibilidade de pensar por si. Vygotsky afirma que essa internalização irá ocorrer como resultado das interações do indivíduo durante o processo de aprendizagem, que não ocorrem necessariamente na sala de aula.

[2] As siglas NDP e NDR utilizadas referem-se a "Nível de Desenvolvimento Real" e "Nível de Desenvolvimento Potencial".

A partir de todo o exposto anteriormente, podemos concluir que a ocorrência no aluno da internalização, defendida por Vygotsky, de uma operação externa poderá ter sucedido mais provavelmente devido à participação do professor ou de outros sujeitos em sua ZDP, fruto das interações do aluno durante o processo de aprendizagem. Essas interações ocorrem mais naturalmente na sala de aula porque, estando fisicamente na escola, as interações face-a-face e o contato social são freqüentes, porém são mais raras ou insuficientes em um ambiente de Educação a Distância para o qual propõe-se o presente trabalho.

Para permitir que esses processos sejam polinizados, o presente trabalho propõe a identificação de grupos naturais de participantes de um ambiente de EAD. Esses grupos estariam proporcionando condições para ocorrência de maiores interações entre esses participantes (através até mesmo de reuniões presenciais), criando condições para que o professor passe a reter maiores conhecimentos sobre os alunos desse ambiente, aumentando assim as possibilidades do mesmo tornar-se o articulador dos conhecimentos produzidos e internalizados pelo grupo.

2.1.2 Revisando a Educação Defendida por Paulo Freire

Essa revisão da literatura apresenta um extrato das propostas de Paulo Freire, educador e filósofo da educação, e um dos mais criativos intelectuais que o Brasil já teve, as quais foram utilizadas no embasamento da proposta da dissertação. Um resumo da trajetória de Freire é apresentada, para depois mostrarmos a base educacional que foi utilizada.

2.1.2.1 A Trajetória de Paulo Freire

Paulo Reglus Neves Freire nasceu em 19 de setembro de 1921, no Recife, Pernambuco. De origem humilde, foi alfabetizado em casa pela mãe e parentes, como ele mesmo disse uma vez: “Fui alfabetizado no chão do quintal de minha casa, à sombra das mangueiras, com palavras do meu mundo, não do mundo maior dos meus pais. O chão foi o meu quadro-negro; gravetos, o meu giz” (FREIRE, 1997).

Freire perdeu o Pai em 1934, na mesma época em que terminou, atrasado, o primário (atualmente, o ensino fundamental). De 1941 a 1944, cursou o secundário (ensino médio) no Recife, o pré-jurídico e o magistério, tornando-se professor de Português. Em 1943, ingressou na Faculdade de Direito do Recife. Em 1947, aceitou o convite para dirigir a Divisão de Educação e Cultura do SESI, o que considerou um marco na história de sua vida como educador, pois foi a partir daí que começou a trabalhar com educação para adultos. Nos anos 50, lecionou História e Filosofia da Educação, na Universidade do Recife, contribuindo com suas idéias para o clima de renovação e esperança respirados na época. Em 1958, teve seu trabalho “A educação de adultos e as populações marginais – o problema dos mocambos” reconhecido no II Congresso Nacional de Educação. Esse fato foi determinante para que ele passasse a ser visto e conhecido como um educador progressista. No começo dos anos 60, iniciou suas experiências com um conjunto de métodos para alfabetização de adultos que marcariam sua trajetória educacional. Seus métodos, conhecidos na época como Sistema Paulo Freire, começaram a serem utilizados em outros estados do nordeste, o que provocou polêmicas sobretudo nas oligarquias do nordeste, talvez porque na época aqueles que não sabiam ler e escrever, não podiam votar.

No início de 1964, assumiu o Programa Nacional de Alfabetização do MEC, introduzindo de seu método com o intuito de alcançar todo o território nacional, o que não conseguiu devido ao golpe militar, quando é preso e exilado, sendo seu método de

alfabetização considerado uma ameaça à ordem, pelos que tomaram o poder. Exilado, vai para o Chile, onde permaneceu por 5 anos com sua família, trabalhando com educação no campo e sendo consultor da UNESCO. Escreveu vários livros no Chile, os quais iriam embasar sua grande obra, "A Pedagogia do Oprimido", lançada em 1970. Realizou conferências no México e nos Estados Unidos, onde passou quase um ano, de 1969 a 1970. Mudou-se com a família para Genebra, onde permaneceu até 1980 como professor na Universidade de Genebra, o que possibilitou difundir suas idéias na Ásia, Oceania e América, mas sobretudo na África, nos países de língua portuguesa. Os anos 70 representaram para Paulo Freire o período mais profundo e mais rico de sua práxis pedagógica. Em Genebra, com outros exilados, funda o IDAC, voltou ao Brasil em 1980, beneficiado pela Anistia, indo viver em São Paulo trabalhando como professor da UNICAMP, de 1980 a 1990, lecionando na PUC após esse período. Ocupou a secretaria de educação do município de São Paulo, de 1989 a 1991. Morreu, de infarto, em maio de 1997 (CENTRO PAULO FREIRE, 2005).

2.1.2.2 As Contribuições de Paulo Freire para a Educação

A obra de Paulo Freire, através de seus métodos e pedagogias educacionais voltadas, principalmente, para a alfabetização, é marcada pela percepção de que a busca pela aprendizagem deve colocar o educando como o sujeito da ação de aprender. Por mais óbvio que isso possa parecer atualmente, nem sempre essa práxis é utilizada, pois na maioria das vezes a escola oprime o indivíduo.

Freire defendeu que é necessário ao educador, conhecer o meio no qual o educando vive, e também suas características únicas, para que o mesmo receba uma educação adequada para que possa aprender e apreender os significados, conseguindo transformá-los. Os educandos, estariam assim, participando de uma aprendizagem transformadora que permitirá

a aquisição da consciência das possibilidades e limitações inerentes aos seres humanos. Em muitas das experiências que vivenciou, Freire percebeu que ao buscar elementos existentes nas relações culturais dos educandos com o meio em que vivem, para utilizá-los na relação de aprendizagem, ocorre uma aproximação da educação que está sendo proposta aos problemas concretos dos educandos, despertando um maior interesse por parte destes no que está sendo colocado para aprender. Essa metodologia permite ainda, uma maior e mais ativa interação e envolvimento do educador, que passa a aprender também com seus alunos, com os educandos (FREIRE *et al*, 1989).

Em suas experiências de educação utilizando técnicas de grupos em substituição ao formato convencional das salas de aula, Freire utilizou a conversa, o grupo de estudo, o fórum, o grupo de debate, o grupo de ação e outros, como ferramentas alternativas que permitiram uma melhor apuração do aprendizado. Foi nesses ambientes de grupo onde os saberes existentes, não percebidos como tais, nos participantes surgem, sobressaem, através do clima de diálogo e discussão inerente a um grupo. É nesse aspecto que aproximamos as idéias de Paulo Freire a nossa proposta, ao entendermos que a utilização de grupos em ambientes de EAD para desempenhar tarefas, possibilitará ao professor, através da observação (e participação) das interações geradas por esses grupos, a identificação, do que chamamos, das variáveis freireanas pertencentes a cada aluno. Essas variáveis estariam exprimindo através de conceitos críticos, sócio-políticos e econômicos de cada um dos participantes dos grupos, o "chão do quintal que serve de quadro-negro", conforme sugerido por Freire. Ou seja, pressupõe-se aqui que o processo de aprendizagem à distância, através de trabalhos em grupo, será potencializado: (i) se introduzirmos variáveis do tipo freireana na análise de agrupamento, (ii) estabelecermos critérios de aproximação na geração dos grupos adequados para uma dada atividade didática que se pretenda e, sobretudo, (iii) se o professor tiver consciência da importância dessas interações não neutras e de natureza política como

atividade pedagógica do processo de aprendizagem. Agindo assim, buscamos uma coerência entre a apropriação das idéias de Freire com a nossa proposta.

2.1.3 Revisando Pedro Demo

Nessa seção estão expostas algumas idéias e concepções de Pedro Demo sobre a associação entre Pesquisa, Ensino e Educação como um fator pedagógico, as quais consideramos interessantes e que esperamos poder refleti-las, na geração dos grupos de alunos para trabalhos cooperativos à distância, através da implementação de nossa proposta. Grande parte da seção utilizou o conteúdo do livro "Pesquisa: Princípio Científico e Educativo", referenciado em (DEMO, 1992).

Pedro Demo é sociólogo e professor titular da Universidade de Brasília, atuando nas áreas de Política Social e Metodologia Científica[3].

2.1.3.1 As Idéias e Concepções

Demo defende a cotidianização da pesquisa, desde o pré-escolar, desmistificando e tornando-a parte do processo normal de formação de pessoas e grupos, vinculada ao ensino como ocorre na Europa. Agindo dessa forma, teria lugar no processo de aprendizagem (onde é comum a atitude do imitador que copia, reproduz e faz prova), a atitude da criação, da elaboração própria, da produção de novos conhecimentos.

Defende ainda, a extinção da sutil separação que vigora entre o ensino e a pesquisa, onde o saber desliga-se do mudar, o que considera causador de estigmas para a pesquisa, principalmente os associados à alienação acadêmica, que surge por conta do distanciamento

das pesquisas do cotidiano das pessoas, e à apropriação do saber, instrumento político que relega a função de transmissão socializada.

Para ele, o pesquisador deve atuar como fenômeno político, buscando sempre servir à sociedade a que pertence, socializando sua pesquisa através da inclusão natural de sua prática ao meio de onde surgiu.

Demo não aceita a forma atual da Educação, resumida a instrução, informação e reprodução. Defende que a pesquisa deve fazer parte de todo o processo educativo, ajudando na motivação da criatividade do educando.

Em uma entrevista à Revista Nova Escola (DEMO, 2001), comenta que a presença virtual, falando da Internet, vai se impor à educação, porém jamais dispensando a presença física do professor que passará a acumular o papel de ordenador das informações. Define o mundo virtual como reprodutivista, estando na esfera da informação e não da formação, porém classificando como motivador, o que é bom para fomentar o aprender.

Além de concordar com boa parte das idéias e concepções de Demo, entendemos que estas são inerentes a qualquer proposta de Educação a Distância que valoriza o trabalho em grupo (CSCL).

---

[3] Maiores informações em http://pedrodemo.sites.uol.com.br/, acessado em fevereiro de 2005.

2.2 Análise de Agrupamento

2.2.1 Introdução

A maioria dos Estatísticos considera Análise de Agrupamento[4] como o conjunto de técnicas que permitem dividir os dados, que normalmente apresentam componentes com observações multivariadas ou multidimensionais, em grupos naturais. A Análise Discriminante[5] do tipo Caixa Preta (*Black Box*), também é referenciada como Análise de Agrupamento, pelo fato de gerar grupos sem seguir uma regra explícita e definida (KRZANOWSKI & MARRIOTT, 1995).

Análise de Agrupamento (originalmente *Cluster Analysis)* é um conjunto de técnicas que tem por objetivo identificar padrões ao formar grupos homogêneos (os mais semelhantes pertencem a um mesmo grupo) a partir de *n* observações ou elementos existentes. A construção dos grupos é feita de modo que as observações de um mesmo grupo, pareçam-se mais entre si do que com as observações existentes nos outros grupos formados (BUSSAB *et al*, 1990). O termo *Cluster Analysis* foi introduzido primeiramente por Tryon, em 1939, mas segundo Bergman & Feser (1998), idéias semelhantes já vinham ocorrendo desde o final do século XIX[6].

Essas técnicas vêm sendo utilizadas em importantes áreas científicas para a identificação de padrões de comportamento nos dados analisados, auxiliando com isso o processo de descoberta do conhecimento, pois a divisão em grupos ou classes facilita a

---

[4] Em livros de Análise Multivariada em inglês, *Cluster Analysis* é o termo utilizado para Análise de Agrupamento.

[5] Segundo Barroso & Artes (2003), Análise Discriminante objetiva diferenciar populações e, conhecendo as populações *a priori*, classificar objetos nas populações pré-definidas.

[6] Idéias iniciadas por Marshall (1890), que destacou a importância das economias externas de escala aproximarem seus negócios aos distritos industriais da época; Weber (1929), ao introduzir a noção de Economias Aglomerativas e elaborar idéias sobre economias de custo; e Hoover (1937) que conseguiu distinguir, através de seus trabalhos, o que ficou conhecido por economias de urbanização e de localidade.

compreensão das observações e o desenvolvimento subseqüente de teorias científicas. Elas são utilizadas ainda na fase exploratória da pesquisa, onde a falta de hipóteses *a priori* sobre as observações e o desconhecido número de grupos, permitem que as mesmas auxiliem na organização dos dados em estruturas significativas de fácil interpretação. Ao gerarem grupos para o pesquisador, este, por ter bastante conhecimento sobre o problema, consegue distinguir e identificar os agrupamentos "bons" dos "ruins".

Em Estatística, os procedimentos exploratórios são úteis para a compreensão da complexa natureza dos relacionamentos multivariados encontrados nas informações analisadas. A busca por grupos naturais ou homogêneos nos dados de uma estrutura é uma importante técnica exploratória representada pela Análise de Agrupamento, pois fornece meios informais para avaliações dimensionais e até mesmo identificação de observações destoantes que podem vir a ser valores atípicos (*outliers*), possibilitando ainda o surgimento de hipóteses interessantes a respeito dos relacionamentos ao se produzirem taxonomias[7] (JOHNSON & WICHERN, 1988).

Além da fácil interpretação, uma outra razão pela opção da utilização de Análise de agrupamento é a não exigência de pressupostos iniciais quanto à distribuição de probabilidade dos dados. Kendall (*apud* KRZANOWSKY, 1995), distingue a procura por grupos naturais nas informações ao invés de subdividi-las meramente por conveniência, criticando essa última por considerar que a escolha subjetiva do número de grupos, significa nenhum compromisso com a busca por uma solução ótima.

---

[7] Taxonomia refere-se a um esquema criado para classificar coisas – objetos, lugares, eventos, etc.

2.2.2 Diferenciando Análise de Agrupamento de Outros Métodos

Cabe ressaltar que, apesar da existência de referências a esse termo, não estamos trabalhando com um método de Classificação, a qual, segundo JOHNSON & WICHERN (1988), pressupõe que se conheçam o número de grupos finais e suas características, sendo seu objetivo operacional associar novas observações a um desses grupos.

Comparada à Classificação, Análise de agrupamento apresenta-se como uma técnica mais primitiva onde não existe assunção sobre a existência e característica dos grupos, tampouco sobre a quantidade dos mesmos. Os agrupamentos são feitos com base nas medidas de distância (coeficientes de similaridade ou dissimilaridade), calculadas entre os indivíduos. Como pré-requisito para aplicar as técnicas de grupamento, além da existência dos dados no conjunto inicial, é necessário definir qual o critério a ser utilizado na definição de proximidade entre os objetos. Na Classificação, os objetos são associados a grupos ou classes pré-definidas, e nos Agrupamentos, os grupos (quantidade dos mesmos e que indivíduos pertencem a cada um) surgem durante o processo. Portanto, a Análise de Agrupamento também pode ser utilizada para auxiliar a uma Classificação que não tenha sido aplicada ainda devido à falta dos grupos, e características dos mesmos.

Foram pesquisados exemplos sobre Classificação, e dentre os vários encontrados apresentamos dois que consideramos interessantes por serem de suma importância para a Astronomia: a classificação das estrelas, introduzida por A. J. Cannon em 1910 e aperfeiçoada por volta de 1920, conhecida por Classificação de Harvard; e a classificação geral das galáxias, elaborada por Hubble, em 1927, que ficou conhecida como Classificação Morfológica de Hubble[8].

---

[8] O Apêndice A apresenta detalhes dessas duas classificações.

Quando da utilização das técnicas de Análise de Agrupamento, não é possível determinar antecipadamente as variáveis dependentes e independentes, ao contrário, as técnicas permitem examinar as relações de interdependência entre todo o conjunto de variáveis, nesse ponto sendo similar a Análise Fatorial, porém diferenciando-se da mesma por tratar os objetos, enquanto que a Análise Fatorial trabalha com as variáveis, buscando reduzir o conjunto das mesmas através da criação de fatores que medirão aspectos em comum (HAIR *et al*, 1998) e (BARROSO & ARTES, 2003).

Diversos autores destacam ainda que as técnicas utilizadas em análise de agrupamento referem-se à aprendizagem não-supervisionada (KRZANOWSKI & MARRIOT, 1995) e dentre estes existem aqueles que ressaltam o fato de não existirem testes de significância estatística envolvidos na sua utilização, mesmo em casos onde existem graus de liberdade (EVERITT, 1974). Esses mesmos autores consideram que as técnicas de análise de agrupamento encontram a solução mais "significativa possível", ao minimizar a variabilidade dentro de cada grupo (em relação aos elementos que os compõem) e maximizar a variabilidade entre os grupos, objetivos semelhantes existentes nas análises Discriminante e Fatorial.

2.2.3 Agrupamentos Simples e Relativização de Variáveis

Bussab *et al* (1990) ressaltam que existem alguns agrupamentos com estruturas facilmente identificáveis através de simples inspeções gráficas, e que uma aplicação de análise de agrupamento teria muita dificuldade na identificação dessas estruturas. Isso ocorre devido ao fato das técnicas partirem de suposições implícitas sobre o tipo de estrutura presente nos dados, cabendo ao analista estar atento a essas suposições. Caso essas suposições não se apliquem ao conjunto de dados, o analista deve aplicar diferentes critérios de

agrupamento, tarefa computacionalmente viável nos dias de hoje, aceitando a estrutura resultante da maior parte deles para posteriores análises.

Para esses agrupamentos de estruturas identificáveis, também caracterizados por apresentarem o número de observações e o total de variáveis bastante reduzidos, a definição do critério de proximidade pode ser feita através da análise visual de um diagrama de dispersão, como o exemplo a seguir, retirado de Barroso & Artes (2003, p.7).

Exemplo 1: A Tabela 2.1 exibe os dados referentes às taxas de delitos por divisão territorial das polícias (Deinter) do Estado de São Paulo, em 2002.

Deinter	**Homicídio doloso**	**Furto**
SJRP	10,85	1.500.80
RP	14,13	1.496,07
Bauru	8,62	1.448,79
Campinas	23,04	1.277,33
Sorocaba	16,04	1.204,02
SP	43,74	1.190,94
SJC	25,39	1.292,91
Santos	42,86	1.590,66
Média (μ)	23,08	1.375,19
DP (σ)	13,69	152,05

Tabela 2.1: Taxa de Delitos[9]

Se o objetivo, associado ao Exemplo 1, é dividir os dados em quatro grupos de regiões homogêneas quanto à incidência de homicídios dolosos e furtos, basta que uma análise visual seja efetuada no diagrama de dispersão, que utiliza apenas essas duas variáveis, mostrado na Figura 2.2:

[9]As taxas apresentam-se divididas por 100 mil habitantes.
Fonte: http://www.ssp.sp.gov.br/estatisticas/porlocal.aspx, acessada em 03/05/2004.

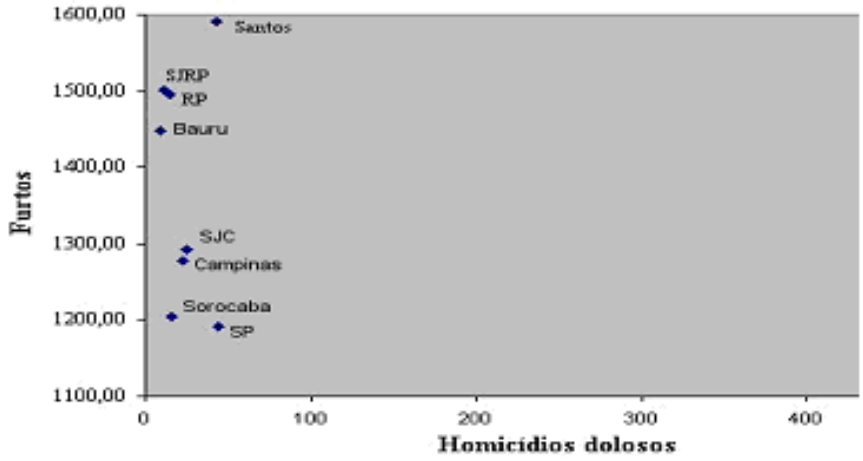


Figura 2.2 – Diagrama de Dispersão possibilitando a formação de grupos

A partir do diagrama anterior, é possível concluir que um critério razoável a ser considerado na formação dos grupos é a proximidade entre os pontos (veremos mais adiante que esse critério se assemelha com um coeficiente de dissimilaridade), ou seja, pontos próximos representam regiões com comportamentos semelhantes no que se referem às variáveis representadas no diagrama. Através desse critério é possível identificar os seguintes grupos: **(Sorocaba, São Paulo); (São José dos Campos, Campinas); (Santos); (São José do Rio Preto, Ribeirão Preto, Bauru)**.

Essa seria a composição dos quatro grupos se os dados das duas variáveis tivessem contribuído da mesma forma para a formação dos mesmos, o que não ocorreu, como podemos comprovar através da análise das distâncias vertical e horizontal do diagrama da Figura 2.2.

Por conta da variabilidade dos dados de *Furto* apresentar-se consideravelmente maior do que a variabilidade dos dados de *Homicídio doloso* – a distância dos dados no sentido vertical é muito maior do que a do sentido horizontal – será necessário relativizar as variáveis, através de transformações nas mesmas, e refazer o diagrama para que ambas venham a contribuir igualmente para a construção dos grupos (se as mesmas permanecerem com seus valores originais, a variável *Furto* influenciará bastante na formação dos grupos).

Nesse ponto pesquisamos quais seriam as justificativas existentes para a utilização dessas transformações. Bussab *et al* (1990) embasaram as justificativas encontradas ao ressaltarem que um aspecto importante, o qual deve ser sempre ponderado, é a homogeneidade entre variáveis de diferentes escalas que venham a participar de agrupamentos, pois a contribuição de uma variável ao coeficiente utilizado dependerá não somente de sua escala, mas também da escala das outras variáveis. A solução para garantir que as variáveis contribuam de forma semelhante para o coeficiente adotado seria homogeneizar suas variâncias, o que somente acontecerá se antes essas variáveis sofrerem transformações com o intuito de relativizá-las ou padronizá-las. Caso isso não seja adotado para variáveis que apresentem variâncias heterogêneas, grupos poderão ser mascarados durante o processo de agrupamento e resultados equivocados poderão ser produzidos[10].

Voltando a análise do diagrama de dispersão contido na Figura 2.2, as observações das duas variáveis, *Homicídios dolosos* e *Furtos,* estarão sendo transformadas através da relativização mais utilizada, a padronização estatística. Essa relativização transforma os valores de cada uma das variáveis, deixando-as com média nula e desvio padrão unitário. Cada valor das variáveis x_{ik} (no exemplo, *i* varia de 1 a 8, representando cada uma das cidades, e *k* varia de 1 a 2 representando as duas taxas de delitos analisadas) é transformado em uma nova variável Z_{ik}, representada pelas fórmula $Z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}$, onde μ_k é a média da k-ésima variável aleatória e σ_k seu desvio padrão (na prática usam-se média e desvio-padrão amostrais). Cabe destacar que existem outros tipos de transformação de variáveis, como

[10] Quando uma variável apresenta uma variância relativamente grande em relação à de uma outra variável, isso indica que a variação de uma unidade da primeira variável terá um significado menor do que uma mesma variação na segunda.

veremos durante a seção 2.2. A transformação efetuada nas variáveis *Furto* e *Homicídio doloso* estão na Tabela 2.2, a seguir:

Deinter	Homicídio doloso	Furto
SJRP	-0,89	0,83
RP	-0,65	0,80
Bauru	-1,06	0,48
Campinas	0,00	-0,64
Sorocaba	-0,51	-1,13
SP	1,51	-1,21
SJC	0,17	-0,54
Santos	1,44	1,42

Tabela 2.2: Variáveis Homicídio doloso e Furto, após relativização

A partir dos dados da Tabela 2.2, é possível construir um novo diagrama de dispersão, já com os quatro grupos identificados, representado pela Figura 2.3.

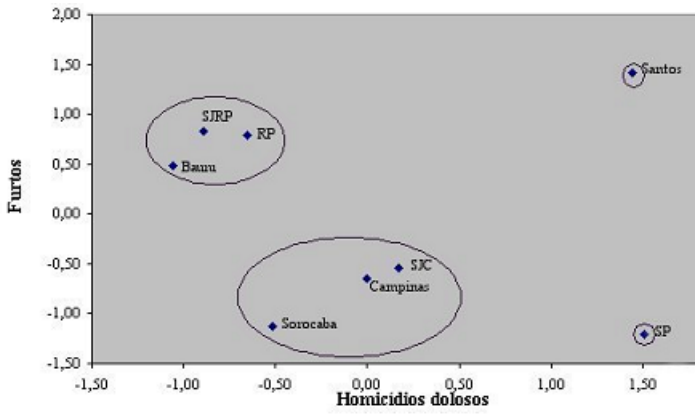


Figura 2.3 – Diagrama de Dispersão para os dados da Tabela 2.2

A partir desse novo diagrama de dispersão, são encontrados novos grupos que poderão, por exemplo, serem utilizados para auxiliar na definição de uma nova política de segurança:

(SJRP, RP e Bauru)*, *(SJC, Campinas e Sorocaba)*, *(Santos)* e *(SP)

É possível notar que três desses grupos diferem daqueles obtidos a partir da Figura 2.2, antes da relativização.

2.2.4 Definindo Alguns dos Principais Componentes da Análise de Agrupamento

2.2.4.1 Coeficientes de Similaridade e Dissimilaridade

Em Análise de Agrupamento, para se construir um simples grupo a partir de um conjunto de elementos é necessário utilizar algum critério de proximidade ou tipo de medida que possibilite a comparação entre os componentes desse conjunto, tornando possível verificar se um dado elemento *A* é mais parecido com *B* do que com *C*. A utilização dessa medida ou critério fornecerá a distância dimensional entre os objetos, permitindo que se quantifique o quanto eles são parecidos. Essa medida é conhecida por Coeficiente de Parecença ou Coeficiente de Similaridade (ou Dissimilaridade), sendo essa última definição a mais adotada na presente dissertação[11]. Porém, em alguns momentos, chamaremos o coeficiente utilizado simplesmente de *distância* ou *medida de distância*.

Se o maior valor calculado, relativo a um coeficiente, for utilizado para indicar que dois objetos são parecidos, estaremos utilizando um Coeficiente de Similaridade; em contrapartida, se o menor valor para um coeficiente indicar que dois objetos são parecidos, estaremos trabalhando com um Coeficiente de Dissimilaridade. A Distância Euclidiana, um dos coeficientes mais conhecidos e utilizados, é um exemplo de coeficiente de dissimilaridade, enquanto que o Coeficiente de Correlação é um exemplo de coeficiente de similaridade. Determinados coeficientes se adaptam melhor a determinados tipos de variáveis e situações e em geral, os coeficientes de dissimilaridade são mais adequados para as variáveis quantitativas, e os de similaridade para as variáveis qualitativas.

---

[11] Muitos autores utilizam o termo Parecença ao invés de similaridade ou dissimilaridade e Medida ao invés de Coeficiente, e alguns utilizam Similaridade, não distinguindo entre dissimilaridade mesmo quando o coeficiente em questão o é. Na presente dissertação optou-se pela colocação sempre dos termos considerados os mais corretos, representativos e simples para a ocasião: Coeficiente de Similaridade e Coeficiente de Dissimilaridade.

Escolhido o coeficiente, torna-se possível construir a Matriz de Distâncias[12] de onde surgirão os grupos.

2.2.4.2 Matriz de Distâncias

A Matriz de Distâncias apresenta em cada célula o valor do coeficiente calculado para os elementos posicionados nas respectivas linha e coluna. Esse valor representa uma medida de distância entre esses dois elementos e dependendo da medida que foi escolhida, esta distância é considerada uma distância "verdadeira" como aquelas recomendadas por Johnson e Wichern (1988)[13]. Essa matriz é quadrada, possuindo dimensão máxima *n x n,* com *n* representando o número de elementos envolvidos. Possui sua diagonal principal nula (correspondente ao valor da distância de um elemento para ele mesmo), e seu triângulo superior vazio (em alguns livros, é o triângulo inferior que se apresenta vazio). A Tabela 2.3 exibe uma Matriz de Distâncias construída a partir das observações de cinco elementos:

	A	**B**	**C**	**D**	**E**
A	0				
B	2,142	0			
C	2,152	0,779	0		
D	1,568	2,576	2,517	0	
E	1,372	2,007	1,857	2,050	0

Tabela 2.3: Exemplo de Matriz de Distâncias

Na matriz da tabela anterior, observa-se que menos da metade dos elementos não estão preenchidos (correspondem ao triângulo superior), pois suas distâncias já foram calculadas em outras células, *dist(A,B) = dist(B,A)*. Uma outra forma de representação das

---

[12] É comum a substituição desse termo por Matriz de Semelhança ou Parecença, semelhante às definições dadas aos coeficientes (Bussab et al, 1990). Nesse trabalho optamos pela utilização de Matriz de Distâncias.

[13] Segundo esses mesmos autores, uma medida de distância é considerada "verdadeira" se satisfaz as seguintes propriedades para o agrupamento de objetos: i) d(A,B) = d(B,A); ii) d(A,B) > 0, se $A \neq B$; iii) d(A,B) = 0, se A=B; iv) $d(A,B) \leq d(A,C) + d(C,B)$ **,** conhecida por Desigualdade Triangular.

distâncias entre os elementos é através da matriz de distâncias reduzidas, que normalmente suprime a primeira linha e a última coluna da matriz original, como pode ser observado na Tabela 2.4:

	A	B	C	D
B	2,142			
C	2,152	0,779		
D	1,568	2,576	2,517	
E	1,372	2,007	1,857	2,050

Tabela 2.4: Exemplo de Matriz de Distâncias Reduzida

Da presente revisão, foi induzido que o número de células[14] preenchidas de uma matriz de distâncias, descontados os valores nulos existentes na diagonal principal, não ultrapassa $\frac{n(n-1)}{2}$, onde n é o número de objetos dessa matriz envolvidos na análise de agrupamento.

2.2.4.3 Dendrogramas

Dendrogramas são estruturas gráficas em forma de árvore, utilizadas para representar as junções (métodos hierárquicos) ou divisões (métodos de partição) que ocorreram a partir de valores provenientes da matriz de distâncias (JOHNSON & WICHERN, 1988). De acordo com Bussab *et al* (1990), para construirmos um dendrograma utilizando os valores da matriz de distâncias com o objetivo de ilustrar as junções, devemos colocar no eixo horizontal os elementos, em uma ordem conveniente de acordo com os grupos formados, de onde partirá de cada um desses elementos uma linha vertical até a altura correspondente ao nível (o valor da distância) em que ocorreu a junção (a um outro elemento ou grupo). Essa altura é marcada no

[14] Conhecer esse número foi importante para aumentar a eficiência da implementação computacional apresentada no Capítulo 4, ao diminuir à metade o tamanho dos vetores utilizados na programação.

eixo vertical. Os dendrogramas são utilizados, principalmente, para a observação dos saltos que ocorrem na formação dos grupos, buscando detectar a formação de grupos heterogêneos. Essas estruturas permitem também a identificação de quanto seria necessário consentir ou "relaxar" a definição de grupos homogêneos[15]. O número ideal de grupos também pode ser inferido do dendrograma. A Figura 2.4 exibe um exemplo de um dendrograma representando os grupos formados através das junções de 27 elementos.

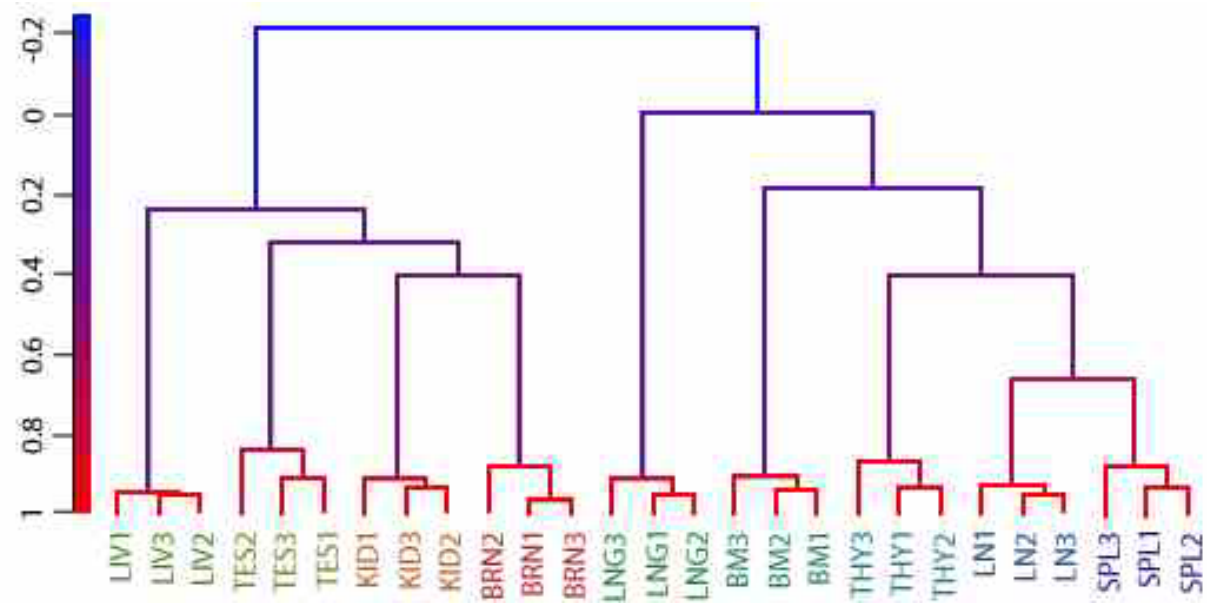


Figura 2.4 – Exemplo de Dendrograma Construído para os Grupos Formados por um Coeficiente de Correlação[16]

2.2.5 Etapas de Aplicação de uma Análise de Agrupamento

As subseções anteriores foram úteis para obtermos uma imagem sobre como ocorre um processo de agrupamento utilizando um simples diagrama de dispersão. Adicionando a essa experiência a literatura consultada (BUSSAB *et al*, 1990) e (BARROSO & ARTES,

[15] Isso está relacionado à escolha do patamar máximo de variação percentual, calculado a partir da formação de um grupo para o anterior. Esse patamar também é conhecido por *threshold.*

[16] A maioria das representações de dendrogramas pesquisadas apresenta o formato da Figura 7, apresentando a "raiz" da árvore na parte de cima do gráfico, entretanto, dendrogramas que invertem essa ordem ou que trocam os eixos também são comuns.

2003), é possível enumerar, nesse momento, as seguintes etapas a serem seguidas para a aplicação de análise de agrupamento:

1. Definição dos objetivos da Análise de Agrupamento, obtenção dos dados e tratamento, se necessário, dos mesmos.
2. Escolha da Técnica de Agrupamento e da medida de distância (coeficiente de similaridade ou dissimilaridade) a ser utilizada[17].
3. Formação dos grupos a partir das definições efetuadas no item anterior.
4. Validação, avaliação e interpretação dos resultados obtidos

Na primeira etapa são definidos os objetivos que conduzirão a aplicação da análise de agrupamento; são conhecidas as características das variáveis (se qualitativas ou quantitativas; discretas ou contínuas; nominais ou ordinais), suas escalas, e se os valores apresentados pelas mesmas terão que sofrer alguma relativização.

É na segunda etapa que são feitas as principais opções da Análise de Agrupamento. Não sendo um dos objetivos da análise a geração de um determinado número de grupos, opta-se, na maioria das vezes, pelas Técnicas Hierárquicas. Se um determinado número de grupos se faz necessário, a opção é utilizar as Técnicas de Partição e, em algumas vezes, utilizar as hierárquicas antes para confirmar se esse número predeterminado de grupos é possível. Feita a escolha da técnica, parte-se para a definição da medida de distância, representada por um coeficiente de similaridade ou dissimilaridade. Em geral, esses coeficientes são utilizados com a finalidade de moldar situações especiais de interesse do pesquisador, existindo uma série ampla dessas medidas.

---

[17] No Exemplo 1, a análise do diagrama de dispersão indicou que se uma análise de agrupamento fosse implementada, provavelmente seria utilizado um coeficiente de dissimilaridade como medida de distância entre os objetos. Nesse mesmo exemplo, a simples inspeção visual no gráfico de dispersão das variáveis padronizadas,

Na terceira etapa é comum a aplicação de diferentes algoritmos nos dados com o objetivo de "sentir" os grupos que são formados, contribuindo assim para a escolha do algoritmo mais adequado. Em muitas ocasiões, o pesquisador, devido ao conhecimento e experiência que possui em relação aos dados, desconfia de um número ideal de grupos ou até mesmo, em alguns casos, já sabe por antemão que somente existirão *p* grupos[18].

Busca-se, na construção de grupos homogêneos, garantir que os elementos de grupos diferentes apresentem comportamentos distintos, e aqueles que se encontram em um mesmo grupo apresentem comportamentos semelhantes. Para confirmar esses objetivos, a última etapa utiliza alguns procedimentos como Dendrogramas, Matrizes Cofenéticas, Gráficos, etc. e uma constante re-análise dos dados existentes que permitirão assegurar que os resultados produzidos poderão ser utilizados.

Na maioria das vezes, a implementação de uma análise de agrupamento é feita computacionalmente, o que nos dias de hoje tornou-se mais simplificado devido ao alto grau de desenvolvimento dos computadores pessoais, softwares e programas existentes. Portanto *n* execuções de análises de agrupamento para um mesmo conjunto de dados, utilizando diferentes algoritmos e até mesmo uma outra medida de distância, passa a ser viável. Sendo até mesmo aconselhável, pois isso permitirá a geração de alguns conjuntos de grupos homogêneos a serem fornecidos ao pesquisador, permitindo que o mesmo possa associar, com maior liberdade, suas percepções em relação aos elementos contidos nos grupos. A ocorrência de uma interação entre o pesquisador ("o principal interessado na construção dos grupos homogêneos") e o responsável pela análise de agrupamento, também é importante para a

---

permitiu a identificação de quatro grupos contendo uma ou mais divisões territoriais. Essa inspeção visual tornar-se-ia bastante complexa e de difícil execução, caso o número de elementos e variáveis fosse bem maior.

[18] Exemplo: os dados a serem analisados referem-se a características de espécimes de insetos; é sabido que existem somente três espécies, o que permitirá a existência de, no máximo, três grupos. Em outras situações o número de grupos será definido *a posteriori* com base nos resultados da análise.

definição de possíveis ajustes a serem feitos antes da execução final da Análise de Agrupamento.

2.2.6 Ilustrando uma Aplicação de Análise de Agrupamento

Antes da apresentação dos tipos de coeficientes que existem, das técnicas hierárquicas e de partição e das ferramentas de análise para confirmação dos resultados, estaremos apresentando um exemplo para mostrar a aplicação direta da análise de agrupamento em dados multivariados. As escolhas efetuadas nesse exemplo não serão objeto de maiores explicações, para não comprometerem as explanações a serem apresentadas logo após a presente subseção.

2.2.6.1 Análise de Agrupamento aplicada ao Índice de Desenvolvimento Humano (IDH)

O IDH é um índice calculado pelo Programa das Nações Unidas para o Desenvolvimento (PNUD), um órgão da Organização das Nações Unidas (ONU), com o objetivo de aferir o avanço do desenvolvimento de uma população, contrapondo-se ao Produto Interno Bruto (PIB), um outro indicador muito utilizado que considera apenas a dimensão econômica (PNUD, 2005).

O IDH é avaliado, para uma população, através das características sociais, culturais, políticas e econômicas que influenciam a qualidade da vida humana dessa população. Seu cálculo é gerado a partir de três indicadores básicos: Educação (taxa de matrícula bruta nos três níveis de ensino e taxa de alfabetização); renda per capita; e Longevidade (esperança de

vida ao nascer). O IDH, o qual varia entre 0 e 1, é montado com base em dados de dois anos anteriores[19].

A Tabela 2.5 a seguir (fornecida pelo IPEA) apresenta os índices IDH e as populações mais recentes (2002, referentes a 2000) dos estados da Região Sudeste[20].

Estado	População Total[21]	IDH
ES – Espírito Santo	3.097.232	0,765
MG – Minas Gerais	17.891.494	0,773
RJ - Rio de Janeiro	14.391.282	0,807
SP – São Paulo	37.032.403	0,82

Tabela 2.5: Índices IDH e Populações dos Estados da Região Sudeste

O objetivo da Análise de Agrupamento aplicada nos dados da tabela anterior é agrupar os estados (não importa o número de grupos) que se apresentam similares, conforme os valores mensurados para IDH e População. As escalas de População Total e IDH diferem-se bastante, portanto, se faz necessário uma relativização dos valores apresentados, caso contrário, qualquer coeficiente calculado será determinado quase que totalmente, apenas pelo valor da variável população. A Tabela 2.6 exibe essa relativização, presente nas variáveis $PopTotal_R$ e IDH_R, construída através da padronização estatística.

Estado	$PopTotal_R$	IDH_R
ES - Espírito Santo	-1,063	-0,993
MG - Minas Gerais	-0,015	-0,690
RJ - Rio de Janeiro	-0,263	0,596
SP - São Paulo	1,342	1,087

Tabela 2.6: Índices IDH e Populações Relativizados

[19] Para aferir a longevidade, o indicador utiliza números de expectativa de vida ao nascer. O item educação é avaliado pelo índice de analfabetismo e pela taxa de matrícula em todos os níveis de ensino. A renda é mensurada pelo PIB per capita em dólar PPC (paridade do poder de compra, que elimina as diferenças de custo de vida entre os países), sendo depois corrigido pelo poder de compra da moeda de cada país. Essas três dimensões têm a mesma importância no índice, que varia de zero a um.

[20] Fonte: http://www.ipeadata.gov.br.

[21] Essa variável apresenta os valores para a População residente total, que exclui os empregados domésticos e pensionistas de seu cálculo (IPEA, 2005).

O coeficiente escolhido para ser utilizado na avaliação da proximidade entre os estados, foi a Distância Euclidiana (DE), um coeficiente de dissimilaridade definido da seguinte forma:

Sejam $X_1, X_2, ..., X_n$ as variáveis p-dimensionadas tal que $X_i^T = (x_{i1}, x_{i2}, ..., x_{in})$. Assim, a Distância Euclidiana (DE) entre a r-ésima e s-ésima observação é dada por

$$DE(X_s, Y_r) = \sqrt{\sum_{j=1}^{p} (x_{rj} - y_{sj})^2} .$$

No caso particular desse exemplo tem-se *n = 4* (Estados) e *p = 2* (variáveis IDH_R e $PopTotal_R$). Utilizando a distância euclidiana para calcular a distância entre os estados ES e MG, temos:

$$DE(ES, MG) = \sqrt{(-1{,}063 + 0{,}015)^2 + (-0{,}993 + 0{,}690)^2} = 1{,}091$$

A partir da definição do coeficiente é construída a Matriz de Distâncias apresentada no Quadro I.

	Espírito Santo	**Minas Gerais**	**Rio de Janeiro**	**São Paulo**
Espírito Santo	0,000			
Minas Gerais	**1,091**	0,000		
Rio de Janeiro	1,779	1,310	0,000	
São Paulo	3,180	2,236	1,678	0,000

Quadro I – Matriz de Distâncias calculada através da Distância Euclidiana

O Quadro I nos fornece o primeiro grupo, identificado pelo menor valor existente na Matriz de Distâncias (utilizamos um coeficiente de dissimilaridade): **1,091.** Esse valor representa a distância euclidiana entre os estados Espírito Santo e Minas Gerais. Então o nosso primeiro grupo homogêneo produzido foi **(Espírito Santo, Minas Gerais)**, que passam a fazer parte da mesma coluna e/ou linha, conforme a re-arrumação adotada para a Matriz. O Quadro II apresenta essa nova matriz já com o primeiro agrupamento produzido.

	Espírito Santo, Minas Gerais	**Rio de Janeiro**
Rio de Janeiro	?	
São Paulo	?	1,678

Quadro II – Matriz de Distâncias a espera do Método Hierárquico

Porém, como iremos calcular a distância euclidiana do grupo (ES, MG) para os estados restantes (RJ e SP)? Isso é feito através da escolha de um algoritmo associado à técnica escolhida – nesse exemplo estaremos utilizando um dos algoritmos existentes nas técnicas hierárquicas, pois o número de grupos não é um pré-requisito. Escolhemos o Método das Médias das Distâncias (*Average Linkage)* – MMD, que define a distância entre dois grupos como a média das distâncias entre os elementos de um para os elementos do outro. As distâncias euclidianas dos elementos do primeiro grupo formado, (ES, MG), para o Rio de Janeiro são: *DE(ES,RJ)=1,779 e DE(MG, RJ)=1,310.* Logo, a distância (D_{MMD}) entre o grupo formado para o Rio de Janeiro, seguindo o Método das Médias das Distâncias será:

$$D_{MMD}((ES,MG), RJ) = (1,779 + 1,310)/2 = 1,545.$$

Aplicando o MMD para ao estado restante, torna-se possível gerar uma nova matriz de distâncias, apresentada no Quadro III.

	Espírito Santo, Minas Gerais	**Rio de Janeiro**
Rio de Janeiro	**1,545**	
São Paulo	2,708	1,678

Quadro III – Matriz de Distâncias calculada após a definição do Método Hierárquico

O que nos permite identificar o próximo grupo, **(Espírito Santo, Minas Gerais, Rio de Janeiro)**, e re-arrumar a matriz de distâncias (utilizando o MMD), no Quadro IV.

	Espírito Santo, Minas Gerais, Rio de Janeiro
São Paulo	2,365

Quadro IV – Matriz de Distâncias para o Grupo Final

O valor **2,365** representa a distância onde todos os elementos, os estados, que participaram da Análise de Agrupamento se reúnem[22]. Cada uma das etapas da Análise de Agrupamento apresentaram os seguintes grupos homogêneos (a princípio):

(Espírito Santo, Minas Gerais), (Rio de Janeiro) e (São Paulo);

(Espírito Santo, Minas Gerais, Rio de Janeiro) e (São Paulo);

(Espírito Santo, Minas Gerais, Rio de Janeiro, São Paulo)

O Quadro V exibe esses mesmos grupos de uma forma mais organizada, comportando as distâncias utilizadas na formação dos mesmos e a diferença (ou variação) percentual para a distância da formação do grupo anterior.

Grupo	Elementos	Distância	Diferença (%)
1	(Espírito Santo, Minas Gerais), (Rio de Janeiro), (São Paulo)	1,091	-
2	(Espírito Santo, Minas Gerais, Rio de Janeiro), (São Paulo)	1,545	41,6
3	(Espírito Santo, Minas Gerais, Rio de Janeiro, São Paulo)	2,365	53,1

Quadro V – Grupos gerados e respectivas distâncias e diferenças percentuais

Observamos que as diferença percentuais dos grupos 2 e 3 para a formação do grupo anterior (42% e 53% aproximadamente), parecem bastante altas. Isso pode ser confirmado, ou não, pela análise dos saltos ocorridos no Dendrograma da Figura 2.5.

[22] É característica das técnicas hierárquicas reunir, ao final da sua aplicação, todos os elementos em um único grupo. Por conta dessa característica, essas técnicas também são chamadas de técnicas aglomerativas.

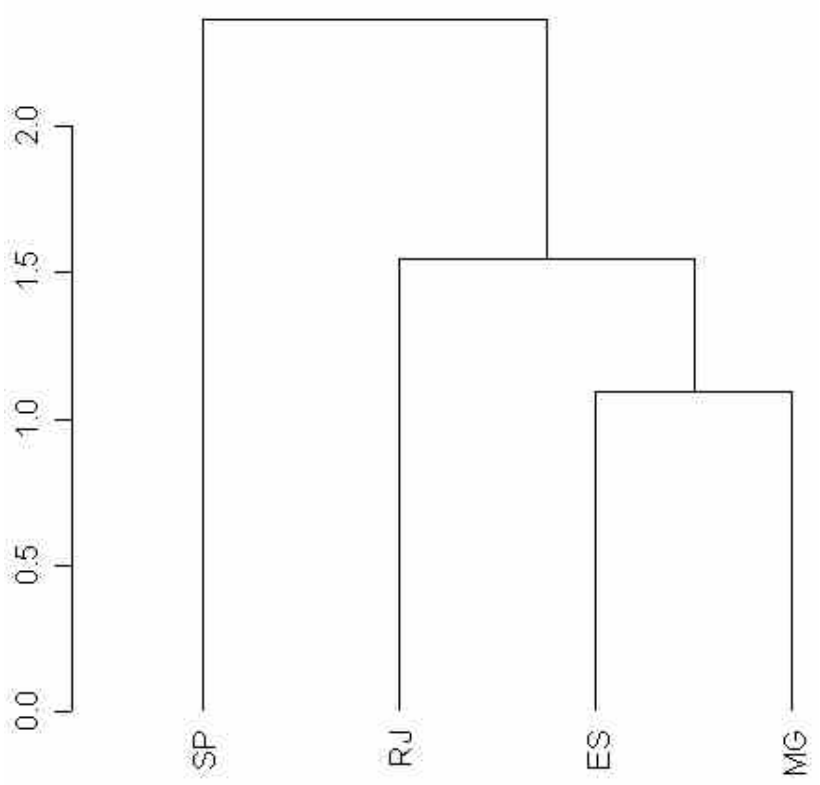


Figura 2.5 – Dendrograma da aplicação do MMD aos dados da Matriz de Distâncias do Quadro I

Observando o dendrograma da Figura 2.5, detectamos um pequeno salto, em relação ao grupo anterior, na formação do grupo (RJ, ES, MG). O grupo (ES, MG), por ser o primeiro grupo formado, é o mais homogêneo entre os três, o que é uma característica dos coeficientes de dissimilaridade, onde um valor 0 calculado para esse coeficiente significaria igualdade entre os elementos. Dependendo da medida trabalhada isso não se traduzirá em poder afirmar que os valores das variáveis desses elementos são iguais. Nas medidas de distância "verdadeiras", como a distância euclidiana, um valor 0 significaria igualdade entre esses valores.

Ainda pelo Dendrograma, observamos que o salto na formação do último grupo (que gerou uma variação de 53% em relação ao anterior) provavelmente indica a formação de um grupo heterogêneo (o fato de ser este o último grupo também influi nessa conclusão). Ao

considerarmos isso, os grupos homogêneos identificados passam a ser expressos apenas pelos dois conjuntos de grupos:

(Espírito Santo, Minas Gerais), (RJ) e (SP);

(Espírito Santo, Minas Gerais, Rio de Janeiro) e (SP).

Nesse momento encerramos o presente exemplo considerando que, através dele, foi possível ilustrar o funcionamento de uma Análise de Agrupamento.

Os próximos passos seriam o cálculo do Coeficiente de Correlação Cofenética e construção de um gráfico para exibir a variação percentual de cada agrupamento formado.

2.2.7 Coeficientes de Similaridade e Dissimilaridade

De acordo com Johnson & Wichern (1988), a maioria dos esforços em produzir um simples grupo a partir de uma estrutura de dados complexa, necessariamente, irá requerer a existência de uma medida de "proximidade" ou "similaridade". Sua definição é fundamental para a utilização das técnicas de Análise de Agrupamento pois, é através dela, que se definem os critérios que avaliam se dois pontos farão parte de um mesmo grupo ou não. Como visto anteriormente, essa medida é dividida em duas categorias: Coeficientes de Similaridade e Coeficientes de Dissimilaridade. De um modo geral, é possível construir uma medida de dissimilaridade a partir de uma de similaridade, e vice-versa.

2.2.7.1 Coeficientes Utilizados para Variáveis Quantitativas

Toda aquela característica ou evento da natureza que apresente mais de uma realização possível poderá ser representada através de uma variável, a qual se refere,

convencionalmente, ao conjunto de resultados possíveis de um fenômeno. As variáveis são classificadas, dicotomicamente, em *Qualitativas* ou *Quantitativas.* Estaremos apresentando nessa subseção, os coeficientes utilizados para variáveis quantitativas.

Uma variável Quantitativa é toda aquela que apresenta como possíveis realizações números oriundos de uma contagem, de uma medição, de uma apuração. São exemplos dessas variáveis os valores referentes a: Hectares de uma fazenda; Idade geológica do continente africano; Total de votos nulos em uma eleição; valores na Escala Richter; etc. As variáveis quantitativas podem ser *Discretas* ou *Contínuas*. Uma variável quantitativa é considerada discreta se o conjunto de valores que ela pode assumir for um conjunto finito ou infinito enumerável. Uma variável quantitativa é considerada contínua se o conjunto de valores que ela pode assumir for um conjunto infinito não-enumerável (BUSSAB & MORETTIN, 2003).

Os coeficientes ou medidas de dissimilaridade são as distâncias mais adequadas quando se trabalha com variáveis quantitativas em Análise de Agrupamento, o que não impede a utilização de coeficientes de similaridade para essas variáveis. Os coeficientes mais adequados ao presente trabalho, identificados na literatura consultada, são apresentados a seguir. Todos eles referem-se a variáveis p-dimensionais $X_1, X_2, ..., X_n$ tal que $X_i^{\,T} = (x_{i1}, x_{i2}, ..., x_{in})$.

Todos eles referem-se a duas observações *p*-dimensionais: $x = [x_1, x_2, ..., x_p]'$ e $y = [y_1, y_2, ..., y_p]'$.

1) Distância Euclidiana (*Euclidean Distance*).

Este é o coeficiente de dissimilaridade mais conhecido e utilizado para indicar a proximidade entre objetos. É, simplesmente, a distância geométrica entre dois objetos em um

espaço multidimensional. A idéia básica é considerar cada observação como um ponto em um espaço euclidiano e, desse modo, calcular o coeficiente que representará a distância física entre os pontos. A fórmula da Distância Euclidiana (DE) foi definida na sub-seção 2.2.6.1.

A Distância Euclidiana, e suas derivadas, podem ser calculadas tanto para os dados puros, crus, quanto para os dados relativizados. Essas distâncias possuem certas vantagens, são elas: apresentam simplicidade de cálculo; a distância entre quaisquer dois objetos não é afetada pela inserção de novos objetos na análise, mesmo que os novos valores sejam classificados como valores atípicos; encontram-se associadas, intuitivamente, ao conceito usual de distância. Entretanto, essas distâncias podem ser bastante afetadas pelas diferenças de escala associadas às dimensões, a partir das quais as distâncias são computadas, o que implicaria na necessidade de relativizar os dados antes da aplicação desse coeficiente[23].

Ao elevarmos a distância Euclidiana ao quadrado, é gerada a Distância Euclidiana Quadrática, sendo essa medida bastante influenciada por aqueles objetos que se encontram mais distantes.

2) Distância Euclidiana Quadrática (DEQ).

A fórmula da Distância Euclidiana Quadrática é expressa por:

$$DEQ(X_s, Y_r) = (x_{s1} - y_{r1})^2 + (x_{s2} - y_{r2})^2 + \ldots + (x_{sp} - y_{rp})^2 = \sum_{i=1}^{p} (x_{ri} - y_{si})^2$$

Uma outra medida derivada da distância Euclidiana muito utilizada é a **Distância Euclidiana Média (DEM)**, expressa pela raiz quadrada da divisão entre o somatório do

quadrado das diferenças, pelo número de variáveis envolvidas. Ela mantém-se robusta mesmo na ausência de dados de algumas variáveis.

3) Distância Euclidiana Média (*Mean Euclidean Distance*):

$$DE(X_s, Y_r) = \sqrt{\sum_{j=1}^{p} (x_{rj} - y_{sj})^2 / p}$$

Distância Manhattan (DMAN) ou Distância Quarteirão é uma outra medida utilizada em análise de agrupamento. Ela é calculada através do somatório do valor absoluto das diferenças entre as observações, e costuma suprimir grandes diferenças que venham a existir devido à presença de *outliers*.

A apresentação dessa medida, a seguir, é aproveitada para introduzir a utilização das ponderações para os pares de variáveis, que podem ser utilizadas em todos os coeficientes apresentados na presente seção.

4) Distância Manhattan ou **Distância Quarteirão.**

$$DMAN(X_s, Y_r) = w_1 \mid x_{r1} - y_{s1} \mid + w_2 \mid x_{r2} - y_{s2} \mid + \ldots + w_p \mid x_{rp} - y_{sp} \mid$$

Os $w_{i's}$, com *i* variando de 1 a *p,* da medida anterior, representam os pesos que podem ser dados a cada dupla de variáveis, sendo normalmente utilizada a equiponderação, com $w_i = 1$ (ou utiliza-se a média, com $w_i = 1/p$).

---

[23] Se uma das dimensões denota um tamanho de medida em centímetros, e é feita uma conversão para milímetros, o resultado da distância Euclidiana (calculada com as múltiplas dimensões) poderá ser bastante afetado e, consequentemente, os resultados provenientes das análises dos grupos serão diferentes.

5) Distância Chebyshev (DCHBY)

Essa distância apresenta o valor absoluto da máxima diferença existente entre as variáveis multidimensionais de dois elementos.

$$DCHBY\ (X_s, Y_r) = MAX\ \{|\ x_{ri} - y_{si}\ |,\ i = 1,...,\ p\}$$

Além das distâncias anteriores, existem outras dentre as quais podemos citar a distância de Mahalanobis, distância de Minkowsky (de onde são derivadas as distâncias 1 a 4), o Coeficiente de Gower, o Coeficiente de Similaridade de Cattel e os Coeficientes para Variáveis Positivas (BUSSAB, 1990).

A Tabela 2.7 a seguir, exibe uma amostra reduzida de informações extraídas do Log de uma plataforma EAD, referente à utilização desta por alunos em um dia específico. Os dados contidos nessa tabela serão utilizados para o cálculo e comparação das cinco distâncias apresentadas nessa subseção.

ALUNO	ACESSOS (A)	MINUTOS DEDICADOS (MD)	ACESSOS (W_A)	MINUTOS DEDICADOS (W_{MD})
A	**1**	**36,1**	**0**	**0,27**
B	**3**	**21,8**	**1**	**0,13**
C	**3**	**21,2**	**1**	**0,13**
D	**3**	**18,5**	**1**	**0,10**
E	**1**	**8,3**	**0**	**0**
F	**2**	**110,4**	**0,5**	**1**
MÍNIMO	1	8,3	0	0
MÁXIMO	3	110,4	1	1

Tabela 2.7: Variáveis Acessos e Minutos Dedicados

Serão utilizados os valores das variáveis ACESSOS (WA) e MINUTOS DEDICADOS (W_{MD}), os quais foram relativizados a partir das variáveis ACESSOS e MINUTOS DEDICADOS, através da seguinte transformação:

$W = \frac{X - \min(X)}{\max(X) - \min(X)}; 0 \leq W \leq 1$, onde *min(X)* é o menor dos valores da variável X, e *max(X)* é o maior valor.

Essa relativização transformou os dados através da divisão efetuada entre o desvio para o menor valor, o numerador, e a amplitude dos dados, o denominador. Essa transformação[24] pode ser aplicada em variáveis discretas e contínuas. A partir dos novos valores, temos as seguintes distâncias:

Distância Euclidiana (DE):

$$DE(A,B) = \sqrt{(0-1)^2 + (0{,}27 - 0{,}13)^2} = \sqrt{1 + 0{,}0196} = 1{,}01$$

Distância Euclidiana Quadrática (DEQ):

$$DEQ(A,B) = (0-1)^2 + (0{,}27 - 0{,}13)^2 = 1 + 0{,}0196 = 1{,}02$$

Distância Euclidiana Média (DEM):

$$DEM(A,B) = \sqrt{[(0-1)^2 + (0{,}27 - 0{,}13)^2]/2} = \sqrt{(1{,}0196)/2} = 0{,}71$$

Distância Manhattan (DMAN):

$$DMAN(A,B) = |0-1| + |0{,}27 - 0{,}13| = 1{,}14$$

Distância Chebyshev (DCHBY):

$$DCHBY(A,B) = MAX\,|A-B| = |0-1| = 1$$

As quatro distâncias calculadas acima podem ser comparadas na Tabela 2.8, com o intuito de escolher qual distância será utilizada em uma possível aplicação de Análise de Agrupamento.

---

[24] Uma conhecida relativização é a que divide os valores apresentados pelas suas médias (BUSSAB *et al*, 1990).

Distância	DE	DEQ	DEM	DMAN	DCHBY
Valor Calculado	1,01	1,02	0,71	1,14	0,9

Tabela 2.8: Comparativo das Distâncias calculadas para os dois alunos escolhidos

Apenas na distância DMAN a variável *Acessos* não é responsável pela quase totalidade da distância calculada. Se for um dos objetivos da Análise de Agrupamento levar em conta a contribuição da variável *Minutos Dedicados*, a Distância de Manhattan passa a ser a mais indicada para ser utilizada na construção da Matriz de Distâncias.

2.2.7.2 Coeficientes e Distâncias mais Comuns para Variáveis Qualitativas

Uma variável é considerada qualitativa quando apresenta como possíveis valores uma qualidade (ou atributo) do indivíduo pesquisado. Dividem-se ainda em nominais, cujos valores não possuem uma ordenação, e ordinais, onde é possível identificar uma ordem em seus resultados (BUSSAB & MORETTIN, 2003).

Na área das Ciências Sociais, é freqüente a utilização de técnicas estatísticas para a análise de variáveis qualitativas. A Análise de Agrupamento aplica as técnicas nos dados qualitativos para gerar outros dados, ou seja, ela utiliza as variáveis qualitativas (e quantitativas) para subsidiar seus objetivos que se apresentam focalizados na identificação daqueles elementos com um mesmo comportamento no conjunto de variáveis analisadas.

Bussab *et al* (1990) dividem os coeficientes para as variáveis qualitativas em Coeficientes para Variáveis Qualitativas Nominais e Coeficientes para Variáveis Qualitativas Ordinais.

Devido à possibilidade da transformação das variáveis nominais e ordinais em variáveis dicotômicas binárias, serão mostradas, nessa revisão, as transformações usuais e as principais medidas para esse tipo de variável.

2.2.7.2.1 Coeficientes para Variáveis Qualitativas Dicotômicas ou Binárias

São consideradas variáveis dicotômicas aquelas que assumem somente um valor (nominal ou ordinal), de dois valores possíveis. Exemplo: O fator RH do sangue de um indivíduo ou é positivo ou é negativo. Quando os possíveis valores de uma variável dicotômica são substituídos pelos números 1 e 0, com o intuito de facilitar análises e a construção de gráficos, elas passam a ser chamadas variáveis binárias. É comum na Estatística a associação do termo SUCESSO para o valor 1 da variável binária, e FRACASSO para o valor 0.

Os coeficientes para essas variáveis normalmente se concentram em medir a similaridade, baseados na contagem das concordâncias (positivas ou negativas), existentes entre os elementos. Poucos são os coeficientes que utilizam como principal elemento de sua medida o número de discordâncias. Segundo Bussab *et al* (1990), os coeficientes que existem para as variáveis qualitativas surgiram das tabelas de contingência ou de dupla entrada.

Por exemplo, a Tabela 2.9 apresenta qual foi ação recomendada de dois avaliadores em relação à leitura (1) ou não (0), de um determinado texto sobre Educação a Distância.

Título do Texto	Avaliador Barra	Avaliador Ilha
Uma Abordagem Multiagente	0	0
Design for Learning	0	1
Contribuições de Conceitos de Comunicação	1	1
Hipermídias Distribuídas e Educação	0	1
Interface de Ambientes Educacionais	1	1
JAVAL? Ambiente para Avaliações Remotas	1	0
Teaching and Learning with Telematics	0	1
Tecnologias de Informação Aplicadas à Educação: construindo uma rede de desempenho em um grupo	1	1
Uma Máquina de Estados Finitos	0	1
Educação a Distância – Padrões para Projetos de Sistemas	1	1

Tabela 2.9: Ação Recomendada dos avaliadores Barra e Ilha

A Tabela 2.10 de dupla entrada (ou de contingência), busca apurar o número de ocorrência dos pares (1,1) – ambos os avaliadores recomendam a leitura; (0,0) – ambos os avaliadores não recomendam a leitura; (1,0) – somente o avaliador Barra recomenda a leitura; (0,1) – somente o avaliador Ilha recomenda a leitura.

		Avaliador Barra		Total
		1	**0**	
Avaliador Ilha	**1**	4 a	4 b	5 a+b
	0	1 c	1 d	5 c+d
Total		5 a+c	5 b+d	**10**

Tabela 2.10: Dupla Entrada para Ação Recomendada dos Avaliadores Barra e Ilha

O par (1,1) é representado na literatura por *a*, e corresponde ao total de concordâncias positivas ocorridas entre os dois elementos, no caso, os avaliadores. O par (0,0), por sua vez, é o total de concordâncias negativas e é representado por *d.* O par (1,0) é *b* e o par (0,1), é *c,* esses dois pares representam as discordâncias entre os avaliadores.

Uma análise superficial dos dados da tabela de dupla entrada anterior, mostra que os avaliadores Barra e Ilha concordaram várias vezes em recomendar um determinado texto, mas quase nunca concordaram na não recomendação de um texto para leitura. Suas avaliações são semelhantes? Como verificar se realmente existe essa semelhança? Através de algum coeficiente que estará medindo a proximidade das variáveis qualitativas.

O Quadro VI a seguir, compilado a partir de Bussab *et al* (1990) e Gower (1985), fornece quase todos os coeficientes existentes e suas características principais. Vários desses coeficientes não são aplicáveis quando suas variáveis geram uma indefinição matemática por conta de valores nulos:

Coeficiente	**Expressão**	**Intervalo**	**Características**
Distância Binária de Sokal	$\sqrt{\frac{b+c}{a+b+c+d}}$	[0,1]	Fornece a proporção de discordâncias nos dois elementos. É um Coeficiente de Dissimilaridade.
Concordância Simples – Matching (Sokal e Michener, 1958)	$\frac{a+d}{a+b+c+d}$	[0,1]	É a proporção de concordâncias (positivas e negativas) entre os elementos. Similaridade.
Jaccard (1908)	$\frac{a}{a+b+c}$	[0,1]	É a proporção de concordâncias positivas entre os elementos. Similaridade.
Rogers e Tanimoto (1960)	$\frac{a+d}{a+2(b+c)+d}$	[0,1]	Similaridade. Oposto ao coeficiente de Sneath e Sokal, proporciona peso 2 para o total de discordâncias. Semelhante ao coeficiente *antiDice*.
Sneath e Sokal (1962)	$\frac{2(a+d)}{2(a+d)+b+c}$	[0,1]	Similar ao coeficiente de concordância simples diferindo apenas por fornecer peso 2 à soma das concordâncias. Similaridade.
Russel e Rao (1940) Concordâncias Positivas	$\frac{a}{a+b+c+d}$	[0,1]	Fornece a proporção das concordâncias positivas. Similaridade.
Ochiai (1957)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0,1]	Similaridade. Considera as concordâncias positivas.
Baron-Urbani-Buser	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	[0,1]	É uma outra visão para a proporção de concordâncias (positivas e negativas). Similaridade.
Hamann (1961)	$\frac{(a+d)-(b+c)}{a+b+c+d}$	[-1,1]	É o total de concordâncias menos o total de discordâncias dividido pelo total. Varia de –1, total discordância, até 1, concordância perfeita. O dobro do coeficiente de concordância simples menos um. Similaridade.

Yule (1950)	$\frac{ad-bc}{ad+bc}$	[-1,1]	Similaridade. Assumirá 1 quando b+c=0, significando total concordância. Assumirá –1 quando a+d=0, significando completa discordância. Porém, quando ad-bc = 0, assumirá 0 não permitindo uma análise.
Dice (1945) Czekanowski (1932) Sørensen (1948)	$\frac{2a}{2a+b+c}$	[0,1]	Similar ao Jaccard, porém permitindo peso 2 para concordâncias. Caso as variáveis envolvidas sejam nulas é indefinido e portanto não recomendado. Similaridade.
Kulczynski I	$\frac{a}{b+c}$	$[0,+\infty]$	
Kulczynski II	$\frac{\left(\frac{a}{a+b}+\frac{a}{a+c}\right)}{2}$	[0,1]	Similaridade. Os denominadores (a+b) ou (a+c) não podem ser nulos.
antiDice (Andeberg, 1973)	$\frac{a}{a+2(b+c)}$	[0,1]	Similaridade. Não considera as concordâncias negativas.
Gower2 (1985)	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	[0,1]	Similaridade. Considera as concordâncias positivas e negativas. Esse coeficiente também é associado a Ochiai na literatura consultada.
Anderberg	$\frac{\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{c+d}+\frac{d}{b+d}\right)}{4}$	[0,1]	Similaridade. Considera as concordâncias positivas e negativas.
Pearson (ϕ)	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	[-1,1]	Esse coeficiente mede a força das concordâncias em relação às discordâncias. Quanto mais próximo de 1, maior será a similaridade entre os elementos quando da concordância. Quanto mais próximo de –1, maior semelhança em relação à discordância.
McConnaughy	$\frac{a^2-bc}{(a+b)(a+c)}$	[-1,1]	É um coeficiente de similaridade. Considera, principalmente, as concordâncias negativas, representada pelo par (0,0), medindo a força das concordâncias em relação às discordâncias entre os elementos que estão sendo avaliados.

Quadro VI – Coeficientes de Semelhança para Variáveis Binárias

A aplicação do coeficiente de Jaccard para avaliar as concordâncias positivas, na Tabela 2.10, nos forneceria o seguinte valor:

$d(AvaliadorBarra, AvaliadorIlha) = \frac{a}{a+b+c} = \frac{4}{6} = 0{,}67$. Ao avaliarmos o que representa o valor 0,67 para esse coeficiente de similaridade, podemos concluir que os dois avaliadores apresentam avaliações semelhantes ao recomendarem um determinado artigo para leitura.

Se por outro lado utilizássemos o coeficiente de Concordância Simples, o qual leva em conta as concordâncias existentes, teríamos:

$d(AvaliadorBarra, AvaliadorIlha) = \frac{a+d}{a+b+c+d} = \frac{5}{10} = 0{,}50$. Esse valor não nos permitiria afirmar sobre uma possível semelhança entre os dois avaliadores.

Já o coeficiente de dissimilaridade Distância Binária de Sokal nos forneceria:

$d(AvaliadorBarra, AvaliadorIlha) = \sqrt{\frac{b+c}{a+b+c+d}} = \sqrt{\frac{2}{10}} = \sqrt{0{,}20} = 0{,}45$; o que nos permite concluir que os Avaliadores são um pouco semelhantes ao discordarem um do outro quanto à recomendação do artigo.

Segundo Clifford & Stephenson (1975), a escolha dos coeficientes restritos ao intervalo [0,1] é mais adequada, pois os índices que tendem ao infinito são sensíveis a pequenas mudanças, especialmente em *a*.

Os coeficientes de Yule, Pearson e McConnaughy são considerados coeficientes de associação, pois medem quanto os elementos concordam, positiva ou negativamente. Quanto mais próximo de 1, mais semelhantes serão os valores gerados pelos elementos.

2.2.7.2.2 Transformando Variáveis Qualitativas Nominais

Quando uma variável qualitativa possui mais de dois níveis é possível transformá-la em uma binária através da criação de variáveis fictícias. Seja y' uma variável qualitativa, formada por um vetor de variáveis qualitativas nominais:

$y' = (y_1, y_2, ..., y_n)$; onde o i-ésimo componente assume l_i níveis codificados da seguinte forma: $y_i = j$, com $j = 1,2,..., l_i$.

Ao transformar essa variável em uma variável binária, cada componente (critério qualitativo) contribuirá para a geração de l_i variáveis binárias $x_k\ (i)$ tal que

$$x_k(i) = \begin{cases} 1 & \text{se } y_i = k, \\ 0 & \text{caso contrário} \end{cases}$$

Assim o vetor y de dimensão n, é transformado no vetor x de dimensão p, formado apenas por componentes binários, e por conseqüência, y' se transforma em x':

$$y' = (y_1, y_2, ..., y_n) \rightarrow x' = (0,...,1...,0;.....;0,...,1,...,0)$$

O exemplo 3 e a Tabela 2.11 mostram uma transformação desse tipo.

Título do Texto	Critério	Avaliador Barra	Avaliador Ilha
Uma Abordagem Multiagente	Originalidade	4	1
	Relevância	3	2
	Ação Recomendada	4	4

Tabela 2.11: Originalidade, Relevância e Ação Recomendada dos avaliadores

Exemplo 3: A Tabela 2.11 exibe as avaliações do artigo *Uma Abordagem Multiagente*, geradas por Barra e Ilha, através dos critérios representados pelas variáveis qualitativas nominais Originalidade, Relevância e Ação Recomendada. Assumimos, nesse exemplo, que as duas primeiras variáveis possuem 5 níveis cada, e a última possui 4 níveis. A partir das características desses avaliadores, deseja-se medir a semelhança entre ambos:

Então podemos associar a cada avaliador os seguintes vetores:

$$y'(AvaliadorBarra) = (4,3,4)$$
$$y'(AvaliadorIlha) = (1,2,4)$$

O que nos permite gerar as seguintes variáveis binárias:

$$x'(AvaliadorB\ arra) = (0,0,0,1,0;\quad 0,0,1,0,0;\quad 0,0,0,1)$$

e

$$x'(AvaliadorI\ lha) = (1,0,0,0,0;\quad 0,1,0,0,0;\quad 0,0,0,1)$$

Esses vetores de variáveis binárias nos permitem gerar uma tabela de dupla entrada, exibida na Tabela 2.12.

		Avaliador Barra		**Total**
		1	**0**	
Avaliador Ilha	**1**	1 (a)	2 (b)	3 (a+b)
	0	2 (c)	9 (d)	11 (c+d)
Total		3 (a+c)	11 (b+d)	14

Tabela 2.12: Dupla Entrada para três variáveis qualitativas nominais geradas pelos Avaliadores Barra e Ilha

Utilizando os mesmos coeficientes do exemplo 3 (Jaccard, Concordância Simples e Distância Binária de Sokal), temos os seguintes resultados:

JACCARD: $d(AvaliadorBarra, AvaliadorIlha) = \frac{a}{a+b+c} = \frac{1}{5} = 0{,}20$. Esse valor não indica existir semelhança entre os dois avaliadores, quando da recomendação positiva do texto.

Concordância Simples: $d(AvaliadorBarra, AvaliadorIlha) = \frac{a+d}{a+b+c+d} = \frac{10}{14} = 0{,}71$.

Esse valor traduz semelhança entre os dois avaliadores quando das concordâncias, porém esse valor pode estar poluído (enviesado) devido ao elevado número de coincidências

de zeros, já esperado por conta do tipo de transformação. Esse coeficiente não seria adequado para variáveis binárias oriundas de qualitativas de mais de dois níveis.

Distância Binária de Sokal:

$$d(AvaliadorBarra, AvaliadorIlha) = \sqrt{\frac{b+c}{a+b+c+d}} = \sqrt{\frac{4}{14}} = \sqrt{0{,}29} = 0{,}54$$

Esse coeficiente nos permite concluir que os avaliadores, provavelmente, não são semelhantes ao discordarem um do outro quanto à avaliação feita sobre o texto. Porém, novamente, o total de concordâncias negativas (nove ao todo) influencia o resultado. Caso essa distância desconsiderasse esse total, a conclusão seria que os dois avaliadores são semelhantes sim quando discordam, ou seja, quando um discorda o outro concorda, e vice-versa.

Devemos então utilizar pelo menos um coeficiente que não envolva o total de ocorrência do par (0,0), pois o valor desse par costuma ser, significativamente, maior que os outros quando é executada uma transformação desse tipo em variáveis qualitativas nominais.

2.2.8 Técnicas Existentes para a Formação e Avaliação dos Grupos

Em Análise de Agrupamento existem duas grandes famílias de técnicas que são utilizadas para a formação dos grupos: Hierárquicas e de Partição[25]. Essas técnicas se distinguem, principalmente, pela metodologia utilizada na construção desses grupos. A escolha de um determinado método dessas técnicas, exige o conhecimento das propriedades desse particular algoritmo aliado aos objetivos da pesquisa (BUSSAB *et al,* 1990).

---

[25] Alguns livros dividem as técnicas hierárquicas em métodos aglomerativos e métodos divisivos (de partição), enquanto outros não fazem essa distinção, apresentando todos os métodos como sendo os que existem em Análise de Agrupamento. Seguimos, nesse trabalho, a organização dada em Bussab *et al* (1990).

Estaremos mostrando na presente revisão, principalmente, as técnicas hierárquicas, bastante utilizadas nos estudos de caso desenvolvidos e as mais adequadas para o contexto do trabalho. Iniciaremos mostrando como essas técnicas funcionam, em linhas gerais, para depois apresentarmos um exemplo desse funcionamento, se adequado.

Em seguida, mostraremos como é calculado o Coeficiente de Correlação Cofenético a partir da Matriz Cofenética, cuja construção também será exibida. Esses dois elementos fazem parte da avaliação, escolhida para o presente trabalho, que foi adotada para a confirmação do processo de Análise de Agrupamento. O Coeficiente Cofenético é utilizado pelas técnicas hierárquicas e de partição.

Finalizando a subseção, serão comentadas algumas técnicas de partição, pois a implementação computacional da interface, apresentada no capítulo 4, estará sendo modificada futuramente para comportar uma delas.

2.2.8.1 Técnicas Hierárquicas para Análise de Agrupamento

Os algoritmos das técnicas hierárquicas são utilizados no momento de reconstrução da Matriz de Distâncias, logo após a formação de um grupo (que ocorre de acordo com o coeficiente adotado). Esses algoritmos contribuem para a formação dos grupos da seguinte forma: cada um dos elementos que participarão do processo de Análise de Agrupamento são considerados como um grupo, que estarão se juntando a um outro elemento ou a um outro grupo, de acordo com os valores existentes na Matriz de Distâncias (que é atualizada a cada agrupamento que surge). Ao fim desse processo, todos os elementos estarão reunidos em um único grupo. A Figura 2.6 ilustra o andamento de um desses algoritmos aplicado nos dados de cinco elementos.

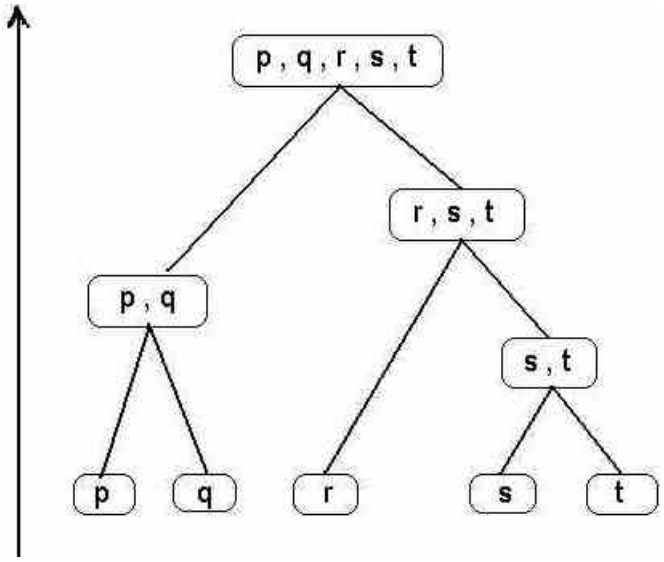


Figura 2.6 – Comportamento de uma Aplicação de um Algoritmo Hierárquico

Na figura anterior, a direção da seta indica o sentido da formação dos grupos. Devemos notar a semelhança dessa figura com os dendrogramas.

É característica das técnicas hierárquicas não utilizar o número de grupos a serem formados (caso se saiba ou se desconfie desse número) e, por conta disso, essas técnicas são as mais indicadas para a fase inicial e exploratória da pesquisa, onde se buscam informações e características associadas a uma população.

As técnicas hierárquicas mais indicadas para o nosso contexto, resumem-se a cinco métodos:

- Método da Ligação Simples (*Single Linkage*)
- Método das Médias das Distâncias (*Average Linkage*)
- Método da Ligação Completa (*Complete Linkage*)
- Método da Centróide (*Centroid Distance)*
- Método de Ward

2.2.8.1.1 Método da Ligação Simples (*Single Linkage*)

Esse método também é conhecido como Método do Vizinho mais Próximo ou da Distância Mínima. Quando aplicado para fornecer a distância entre um grupo (conjunto de elementos) e outro, seleciona a distância correspondente a **maior semelhança** entre os elementos de grupos distintos. Se o coeficiente, conforme definições exibidas na subseção 2.2.4.1, utilizado na construção da Matriz de Distâncias for um coeficiente de similaridade, essa maior semelhança será representada pela maior distância existente entre esses elementos, caso o coeficiente for de dissimilaridade, essa maior semelhança será a menor distância existente. Por exemplo, se na seção 2.2.6.1 estivéssemos utilizando o Método da Distância Mínima para re-arrumar a matriz de distâncias (criada a partir de um coeficiente de dissimilaridade), ao calcular a distância entre o primeiro grupo formado, Espírito Santo e Minas Gerais, e Rio de Janeiro, escolheríamos a menor distância existente entre esses elementos, ou seja, sabendo que *DE(ES,RJ)=1,779 e DE(MG,RJ)=1,310,* a distância calculada através do Método da Ligação Simples, D_{MLS}, seria $D_{MLS} = min \{DE(ES,RJ); DE(MG, RJ)\} = 1,310$.

A Figura 2.7 ilustra o Método da Ligação Simples aplicado a dois grupos formados através de um coeficiente de dissimilaridade. A reta ligando os dois pontos mais próximos indica a distância (no caso, mínima) que foi utilizada para indicar a maior semelhança entre os grupos.

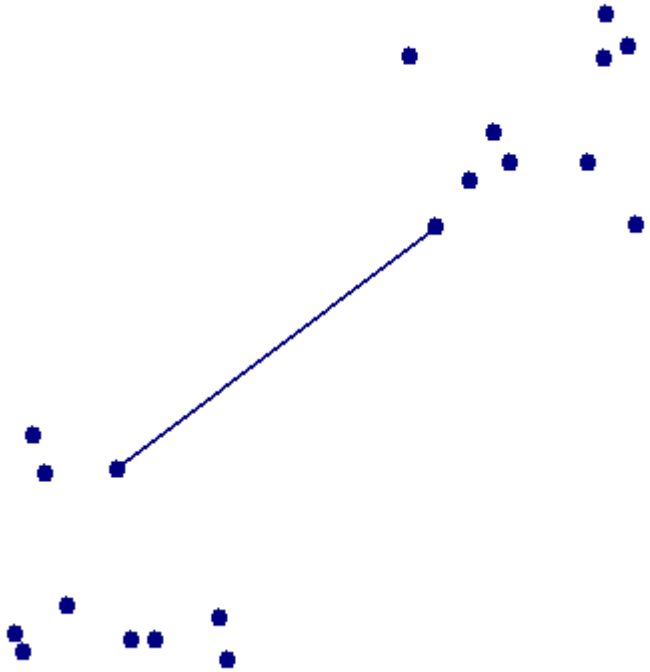


Figura 2.7 – Método da Ligação Simples para Grupos Formados Através de Dissimilaridade

Como esse método une dois grupos através de um valor que está indicando a menor distância entre eles, os grupos formados são menos homogêneos se comparados aos outros métodos hierárquicos, podendo apresentar elementos bem distintos em um mesmo grupo. Isso ocorre porque no momento da junção desses grupos, possivelmente a maioria das distâncias de seus elementos serão menos semelhantes do que a escolhida. A Ligação Simples tende a formar grupos com vários elementos, enquanto mantém isolados outros elementos que ainda não foram anexados. Este método foi popularizado através dos trabalhos de Sokal & Sneath em 1963, e Jardine & Sibson em 1971 (*apud* KRZANOWSKI & MARRIOTT, 1995).

2.2.8.1.2 Método das Médias das Distâncias (*Average Linkage*)

O Método das Médias das Distâncias, como já visto na subseção 2.2.6, utiliza a média das distâncias entre todos os pares (cada elemento de um par pertence a grupos

distintos) de elementos formados a partir dos dois grupos, para gerar a nova distância a ser utilizada na Matriz de Distâncias. Esse método é mais similar ao Método da Ligação Completa apresentando, por conta da utilização das médias de todas as distâncias envolvidas, as mesmas junções, ocorridas porém em passos distintos das que ocorrem na Ligação Completa. Esse método forma grupos mais homogêneos do que os formados pela Ligação Simples, porém menos homogêneos que os formados através da Ligação Completa. A Figura 2.8 ilustra o Método das Médias das Distâncias aplicado a dois grupos.

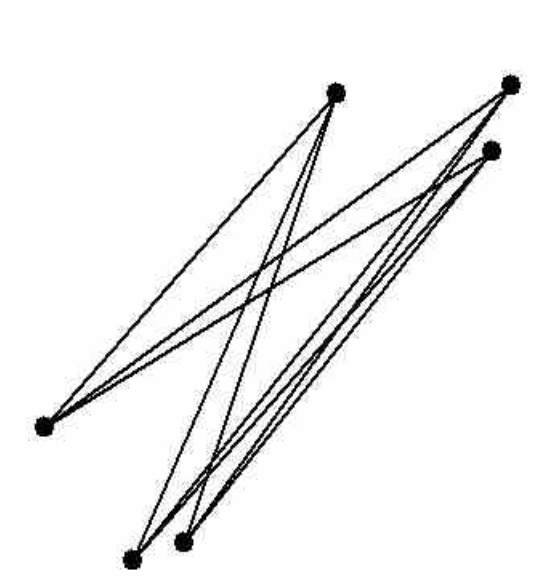


Figura 2.8 – Método das Médias das Distâncias aplicada em Grupos com três elementos cada

Esse método é afetado pela mudança de coeficiente utilizado na construção da Matriz de Distâncias, independente se esse novo coeficiente manteve a ordem dessas distâncias (JOHNSON & WICHERN, 1988). Existe um outro método muito semelhante ao das Médias das Distâncias, que utiliza a Mediana no lugar da média[26].

[26] Esse método é conhecido por WPGMC, *weighted center of mass distance*.

2.2.8.1.3 Método da Ligação Completa (*Complete Linkage*)

Conhecido como Método do Vizinho mais Distante, o Método da Ligação Completa forma grupos mais homogêneos que o método da Ligação Simples, sendo oposto a ele. A aplicação desse método busca encontrar a distância entre os elementos de um grupo e de outro, que irá representar a **menor semelhança** entre esses elementos. Se o coeficiente utilizado na construção da Matriz de Distâncias for de similaridade, a distância escolhida será o menor valor dentre as existentes nos dois grupos distintos, caso o coeficiente for de dissimilaridade, a distância escolhida será a maior existente. Voltando ao exemplo utilizado na seção 2.2.6.1, o Método da Ligação Completa forneceria, ao calcular a distância entre o grupo formado por Espírito Santo e Minas Gerais, e Rio de Janeiro, a maior distância entre esses elementos, ou seja, sabendo que *DE(ES,RJ)=1,779 e DE(MG, RJ)=1,310,* a distância calculada através do Método da Ligação Completa, D_{MLC}, seria $D_{MLC} = max \{ DE(ES,RJ); DE(MG, RJ) \} = 1,779$.

A Figura 2.9 ilustra o Método da Ligação Completa aplicado a dois grupos formados através de um coeficiente de dissimilaridade. A reta ligando os dois pontos mais distantes indica a distância que foi utilizada para designar a menor semelhança entre os grupos.

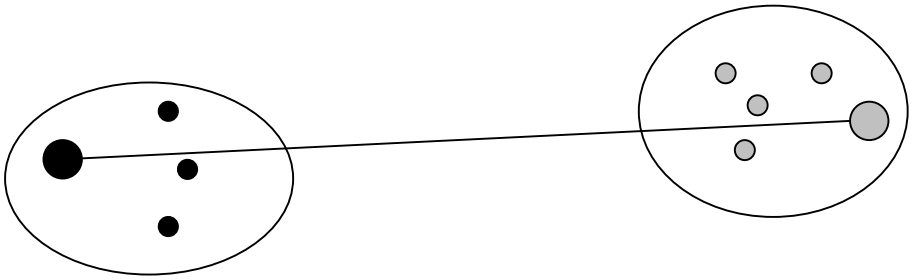


Figura 2.9 – Método da Ligação Completa para Grupos Formados Através de Dissimilaridade

A cada adição de um elemento a um grupo através da Ligação Completa, esse grupo torna-se mais distinto em relação aos outros. É característico desse método iniciar com pequenos agrupamentos os quais, durante as etapas do processo, formarão grupos maiores.

Segundo Krzanowski & Marriott (1995), a Ligação Completa é indicada para a formação de grupos com tamanhos semelhantes.

O Método da Ligação Completa assegura que as distâncias, calculadas pelo coeficiente adotado, entre todos os componentes do grupo estarão inseridos na distância utilizada em sua formação (JOHNSON & WICHERN, 1988).

Os métodos da Ligação Completa e Simples produzem grupos que não se modificam, mesmo quando um outro coeficiente substitua o que foi utilizado anteriormente, mantendo a prévia ordenação das distâncias.

2.2.8.1.4 Método da Centróide

O Método da Centróide caracteriza-se por ser bem direto na geração da distância entre um elemento e um grupo, ou entre dois grupos. Ele gera, para o novo grupo formado, uma distância representada por um único ponto, representada pela média de todas as coordenadas de seu centro.

Para exemplificarmos o seu funcionamento, novamente iremos utilizar o exemplo da seção 2.2.6.1. Espírito Santos e Minas Gerais formaram o primeiro grupo através da distância **1,091.** Através do Método da Centróide, esse novo grupo apresentaria os seguintes valores para as suas variáveis População e IDH, calculados a partir dos dados existentes na Tabela 2.6:

$$\text{Pop. Total}_{ES,MG} = (-1{,}063 + (-0{,}015))/2 = -0{,}539$$

$$\text{IDH}_{ES,MG} = (-0{,}993 + (-0{,}690))/2 = -0{,}842$$

Esses serão os valores que serão utilizados no cálculo da distância desse grupo para os outros elementos ou grupos. Devemos observar que esse cálculo utiliza o coeficiente

adotado na construção da Matriz de Distâncias, diferindo assim dos outros métodos hierárquicos apresentados, Segue abaixo o cálculo efetuado para a distância entre o grupo (ES, MG) e o Rio de Janeiro:

$$DE((ES, MG), RJ) = \sqrt{(-0{,}539 + 0{,}263)^2 + (-0{,}842 - 0{,}596)^2} = 1{,}464$$

A Figura 2.10 ilustra o Método da Centróide aplicado a dois grupos, as cruzes indicam o centro de cada um dos grupos, calculados a partir da média dos valores associados às variáveis dos elementos que os compõem.

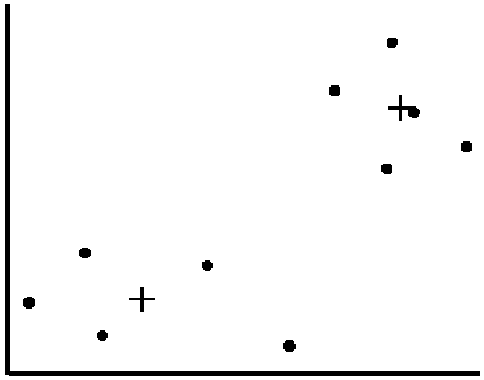


Figura 2.10 – Método da Centróide aplicado em Grupos

Os métodos da Centróide e das Médias das Distâncias são muito parecidos, e por conta disso, na maioria das vezes apresentam agrupamentos muito semelhantes. Segundo Bussab *et al* (1990), a maior dificuldade associada a esse método é representada pela necessidade de recuperação dos dados originais, a cada grupo, para recalcular os valores das distâncias para refazer a matriz, o que poderá dificultar a programação desse tipo de método.

2.2.8.1.5 Método de Ward

O Método de Ward caracteriza-se pela formação de grupos com alta homogeneidade interna. Este método utiliza como medida de homogeneidade a soma total dos quadrados de

uma análise de variância calculada para cada uma das variáveis. O método é iniciado construindo (*n-1)* grupos a partir dos *n* elementos existentes (um dos grupos terá dois elementos). A partir daí são calculadas as seguintes soma de quadrados em relação aos grupos formados (estaremos considerando abaixo somente a primeira variável, X_1, do vetor de observações *X*):

$SQT(1) = SQE(1) + SQD(1)$, onde SQT(1) é a soma de quadrados total da variável X_1, composta pela soma de SQE(1), a soma de quadrados da variável X_1 calculada entre grupos, com SQD(1) que é a soma de quadrados da variável X_1 intra grupos.

SQD(1) mede o grau de homogeneidade interna dos grupos em relação a X_1, enquanto que SQE(1) mede o grau de heterogeneidade entre os grupos. Um bom grupo para X_1 seria aquele que minimizasse SQD(1) e maximizasse SQE(1). Aplicando em todas as variáveis do vetor *X* o método de Ward, poder-se-ia utilizar SQDP como uma medida global para a avaliação da homogeneidade da Análise de Agrupamento. Associada a essa medida, utilizaríamos um critério de seleção que estaria indicando o melhor agrupamento aquele que apresentasse o menor valor para SQDP.

$SQDP = \sum_{i=1}^{p} SQD(i)$, onde p é o número de variáveis distintas dos elementos.

A implementação computacional do Método de Ward é de complexidade alta, para o qual Artes & Barroso (2003, p. 24) apresentam um bom exemplo de sua utilização.

2.2.8.2 Coeficiente de Correlação Cofenética

O Coeficiente de Correlação Cofenética é um coeficiente de correlação calculado através dos valores de duas matrizes: a matriz de distâncias original (aquela primeira que é

formada) em sua forma reduzida; e a Matriz Cofenética. Esse coeficiente permite medir o grau de associação linear entre essas duas matrizes, ele equivale ao coeficiente de Pearson e foi proposto como uma medida de concordância entre os agrupamentos obtidos e a matriz de distâncias original (BUSSAB *et al*, 1990). Antes de mostrarmos a fórmula utilizada para calcular o Coeficiente Cofenético, será apresentada a Matriz Cofenética e sua construção.

A Matriz Cofenética é construída ao substituirmos os valores da Matriz de Distâncias (reduzida) pelos valores correspondentes à distância que ocorreu a junção real entre dois elementos, utilizando para isso os mesmos valores que permitiram a construção do Dendrograma.

Por exemplo, a Matriz de Distâncias relativa ao exemplo da subseção 2.2.6, representada no Quadro I, permitiu a construção do Dendrograma naquela subseção, através das seguintes informações localizadas na Tabela 2.13 a seguir.

Elementos	**Distância**
(Espírito Santo, Minas Gerais)	1,091
(Espírito Santo, Minas Gerais, Rio de Janeiro)	1,545
(Espírito Santo, Minas Gerais, Rio de Janeiro, São Paulo)	2,365

Tabela 2.13: Dados utilizados para a construção do Dendrograma da Figura 2.5

Esses dados nos permitem construir a Matriz Cofenética (a partir da Matriz de Distâncias reduzida do Quadro I) representada no Quadro VII.

	Espírito Santo	**Minas Gerais**	**Rio de Janeiro**
Minas Gerais	**1,091** (1,091)		
Rio de Janeiro	1,545 (1,779)	1,545 (1,310)	
São Paulo	2,365 (3,180)	2,365 (2,236)	2,365 (1,678)

Quadro VII – Matriz Cofenética calculada para o exemplo da seção 2.2.6

Apesar da distância entre Espírito Santo e São Paulo, existente na Matriz de Distâncias do Quadro I, ser expressa pelo valor **3,180** (entre parênteses), na Matriz Cofenética esse valor foi alterado para **2,365**, correspondente à distância onde os dois estados se juntaram

em um mesmo grupo. No exemplo utilizado, somente a distância entre o primeiro grupo que foi formado apresentou o mesmo valor que a matriz de distâncias original.

O Coeficiente de Correlação Cofenética entre as matrizes X (Matriz de Distâncias reduzida) e Y (Matriz Cofenética relativa a X), COF, é expresso pela fórmula:

$$COF(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$, *i = 1,...,n* onde *n* é o número de

elementos preenchidos na matriz X. Cada x_i corresponde a um dos valores da matriz X e cada y_i corresponde a um dos valores da matriz Y localizado na mesma posição que x_i se localiza na matriz X. Os valores $\bar{x}$ e $\bar{y}$ correspondem as médias dos elementos existentes nas matrizes X e Y. Quanto mais próximo da unidade estiver o valor de COF, mais próximas estarão essas duas matrizes e melhor será a representação fornecida pelo Dendrograma, e por conseqüência, estaremos diante de uma Análise de Agrupamento que apresentará bons resultados (mesmos que os grupos gerados não atendam aos critérios adotados de homogeneidade). Seguimos no presente trabalho o valor utilizado por Bussab *et al* (1990), que sugerem aceitar Análises de Agrupamento que produzam um Coeficiente Cofenético a partir de **0,80**[27].

O Coeficiente Cofenético calculado para o exemplo da seção 2.2.6 é expresso pelo valor **0,743**, perto mais nem tanto de 0,80, porém aceitável em virtude dos poucos estados que participaram do exemplo.

---

[27] Esse valor é considerado alto para as ciências sociais (que admite um valor na casa de 0,70 e até menos), sendo que a pesquisa feita sobre trabalhos que tenham utilizado o coeficiente cofenético, encontrou coeficientes sendo aceitos a partir de 0,60;

Devemos ressaltar que não existe somente o Coeficiente de Correlação Cofenética para avaliar a qualidade da Análise de Agrupamento efetuada. Cormack (1971, *apud* BUSSAB *et al,* 1990) enumera as medidas de distorção de Sokal & Rohlf, Guttman, Gower, Jardine, Hartigan, Anderson, Shepard e Sammon. Artes e Barroso (2003) por sua vez, exibem a utilização de métodos de avaliação que utilizam Gráfico de Silhueta, Gráfico de Perfil e Gráfico de Radar.

2.2.8.3 Técnicas de Partição para Análise de Agrupamento

As técnicas de Partição, como o próprio nome já informa, caracterizam-se por produzirem os agrupamentos através de partições do conjunto de elementos existente. A Figura 2.6, com a seta no sentido oposto (para baixo) ilustra a atuação das técnicas de partição. Essas partições seguem as seguintes premissas: coesão interna dentro de cada grupo formado; isolamento entre os grupos e necessidade da definição do número de grupos finais. As técnicas diferem-se uma das outras em relação a um ou mais procedimentos adotados, e as principais são representadas pelos métodos k-Means e k-Medóides[28]. Estaremos mostrando no presente trabalho apenas o Método k-Means.

2.2.8.3.1 Método k-Means

O *k-Means* é uma técnica de partição que aloca cada um dos elementos existentes em um dos *k* grupos pré-definidos, ou seja, o número de grupos a ser formado deve ser definido *a priori*. Este método objetiva minimizar a soma dos quadrados residuais dentro de cada grupo (sendo bastante semelhante à Análise de Variância) com o intuito de aumentar a

homogeneidade do mesmo, ao mesmo tempo em que busca maximizar essa soma entre grupos, aumentando a diferença entre eles. Este método foi introduzido por J. B. MacQueen em 1967 (*apud* JOHNSON & WICHERN, 1988).

O k-Means, inicialmente, distribui um elemento para cada um dos *k* grupos. Essa distribuição pode ser feita aleatoriamente (o menos indicado), ou através dos elementos que apresentem os valores mais distantes de uma variável escolhida (a forma mais utilizada)[29]. Cada um desses *k* elementos será o elemento central do grupo a que pertence, representando as sementes dos grupos no momento inicial. Quando, no decorrer do método, os grupos recebem elementos, o elemento central passa a ser a média destes. Segue-se então designando cada novo elemento para um grupo, já existente, que apresente o elemento central mais próximo desse que está sendo designado. Ao fim da distribuição de todos os elementos nos *k* grupos, é calculada a soma dos quadrados residuais de cada grupo:

$SQ\text{Res}(g) = \sum_{i=1}^{n_g} (x_{ig} - \bar{X}_g)^2$, g = 1,...,k; e n_g é o número de elementos do g-ésimo grupo.

Após calcular todos os *k* SQRes(*g*), calcula-se o somatório desses, $SQ\text{Res} = \sum_{i=1}^{k} SQ\text{Res}(i)$, quanto menor esse valor, mais homogêneos serão os grupos formados.

São iniciadas, após essa primeira etapa, as iterações onde ocorrerão as movimentações dos elementos de um grupo para o outro. Após cada movimento, são recalculadas a média do grupo e a soma de quadrados correspondentes SQRes(*k*) e SQRes. Se

---

[28] Artes & Barroso (2003, p.32) apresentam um exemplo que utiliza o Método dos k-Medóides.

SQRes diminui, a movimentação que ocorreu de um objeto saindo de um grupo e indo para outro é mantida, caso contrário, o objeto volta para o grupo original ou move-se para um outro grupo, iniciando o ciclo novamente. Quando SQRes não mais diminui ou o máximo de iterações pré-definidas foi ultrapassado, o processo termina e os grupos formados são apresentados. A Figura 2.11 exibe a aplicação do método k-Means dentro da área de computação gráfica, com o objetivo de agrupar os *pixels* de cores semelhantes ao padrão RGB.

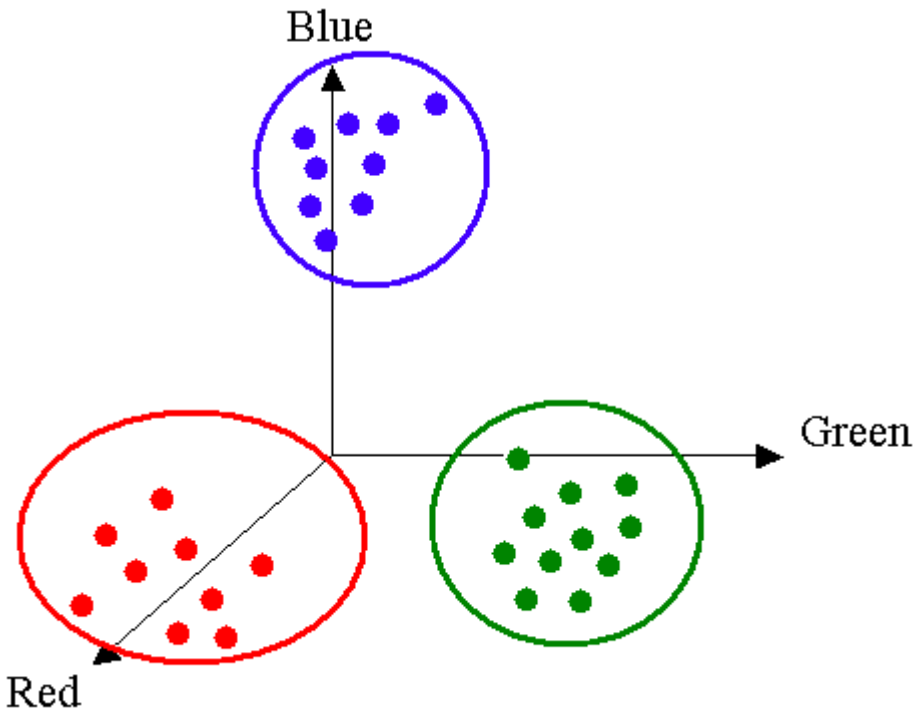


Figura 2.11 – k-Means para agrupar *pixels* de cores semelhantes

Na presente dissertação a utilização do k-Means não será necessária, porém é um importante algoritmo que provavelmente será implementado na plataforma Pii, pois é admissível que o professor de um curso qualquer, venha a necessitar de um número exato de grupos de alunos homogêneos para a implementação de uma determinada tarefa educacional.

---

[29] Uma outra forma utiliza os valores centrais de uma variável escolhida, semelhante à Mediana.

Bussab *et al* (1990, p. 58) apresenta um exemplo, com objetivos didáticos, da utilização desse método em uma massa de dados.

2.2.8.4 Conclusões

Informações atípicas não são formalmente tratadas nos métodos hierárquicos, o que significa que eles são sensíveis a esses valores ou pontos de ruído. Os métodos hierárquicos não mudam elementos de um grupo para outro, e quando um elemento se une a um outro dentro de um grupo, permanece junto a este durante o restante do processo hierárquico.

Testar diferentes métodos de agrupamento e, num dado método, utilizar diferentes distâncias (ao menos duas), produzirá uma escolha eficaz para a aplicação final da Análise de Agrupamento. Um meio de testar a estabilidade de uma solução obtida, é traduzido por aplicar o algoritmo gerando os grupos, após isso introduzir dados novos (ruído) aos dados originais, e refazer a análise para observar os grupos que foram gerados com esses novos dados.

Para uma grande massa de dados, Johnson & Wichern (1988) aconselham a utilização do k-Means ao invés de métodos hierárquicos, por ser o primeiro um método mais adequado a essas situações. Entretanto, as pesquisas que efetuamos na literatura escolhida sobre Análise de Agrupamento e as executadas através de *sites* de busca, não encontraram indícios referentes à magnitude do tamanho dessa massa de dados a que se referem esses dois autores.

2.3 Análise sobre LOG

2.3.1 Introdução

Em um ambiente EAD, como poderíamos solucionar a necessidade de avaliar o comportamento do aluno remoto, a partir da observação de suas interações com o material pedagógico disponível e com a plataforma educacional utilizada? Em contrapartida, de acordo com Souto (2003) "em um processo de aprendizagem presencial, o professor presta assistência pedagógica aos seus alunos a partir da observação e da avaliação que ele faz a respeito das interações entre: o professor e o aluno, aluno e o material pedagógico e aluno-aluno."

Na área de comércio eletrônico (*e-commerce)* o conteúdo dos Logs (arquivos contendo registros de eventos que ocorreram) referentes a servidores *Web* são constantemente analisados, através de técnicas de mineração de dados – *datamining* (DM), objetivando extração de padrões de comportamentos de acesso relacionados aos usuários (clientes) que visitam sites *e-commerce.* A característica desse tipo análise pertence a uma área em desenvolvimento, conhecida por *Webmining* (KOSALA & BLOCKELL, 2000) e (COOLEY, MOBASHER & SRIVASTAVA, 2000). De acordo com Kohavi (2001) essas análises tendem a beneficiar as organizações atuantes na área de *e-commerce*, pois permitem recomendações inteligentes ao identificar que determinado usuário apresenta um comportamento semelhante a outros usuários analisados. Kohavi destaca ainda que o conteúdo desses arquivos de Log analisados, serve como um sistema de alerta inicial para padrões emergentes e um laboratório de experimentação. A Figura 2.12 exibe um aplicativo que filtra o conteúdo do Log de um servidor, com o intuito de detectar a presença de conteúdo gerados por vírus ou similares.

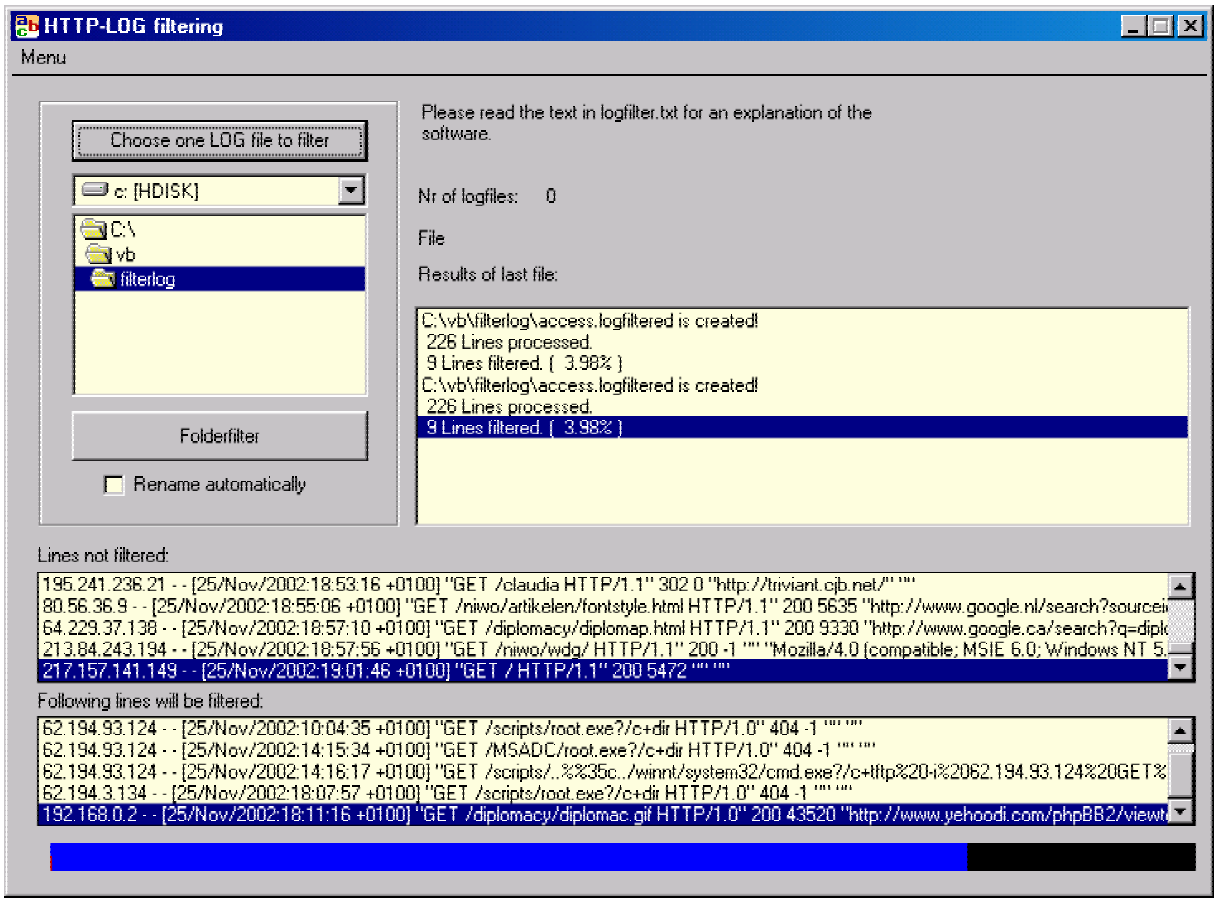


Figura 2.12 – Programa LogFilter, destinado a remover informações do Log geradas por vírus[30].

Zaiane (2000), destaca que o crescente esforço significativo das pesquisas na busca pela modelagem de padrões de comportamento na área de comércio eletrônico, tem como principal objetivo o aumento das vendas e do lucro existente na área de *e-commerce*. Alerta que nessa área muitos estudos se fazem necessários, principalmente aqueles relacionados com a descoberta de padrões de comportamento do aluno nesses ambientes.

Respondendo a pergunta contida no primeiro parágrafo da introdução da presente seção, para avaliar o comportamento de um aluno remoto em um ambiente EAD seria necessário coletar e analisar os dados contidos na principal fonte de informação que é o

arquivo de Log de navegação, localizado fisicamente no servidor *web* da plataforma EAD. Corroborando Souto (2003), atualmente existe uma forte tendência das pesquisas na área de EAD direcionada no sentido da descoberta de padrões de acesso dos alunos ao material pedagógico acessado, buscando assim mapear seu comportamento. Objetiva-se com isso a compreensão do comportamento do aluno remoto, o que permitiria um fornecimento adequado de assistências pedagógicas, tornando o processo de aprendizagem mais próximo do que ocorre numa sala de aula.

Neste capítulo estão sumarizadas as principais características dos arquivos Logs, consideradas pertinentes aos objetivos da dissertação. As metodologias de recuperação, tratamento e análise de arquivos gerados por servidores *Web*, denominados Logs, são também comentadas. Foram analisados diversos artigos referentes ao assunto, sendo selecionados para a presente revisão da literatura alguns que apresentaram um conteúdo na direção desejada.

2.3.2 Características de um arquivo de Log

Segundo Kohavi (2001), apesar do objetivo inicial do arquivo Log de um servidor ser depurar os eventos que ocorrem no mesmo, o conteúdo da maioria deles apresenta informações iniciais insuficientes, até mesmo não registrando determinados eventos significativos, o que provoca a utilização de várias técnicas (chamadas de heurísticas por ele) para reconstruir alguns eventos. Alerta ainda que uma apropriada arquitetura a ser adotada pelos *sites* diminuiria a deficiência desses arquivos. Ele categoriza os atuais ambientes de sites comerciais em três tipos: *The Good*, *The Bad* e *The Ugly*, sendo apenas os dois primeiros possíveis de execução de processos de DM.

---

[30] http://hace.dyndns.org/index.php

Kohavi enumera o que deveria existir, a partir de uma visão macro, nos arquivos de Log dos sites de *e-commerce* para serem aplicadas as técnicas de DM: 1 - O conteúdo desses arquivos deve apresentar uma grande quantidade de informações (garantindo a significância estatística e reduzindo a verossimilhança); 2 - Cada registro contido nesses arquivos deve ter vários atributos (campos) - se são poucos atributos então outras técnicas são preferíveis à utilização de DM; 3 - Os dados contidos nos Logs devem estar limpos (*clean data*) com um mínimo de ruído.

Existem diversos formatos de arquivos de Log utilizados por milhares de servidores existentes na *internet*, dentre esses existem seis formatos considerados principais (MICROSOFT, 2003) sendo suportados pelo aplicativo Internet Information Services (IIS) 6.0:

- W3C Extended log file format.
- National Center for Supercomputing Applications (NCSA) Common log file format.
- IIS log file format.
- ODBC logging.
- Centralized binary logging.
- HTTP.sys error logging.

É grande a incidência de servidores que apresentam o seu próprio formato particular de arquivo Log, baseado ou não nos seis tipos acima listados. A maioria dos servidores permite que sejam adicionados novos campos a seus Logs, como *cookies* e dados de performance.

A capacidade dos arquivos de Log deve ser observada (KOHAVI, 2001) pois é importante que suportem uma grande quantidade de registros. Os arquivos que apresentam essa característica são conhecidos por arquivos *flat*, e normalmente apresentam-se no formato associado a um banco de dados. Dos seis tipos de arquivo de Log listados anteriormente são do tipo texto (*text file*) os arquivos que seguem os formatos W3C, NCSA e IIS. O formato NCSA é um dos mais comuns, pertencendo ao formato padrão CLF (*common log format*). Um arquivo no formato NCSA registra os seguintes dados:

- Remote host name (o endereço IP do cliente)
- Remote log name (sempre um hífen).
- Username (o nome do usuário que está atuando como cliente do servidor).
- Data, hora e a diferença entre o horário local e o de *Greenwich* (GMT).
- Request and protocol version (qual comando foi emitido pelo cliente, sob que protocolo e versão).
- Service status code (um número que indica o *status* da requisição).
- Bytes sent (quantidade de bytes enviados).

A Figura 2.13 exibe um fragmento de um arquivo de Log de formato NCSA.

Figura 2.13 – Log no formato NCSA

O fragmento do arquivo Log acima informa que o cliente “John Doe” identificado pelo IP “172.21.13.45” requisitou, em “08/04/2001” às “17h39min04s – 0800” - 8h a menos

que o horário registrado em Greenwich, um comando “GET” para o arquivo “Ism.dll” sob protocolo “http" cuja versão é a “1.0”, recebendo “3401” bytes por conta da requisição. Existe ainda Logs que seguem o formato ECLF (*extended common log format*) que acrescenta dois campos ao formato CLF, ao final da linha do registro: “*referer*” (a URL onde o cliente se encontrava antes de fazer a requisição) e “*user_agent*” (o aplicativo que o cliente está utilizando para fazer a requisição, normalmente um *browser)*.

Freqüente são os arquivos de Log que apresentam problemas como a não identificação de sessões ou usuários, separação dos dados transacionais, existência de ruído (atributos com valores incorretos) ou dados corrompidos, informações redundantes, etc. (COOLEY, MOBASHER & SRIVASTA, 2000) e (KOHAVI, 2001). Para esses Logs se faz necessário um pré-processamento de dados (conhecido também por *data cleaning)*, que objetiva aprimorar a qualidade dos dados analisados, porém tal procedimento não costuma ser trivial (BATISTA, 2003). Os ruídos e os dados corrompidos que porventura venham a existir em um arquivo Log contribuem para que as predições tornem-se mais difíceis de se realizar pois acabam por esconder os padrões procurados. Estão associados à produção de ruídos e dados corrompidos a entrada manual de dados e a integração com sistemas legados.

A coleta direta e eletrônica de dados, diretamente na fonte, aliada a um processo de validação *on-line* (através de um tratamento inicial das informações equivocadas) provê uma qualidade superior aos dados de um *log*, além de torná-los altamente confiáveis.

Quando os registros de um *Log* apresentam dúzias ou centenas de atributos, técnicas automatizadas podem ser utilizadas para peneirar os dados e identificar os fatores importantes a serem utilizados em futuros processos de extração de padrões. Análise Fatorial é uma dessas técnicas mais comuns.

Para um arquivo de *log*, armazenado em um servidor e associado a um ambiente de EAD, que contenha atributos (constantemente chamados de variáveis) representando os

acessos feitos pelos alunos remotos às páginas, Souto (2003) classifica como atributos originais e a partir dos quais o conhecimento é construído, aqueles representados pelo conjunto *endereço de IP*, *data e hora de acesso*, e *recurso solicitado* (*URL*).

Os dados registrados no *Log* representam a navegação real do aluno pelas páginas que compõem o ambiente educacional remoto, mas não registram o comportamento e o relacionamento existentes entre esses dados, porém servem de matéria prima para a construção dessas informações. De acordo com Kohavi (2001), *logs* de servidores *web* servem como alerta inicial para padrões emergentes além de serem um proveitoso laboratório de experimentação.

2.3.3 Problemas mais comuns encontrados em arquivos de Log

Kosala & Blockell (2000) e Kohavi (2001), discorrem sobre os principais e mais comuns problemas encontrados, enumerados a seguir, em arquivos Log. Esses problemas dificultam e podem até a inviabilizar a utilização do conteúdo desses arquivos:

1) A identificação de sessões ou usuários, não são retidas no *Log* por conta do protocolo HTTP não disponibilizar tal informação. Condensar as requisições que formam uma sessão é atualmente um tópico ativo e importante nas pesquisas atuais. As técnicas de hoje utilizam muito as informações de *cookies*, do tempo gasto em cada evento relacionando-os sempre a IPs e suas sessões. Problemas podem ocorrer também em decorrência da existência de *caches* de *proxies*, IP dinâmico, *browsers* que impedem a utilização de *cookies*, etc.

2) A dissociação dos dados transacionais do conteúdo dos arquivos de *log*, gerando uma dificuldade na junção desses dados através da identificação dos

relacionamentos existentes, que caso não seja possível, tornaria deficiente o conteúdo existente nos *log*.

3) Deficiência na apuração de certos eventos não registrados. Eventos em sites comerciais tais como "adicionar ao carrinho", "deletar item", ou "mudar quantidade", não estão disponíveis. Uma das mais importantes métricas para o *e-commerce* é conseguir avaliar o nº de compras abandonadas e tentar associar motivos para a ocorrência desses abandonos. É importante que nenhum evento seja dispensado de registro no arquivo *log*.

4) Informações preenchidas em formulários normalmente não são retidas no *log*. É importante saber que informação está sendo inserida, pois poderá ser útil para o aperfeiçoamento do formulário e do próprio *site* por extensão.

5) O *Log* retém a *URL* porém a informação semântica do *site* não é guardada. Informações como qual imagem é exibida na página principal, que outras páginas fazem parte do processo de *checkout* ou registro, se a página principal possui múltiplas versões em diferentes línguas, etc. Isso tudo é importante para que seja possível a utilização de um mapeamento reverso.

6) Páginas que possuem conteúdo dinâmico podem dificultar o registro de informações em arquivos de *log*, caso esses não estejam devidamente preparados. Esses sites que possuem páginas desse tipo são compostos por um pequeno conjunto pequeno de padrões (conhecidos por *templates*).

7) A infra-estrutura dos arquivos *Log* tem ser condizentes com a complexidade e tamanho dos mesmos. A maioria dos servidores *web* gera arquivos no formato ASCII, os quais são ineficientes para armazenar grandes quantidades de registros complexos.

A utilização de um arquivo *Log* para análise pressupõe um esforço que buscará diminuir consideravelmente os problemas acima relatados. Quanto menos problemas um arquivo *Log* apresentar maiores serão as chances do mesmo subsidiar a descoberta de informações úteis para os pesquisadores.

Outros eventos estão associados à dificuldade da utilização das informações contidas em um arquivo de *log*, sendo representados pelo conjunto formado por Crawlers, Bots, Spiders e Robots, que são pequenos programas que são executados com a intenção de, automaticamente ou semi-automaticamente, coletarem dados ao visitarem páginas *web.* Esses programas, principalmente *bots* e *crawlers,* interferem significantemente nos padrões de *clickstream,* poluindo um arquivo *Log* com informações que não foram geradas por um verdadeiro usuário que esteja navegando através da *internet.* Um exemplo desse tipo de situação pode ser conferida no *site* da Keynote[31], que monitora o acesso de determinados *websites* fornecendo medições relativas a esse monitoramento. O *bot* dessa empresa pode gerar várias requisições por minuto, 24 horas por dia, 7 dias por semana, distorcendo com isso as estatísticas sobre o número de sessões, visitas nas páginas, e última página visitada em uma sessão. Existem outros mecanismos, como o *mySimon* que busca automaticamente na web preços de produtos. Com isso são criados padrões e clickstreams adicionais, e a tarefa de identificar e filtrar o que foi gerado por bots torna-se complexa.

A impossibilidade de transformar os dados existentes, a complexidade da captura e operacionalização das descobertas feitas e, finalmente, a escalabilidade dos algoritmos existentes de datamining, são, em menor grau e ocorrência, fatores que inviabilizariam a extração de informações relevantes dos dados existentes nos registros de um arquivo *log.*

---

[31] *The Keynote Method*, http://www.keynote.com/keynote_method/keynote_method_main_tpl.html.

Todos esses problemas exemplificados ratificam o exposto por Kosala & Blockell (2000), ao afirmarem que a falta de uma estrutura que permeie as fontes de informação que utilizam o protocolo WWW como via de comunicação com o usuário, torna árdua a descoberta de informações na *web*.

2.3.4 Alternativas à utilização de um arquivo de Log

Quando não é possível a utilização de um arquivo *log*, pela falta do mesmo ou se existente apresentar várias deficiências, para a implementação de análises se faz necessário buscar outra estratégia para que sejam retidas as informações relativas ao passo-a-passo dos usuários de um determinado site. Todas as informações relativas à interação do usuário com uma determinada página, conhecido por informações de Clickstream, são passíveis de serem coletadas rapidamente, mesmo em pequenos sites. Por exemplo, um *website* comercial que tenha por volta de 1,4 milhão de páginas visitadas por mês, assumindo uma taxa de 2% de vendas efetuadas em relação ao total de visitas, onde cada visita apresenta uma média de acesso a oito páginas por sessão, produziria pelo menos cerca de 930 vendas as quais estariam ricas em informações passíveis de análise. O popular site *Yahoo!* possui cerca de um bilhão de páginas visitadas ao dia, o que implica na produção de um *Log* cujo tamanho pode chegar a 10 GB a cada hora. Com tanta informação assim talvez seja impossível conseguir analisar tudo, uma solução seria utilizar estratégias de amostragem, como amostragem por usuários (baseadas nos *cookies*), ou, menos preferível, por sessões, o que permitiria uma alavancagem para uma análise exploratória.

As duas opções mais comuns que vêm a suprir a ausência de um *Log* suficiente são: A utilização de *Packet Sniffing*; Análise do *clickstream* através da camada de aplicação.

O *packet sniffing* seria o registro dos pacotes que trafegam entre o servidor e o cliente, normalmente associado a métodos de espionagem. Isso permite um meio não-intrusivo de ampliação do *Log* do servidor através de informação adicional. Porém se a informação está encriptada, não é possível utilizar esse método. Esse método não é tão utilizado atualmente.

Utilizar uma camada de aplicação para analisar o *clickstream* pode ser muito dispendioso, pois fatalmente implica uma reengenharia de todo o sistema associado aos sites. A terceira opção seria um método semi-intrusivo, funcionando dentro da *webpage*. O aplicativo *WebTrends* é um exemplo disso. A grande vantagem dessa opção é que não são necessárias significativas modificações na programação das páginas *web* para permitir a captura dos eventos. As principais desvantagens são:

a) Programação adicional em *Javascript* para a maioria dos eventos que ocorrem no site, duplicando praticamente o código existente. Todas as informações semânticas terão que ser rescritas.

b) Os usuários, através da configuração dos *browsers*, conseguem desligar a compatibilidade para o Javascript, impedindo o funcionamento desses aplicativos.

c) A utilização de *Javascript* impacta no nível de privacidade, podendo ser impedida caso a privacidade do *browser* estiver configura como alta ou muito alta.

d) Os dados são coletados em um terceiro sistema (não nos servidores *web* ou em uma aplicação de banco de dados), o que implicaria em mais um esforço para fundir os dados que estão em diferentes fontes de informações.

e) Vários eventos relativos às camadas das aplicações são impossíveis de rastrear.

O *Log* de um servidor web pode vir a ter várias deficiências, as quais podem ser superadas se o site possuir uma arquitetura adequada a qual venha a permitir que o registro do *Log* seja feito nas camadas da aplicação. Existem sites que gerenciam suas *webpages* ou utilizam *packet sniffers*, porém eles têm que fundir vários formatos de informação e mesmo assim os dados coletados perdem importantes informações.

2.3.5 Conclusões Preliminares

Um *website* proporciona a seus proprietários e responsáveis o acesso privilegiado a um "laboratório" para condução de experimentos, de estudos do comportamento dos usuários e de aprendizagem sobre tendências detectadas. As modificações a serem testadas em páginas *web* são de rápida implementação em sua grande maioria e após esse passo, torna-se quase que imediata à avaliação do que foi modificado, permitindo a rápida medição da reação dos usuários frente às recentes mudanças.

Segundo Kosala & Blockell, a *web* pode ser encarada como um laboratório experimental, o que adicionamos: o qual poderá ser utilizado, através de várias técnicas, para a coleta de importantes informações que invariavelmente geram induções ou deduções, verdadeiras ou não, podendo até ser difundidas para outros canais exteriores ao ambiente *internet.*

Muitos autores sugerem que o valor de um *website* é avaliado não somente pela receita imediata que ele gera, mas também pelo seu valor como um laboratório a ser utilizado como centro de pesquisa e experimentação.

Dentro do contexto a que se refere essa revisão, as principais características que são necessárias a um arquivo referente ao Log de um ambiente educacional, que permitirão a aplicação de Análise de Agrupamento no mesmo, estão relatadas a seguir:

- O arquivo *Log* deverá ter a identificação das sessões e dos usuários.
- O tipo do formato dos arquivos logs, formato texto ou não, deverão estar condizentes com o volume de informações geradas pelo ambiente educacional, sendo recomendável a existência dos atributos originais, conforme classificação de Souto (2003). Um formato do tipo *flat* e que seja lido por programas de Banco de Dados permitirá uma rápida e eficiente programação das manipulações necessárias às análises.
- Os principais eventos relacionados ao ambiente EAD deverão estar registrados no arquivo de *Log* e, caso seja possível, todos os eventos existentes poderão também pertencer a esse arquivo.
- Caso seja necessário, a arquitetura do ambiente EAD deve ser estendida aos arquivos de *log,* exemplificando: um ambiente EAD que possua diversos cursos também poderá manter um arquivo *Log* específico para cada um deles, com isso o acesso a um deles somente torna-se mais rápido além de propiciar uma melhor segurança para os dados em momentos críticos.
- O arquivo de *Log* deverá apresentar flexibilidade para modificação de sua estrutura, para que assim torne-se mais eficiente e abrangente.

Consideramos portanto que um Log, ao apresentar as características listadas anteriormente, pode ser classificado como um Log mínimo e suficiente para a aplicação de Análise de Agrupamento.

3 METODOLOGIA

Neste capítulo estaremos apresentando a Metodologia que foi seguida no sentido de alcançar os objetivos do presente trabalho. Essa metodologia foi composta de dois estudos de caso os quais permitiu-nos: (i) consolidar, na seção 3.3, as propostas iniciais da dissertação, apresentadas no primeiro capítulo, (ii) embasar o desenvolvimento da interface de análise de agrupamento, apresentada no capítulo 4.

3.1 Estudo de Caso 1 – Avaliações de Artigos na Disciplina Estudo Dirigido II

3.1.1 Introdução

Em 2003, no programa de pós-graduação *stricto sensu* do NCE, aconteceu a disciplina Estudo Dirigido II, em continuação à disciplina Estudo Dirigido I, que estabeleceu as principais necessidades dos participantes de um ambiente de ensino à distância e as características mais relevantes de uma plataforma para que a mesma possa ser considerada ideal (ROQUE *et al*, 2004). Estudo Dirigido II consistiu na avaliação e discussão de 10 artigos sobre Educação a Distância, pertencentes à seleção ocorrida no Estudo I. Participaram da disciplina alunos do mestrado sob a orientação dos mesmos dois professores, Claudia Motta e Marcos Elia, do Estudo Dirigido I.

Para finalizar o Estudo Dirigido, composto por essas duas disciplinas, foi promovida uma Mesa Redonda que teve Plataformas EAD como tema principal. Essa mesa foi composta por dois autores dos artigos selecionados, dos professores que orientaram o estudo e por um representante dos alunos que participou das duas disciplinas.

Foram efetuadas Análises de Agrupamento nos dados gerados pela disciplina, objetivando:

I – A geração dos grupos naturais e fornecimento dos mesmos para os professores.

II – A utilização de diferentes coeficientes de similaridade.

III – Verificação da viabilidade de uma implementação no aplicativo que suportou as avaliações geradas, das técnicas de Análise de Agrupamento.

3.1.2 Descrevendo o Andamento da Disciplina Estudo Dirigido II

O presente estudo de caso concentra-se no Estudo Dirigido II, principalmente nas avaliações dos 10 artigos. A disciplina ocorreu da seguinte forma:

1) Em sua primeira aula os alunos foram divididos em 5 duplas, onde cada dupla seria responsável pela apresentação de dois artigos em uma data pré-determinada conforme o cronograma da disciplina. Os artigos de cada dupla estavam associados a uma ou duas das seguintes categorias: *Interface; Navegação; Avaliação; Recursos Didáticos; Comunicação/Interação; Coordenação; Apoio Administrativo*. Essas associações criadas no Estudo Dirigido I não foram levadas em conta no estudo de caso, sendo exibidas aqui apenas como ilustração.

2) A cada encontro quinzenal após a primeira aula, todos os participantes voltavam a se reunir para que a dupla da semana apresentasse os artigos que lhe competiam para logo em seguida ocorrer uma discussão acerca do que foi apresentado e do conteúdo dos textos. Esperava-se que, antes do encontro, todos os alunos tivessem emitido sua análise particular sobre o artigo através da Plataforma TeamWorks (TW), um ambiente de CSCW (MOTTA & BORGES, 2000), que além de suportar essas avaliações também armazenava os artigos dentre outras funcionalidades as quais não serão descritas no presente estudo. Participaram de algumas dessas reuniões como ouvintes outros alunos interessados no tema do dia.

Após a reunião quinzenal o ambiente TW permitia, até uma certa data limite (definida pelos professores), que as avaliações já registradas sobre os artigos da última reunião fossem alteradas ou que novas fossem incluídas, caso um aluno não tivesse emitido a sua antes do encontro quinzenal.

As avaliações eram realizadas para cada artigo através da atribuição de escores aos critérios WIMPE (NICOL, 2001) que originalmente são: *Originality, Tech. Merit, Readability, Relevance, Confidence, Overall* e *Action*. O critério *Confidence* não foi utilizado mas poderia ter sido, pois o ambiente TW é bastante parametrizável nesse sentido. A seguir, na Tabela 3.1, encontra-se um exemplo de uma avaliação feita por um dos participantes para o artigo *Uma Abordagem Multiagente*:

Uma Abordagem Multiagente	Originalidade	4
	Mérito Técnico	2
	Legibilidade	1
	Relevância	3
	Conceito Geral	2
	Ação Recomendada	4

Tabela 3.1: Exemplo de Avaliação Utilizando Critérios WIMPE

Os escores atribuídos para cada um dos cinco primeiros critérios (Originalidade, Mérito Técnico, Legibilidade, Relevância e Conceito Geral) obedecem à escala Likert[32] de 1 a 5, onde o escore 1 representa *discordo muito* e 5 representa *concordo muito*. O escore 3 (*indiferente)* significa que a atitude de avaliação do participante no determinado critério é neutra.

[32] O psicólogo social Rensis Likert (1903-1981) em 1932 introduziu, através de sua tese de doutoramento (*A Technique For The Measurement of Attitudes*) na Universidade de Columbia, EUA, uma escala, a qual é bastante utilizada nos dias de hoje em diferentes campos de pesquisa, para medir a atitude do entrevistado, essa escala ficou conhecida por escala Likert (LIKERT, 1932). Essa escala, normalmente com valores de 1 a 5 ou 1 a 7, estruturalmente é delimitada por dois extremos, do favorável ao desfavorável, com um ponto central neutro reservado para as respostas que refletem uma posição indecisa. É uma escala onde os respondentes são solicitados não só a concordarem ou discordarem das afirmações, mas também a informarem qual o seu grau de concordância/discordância (http://www.thoemmes.com/dictionaries/bdm_likert.htm acessada em 12/02/2005).

O critério Ação Recomendada apresentava uma polaridade diferente dos outros critérios que apresentavam polaridade (ou direção) positiva. O escore 1 para Ação Recomendada significa que o aluno *Concorda Muito* na recomendação do artigo, e o escore 5 significa que ele *Discorda Muito*, ou seja, ele não recomendaria o artigo para leitura. Ação Recomendada possui polaridade negativa. Essa arrumação foi efetuada e justificada na primeira aula através da necessidade do aluno estar bastante atento no momento do preenchimento da Ação Recomendada, que reflete os outros critérios porém apresenta a polaridade invertida aos mesmos. Infelizmente erros ocorreram no preenchimento e nas posteriores utilizações, todos os critérios passaram a ter a mesma polaridade.

3.1.3 Apresentando a Demanda pela Análise de Agrupamento e os Dados Analisados

Os professores orientadores do estudo apresentaram a necessidade de formar grupos de alunos em relação às avaliações feitas dos artigos, dentre os vários motivos alegados por estes encontrava-se o fato de que conhecendo quais são os alunos que apresentam avaliações parecidas, a construção de outros grupos (duplas, trios, etc.) para outras tarefas seria facilitada. Ressaltamos que as duplas iniciais foram construídas seguindo critérios subjetivos dos professores, esses critérios estavam associados principalmente à afinidade entre os integrantes da dupla e facilidade em se reunirem para discutir os artigos que foram destinados a apresentarem.

Na Tabela 3.2, encontram-se os dados que representam as avaliações feitas pelos participantes, no Estudo Dirigido 2, em relação a cada um dos artigos (ausências de avaliações na tabela, significam que determinado participante deixou de efetuar a avaliação do texto em questão). Esses dados foram utilizados para a execução das Análises de Agrupamentos contidas na presente seção.

Artigos	Critérios (WIMPE)	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde	Catete
Uma Abordagem Multiagente	Originalidade	4	2	1	2	1	2	2	3	2	2
	Mérito Técnico	2	1	1	2	1	1	1	2	1	2
	Legibilidade	1	1	1	1	2	3	1	2	3	3
	Relevância	3	2	2	3	1	2	1	3	2	2
	Conceito Geral	2	2	1	2	1	2	1	2	2	2
	Ação Recomendada	4	3	4	4	4	4	4	3	3	4
Design For Learning	Originalidade	2	2	4	3	3		3		2	3
	Mérito Técnico	2	2	5	3	3		3		3	3
	Legibilidade	4	3	4	4	4		4		2	3
	Relevância	2	2	5	3	2		3		3	2
	Conceito Geral	2	2	5	3	3		3		3	3
	Ação Recomendada	4	3	2	3	3		3		3	3
Contribuições de conceitos de comunicação	Originalidade	4	4	3	4	4		4	4	4	
	Mérito Técnico	4	4	4	3	3		4	4	4	
	Legibilidade	4	5	4	4	4		4	5	4	
	Relevância	4	4	4	3	3		4	5	4	
	Conceito Geral	4	4	4	4	3		4	4	4	
	Ação Recomendada	1	1	2	2	2		2	1	1	
Hipermídias Distribuídas e Educação	Originalidade	3	3	4	3	3	3	3	3	3	3
	Mérito Técnico	4	3	5	3	3	4	4	3	4	3
	Legibilidade	3	4	5	4	4	4	5	3	4	4
	Relevância	4	3	4	3	3	3	4	5	3	3
	Conceito Geral	4	3	5	3	3	4	4	3	3	3
	Ação Recomendada	3	3	1	3	3	3	2	3	3	3
Interface de Ambientes Educacionais	Originalidade	4	3	4	3	3	3	4	4	3	3
	Mérito Técnico	4	2	4	4	3	4	5	4	4	3
	Legibilidade	5	3	4	4	4	4	5	5	4	4
	Relevância	4	3	4	4	4	4	5	5	4	4
	Conceito Geral	4	1	4	4	4	4	5	4	4	3
	Ação Recomendada	2	3	2	2	2	2	1	2	2	3
JAVAL? Ambiente para avaliações remotas em EAD	Originalidade	5	3	3	5	5	3	5	4	3	
	Mérito Técnico	4	3	3	4	4	4	4	4	4	
	Legibilidade	4	4	4	4	3	4	4	5	3	
	Relevância	4	3	5	4	4	4	4	5	3	
	Conceito Geral	4	3	3	4	4	4	4	4	3	
	Ação Recomendada	2	3	3	2	1	2	2	2	3	
Teaching and Learning with Telematics	Originalidade	4	4	5	4	4		4	4	4	4
	Mérito Técnico	4	4	5	4	5		5	4	4	5
	Legibilidade	3	3	5	4	5		4	3	4	5
	Relevância	3	4	5	5	5		5	4	4	5
	Conceito Geral	4	4	5	4	5		5	4	4	5
	Ação Recomendada	3	1	1	2	1		1	2	1	1
Tecnologias de Informação Aplicadas à Educação: construindo uma rede de aprendizagem utilizando o ambiente AulaNet.	Originalidade	4	3	4	4	4		4	4	4	4
	Mérito Técnico	4	5	5	4	4		4	4	5	4
	Legibilidade	4	4	4	5	4		5	5	5	4
	Relevância	4	4	4	4	5		4	5	4	5
	Conceito Geral	4	4	4	4	1		4	4	5	4
	Ação Recomendada	2	2	1	1	1		1	2	1	1

Uma máquina de estados finitos para avaliação de desempenho em um grupo de discussão on-line	Originalidade	4	4	4	4	4	5	4	3	3	
	Mérito Técnico	3	3	5	3	3	4	4	4	4	
	Legibilidade	4	1	5	3	2	3	2	3	2	
	Relevância	4	3	4	3	3	4	3	4	3	
	Conceito Geral	4	3	5	3	3	4	3	4	3	
	Ação Recomendada	3	3	1	3	2	2	2	2	3	
Educação a Distância - Padrões para Projetos de Sistemas	Originalidade	4		4	4	3	4	4	4	4	
	Mérito Técnico	4		4	4	3	4	4	4	4	
	Legibilidade	4		4	5	4	3	5	5	5	
	Relevância	5		5	4	4	4	5	5	4	
	Conceito Geral	5		5	4	4	4	4	4	4	
	Ação Recomendada	1		1	2	2	2	1	1	1	

Tabela 3.2: Avaliações emitidas pelos Alunos para os 10 artigos[33].

As análises executadas basearam-se no contexto da seção 2.2.7.2, e tiveram como objetivo principal a geração de grupos contendo os alunos mais semelhantes, em relação a avaliação positiva ou negativa de um artigo. O fornecimento dos grupos encontrados aos professores, além de subsidiá-los de informações acerca da similaridade dos avaliadores, possibilitou que eles verificassem se as características atribuídas pelos professores a cada um dos avaliadores, correspondiam aos grupos.

3.1.4 Análises de Agrupamento Efetuadas

Antes de mostrarmos a metodologia utilizada e os procedimentos que foram feitos, gostaríamos de comentar a escala utilizada nos critérios de avaliação dos textos (artigos): a escala Likert, apesar de ser considerada uma escala ordinal intervalar, pode, em determinados casos, não apresentar a propriedade numérica esperada das escalas intervalares, a qual seria quantificar, precisamente e representativamente, a diferença entre as categorias. No nosso Estudo de Caso isso pode ser exemplificado pelo simples fato de não sabermos expressar o quanto um conceito 5 (Concordo Muito) no critério Legibilidade é maior que um conceito 4 (Concordo). Sabemos que o valor 5 expressa uma maior concordância em relação a 4, porém

não podemos definir o quanto ele é maior, apenas medir que essa diferença é 1 mas não sabemos o que esse 1 significa em relação a variável qualitativa.

3.1.4.1 Metodologia Adotada

Como os professores estavam interessados em verificar os alunos que emitem avaliações semelhantes, consideramos, para a aplicação da Análise de Agrupamento, apenas o critério Ação Recomendada[34], dividindo os escores da escala Likert em *Recomendado* (escores 1 e 2) cujo valor é 1, e *Não recomendado* (os outros escores ou a ausência de escore) cujo valor é 0. Ressaltamos que o escore neutro está sendo considerado como uma ação de não recomendação do artigo, esse foi o critério utilizado nas análises efetuadas, porém existem outros critérios, como a utilização do valor esperado daquele avaliador nos valores ausentes. Essa simplificação, através da dicotomização, é razoável por dois motivos: 1) Normalmente não se recomenda a leitura de um artigo (o que pode ser uma tarefa complexa) quando não se têm opiniões favoráveis sobre ele. 2) Aqueles alunos que não preencheram as avaliações de algum artigo, têm suas Ações Recomendadas consideradas como *Não recomendado*, pois não participaram à disciplina suas opiniões sobre o mesmo.

A Tabela 3.3, mostra os valores relativos ao escore dado para o critério Ação Recomendada, já com essa nova formatação binária (ou dicotômica).

---

[33] Os nomes verdadeiros dos avaliadores foram substituídos por nomes de bairros da cidade do Rio de Janeiro.
[34] Uma outra metodologia seria utilizar todos os seis critérios, ponderando Ação Recomendada com, por exemplo, peso 2.

Artigo	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde	Catete
Uma Abordagem Multiagente	0	0	0	0	0	0	0	0	0	0
Design For Learning	0	0	1	0	0	0	0	0	0	0
Contribuições de conceitos de comunicação	1	1	1	1	1	0	1	1	1	0
Hipermídias Distribuídas e Educação	0	0	1	0	0	0	1	0	0	0
Interface de Ambientes Educacionais	1	0	1	1	1	1	1	1	1	0
JAVAL? Ambiente para avaliações remotas em EAD	1	0	0	1	1	1	1	1	0	0
Teaching and Learning with Telematics	0	1	1	1	1	0	1	1	1	1
Tecnologias de Informação Aplicadas à Educação: construindo uma rede de aprendizagem utilizando o ambiente AulaNet	1	1	1	1	1	0	1	1	1	1
Uma máquina de estados finitos para avaliação de desempenho em um grupo de discussão on-line	0	0	1	0	1	1	1	1	0	0
Educação a Distância - Padrões para Projetos de Sistemas	1	0	1	1	1	1	1	1	1	0

Tabela 3.3: Ação Recomendada dicotomizada dos avaliadores

Da tabela anterior, escolheu-se um determinado par de avaliadores para a construção da tabela de contingência (ou dupla entrada), exibida no Quadro VIII, a seguir.

		Avaliador Glória		Total
		1	0	
Avaliador Ilha	1	**3** **a**	**5** **b**	8 a+b
	0	**1** **c**	**1** **d**	2 c+d
Total		4 a+c	6 b+d	10

Quadro VIII – Dupla Entrada para Ação Recomendada dos Avaliadores Ilha e Glória

O quadro anterior, fornece a incidência dos possíveis pares (1,1), (1,0), (0,1) e (0,0)[35] entre esses avaliadores, através dos quais algumas hipóteses iniciais foram formadas, como por exemplo: 1 – Os dois avaliadores apresentaram poucas concordâncias ao recomendarem positivamente os artigos, três ao todo, e quase nenhuma ao recomendarem negativamente; 2 – Um total de cinco das seis avaliações negativas dos artigos feitas por Glória, repercutiram em

[35] O par (1,1) significa que os dois avaliadores recomendam o texto, (1,0) significa que o primeiro avaliador não recomenda enquanto que o segundo recomenda, o par (0,1) é exatamente o contrário do par (1,0) e finalmente o par (0,0), onde ambos não recomendam o texto.

contrário no avaliador Ilha. 3 – Aparentemente, os dois avaliadores não são semelhantes quanto à forma de avaliar os artigos. Esses totais foram utilizados na construção dos coeficientes escolhidos na revisão da literatura[36].

Como é de interesse dos professores conhecer os participantes que avaliam de forma semelhante, positivamente ou negativamente, as análises de agrupamento efetuadas utilizaram coeficientes que focalizaram em pelo menos um dos índices ***a*** e ***d.*** Os coeficientes escolhidos e seus respectivos valores, encontram-se na Tabela 3.4.

Coeficientes	**Valor calculado**
Concordância Simples (Sokal e Michener)	0,400
Jaccard	0,333
Russel e Rao (Concordâncias Positivas)	0,300
Ochiai	**0,530**
Baron-Urbani-Buser	**0,534**
Hamann	-0,200
Yule	-0,250
Gower2	0,153
Anderberg	**0,448**
Pearson	-0,102

Tabela 3.4: Coeficientes Calculados para os Avaliadores Ilha e Glória com índices a=4, b=5, c=0, d=1

Ressaltamos que os coeficientes que atribuem peso aos índices *a, b, c* e *d* foram descartados, por não haver informação suficiente para se decidir sobre a ponderação desses índices.

Podemos observar na tabela anterior que somente dois coeficientes, Ochiai e Baron-Urbani-Buser, apresentaram valores indicando uma certa semelhança entre os avaliadores, quando estes emitem recomendações positivas ou negativas acerca dos artigos. Porém essa indicação contradiz uma das hipóteses preliminares, acerca desse par de avaliadores. Esses

[36] Vide o Quadro III – Coeficientes de Semelhança para Variáveis Binárias, existente na Revisão da Literatura.

dois coeficientes ficaram muito próximos um do outro, sendo que o coeficiente de Ochiai não envolve no seu cálculo o índice *d*, responsável pelo número de vezes que os avaliadores concordaram na não-recomendação do artigo. Por conta dessas observações, as análises de agrupamento efetuadas não utilizaram esses dois coeficientes, Ochiai e Baron-Urbani-Buser, sendo excluído também o coeficiente de Anderberg, pelos mesmos motivos.

A partir dos coeficientes Concordância Simples (Sokal e Michener), Jaccard, Russel e Rao (Concordâncias Positivas), Hamann, Yule, Gower2 e Pearson foram efetuadas as aplicações das técnicas de análise de agrupamento.

A escolha do coeficiente a ser utilizado na análise de agrupamento depende dos objetivos do agrupamento e, ao mesmo tempo, é um pouco subjetiva, sendo comum a execução da análise repetidas vezes com diferentes coeficientes, obtendo assim uma maior segurança na escolha do coeficiente final.

Como esse estudo de caso apresenta a peculiaridade de estar sendo executado, adicionalmente aos objetivos dos professores, para que as técnicas de análise de agrupamento sejam testadas, foram aplicadas sete análises de agrupamento nos dados, cada uma delas utilizando um coeficiente de similaridade distinto.

3.1.4.2 Análise de Agrupamento para o Coeficiente de Pearson

Estará sendo mostrado adiante, algumas etapas iniciais e a etapa final da análise que utilizou o coeficiente de Pearson, o qual mede a força das concordâncias em relação às discordâncias. Quanto mais próximo da unidade estiver o valor calculado para este coeficiente, maior será a similaridade entre os elementos quando da recomendação (positiva ou negativa) do artigo. Quanto mais próximo de –1, maior a discordância entre os elementos, indicando que enquanto um recomenda, o outro não recomenda, e vice-versa. Valores

próximos a 0, indicam que os participantes não apresentam alguma semelhança significativa ao recomendarem os artigos. O coeficiente de Pearson é expresso pela fórmula: $\frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$.

Em relação às outras análises de agrupamento, somente serão exibidas, na presente seção, os grupos homogêneos e partes dos dados gerados. As informações relativas às gerações dos grupos (matriz de distâncias, tabela de agrupamentos, dendrogramas e matriz cofenética e coeficiente cofenético) encontram-se no Apêndice B.

O Quadro IX mostra a Matriz de Distâncias construída através do coeficiente de Pearson, aplicado às tabelas de contingência dos avaliadores.

	Barra	**Tijuca**	**Ilha**	**Centro**	**Leblon**	**Glória**	**Urca**	**Ipanema**	**Saúde**	**Catete**
Barra	0									
Tijuca	0,218218	0								
Ilha	0	0,327327	0							
Centro	0,816497	0,534522	0,102062	0						
Leblon	0,654654	0,428571	0,218218	0,801784	0					
Glória	0,408248	-0,65465	-0,10206	0,25	0,534522	0				
Urca	0,5	0,327327	0,375	0,612372	0,763763	0,408248	0			
Ipanema	0,654654	0,428571	0,218218	0,801784	**1**	0,534522	0,763763	0		
Saúde	0,6	0,654654	0,5	0,816497	0,654654	0	0,5	0,654654	0	
Catete	0	0,763763	0,25	0,408248	0,327327	-0,40825	0,25	0,327327	0,5	0

Quadro IX – Matriz de Distâncias utilizando o Coeficiente de Pearson

Como o coeficiente de Pearson é uma distância de similaridade, então os maiores valores expressam uma maior semelhança entre os elementos. Nessa primeira etapa, os grupos foram formados buscando identificar aqueles avaliadores que apresentaram uma maior recomendação positiva, por conta disso, somente foram considerados em cada formação dos grupos, aquele valor que traduz essa maior similaridade.

Observando os valores da Matriz de Distâncias do Quadro IX, podemos identificar que os avaliadores mais semelhantes são exatamente Leblon e Ipanema, que apresentaram o valor máximo para o Coeficiente de Pearson, expresso por 1. Então o primeiro grupo

homogêneo formado é composto por (Leblon, Ipanema) e os outros avaliadores permanecem, cada um deles, sendo um único grupo. Esse primeiro grupo homogêneo será utilizado para substituir o participante Leblon ou Ipanema nas outras distâncias para os outros avaliadores, que terão que ser recalculadas. Nesse momento que antecedeu o recálculo das novas distâncias, ocorreu a escolha do algoritmo de agrupamento. Para que fossem formados grupos mais homogêneos, optamos pelo Método da Ligação Completa ou Vizinho mais Longe, onde se escolhe a distância que representa a maior diferença entre os elementos. Como o primeiro grupo alcançou o valor máximo para o coeficiente de Pearson, significando que esses dois elementos apresentaram avaliações idênticas, não é preciso recalcular os valores da matriz de distâncias, que terá sua dimensão diminuída como mostrado no Quadro X.

	Barra	**Tijuca**	**Ilha**	**Centro**	**Leblon, Ipanema**	**Glória**	**Urca**	**Saúde**
Tijuca	0,218218							
Ilha	0	0,327327						
Centro	**0,816497**	0,534522	0,102062					
Leblon, Ipanema	0,654654	0,428571	0,218218	0,801784				
Glória	0,408248	-0,65465	-0,10206	0,25	0,534522			
Urca	0,5	0,327327	0,375	0,612372	0,763763	0,408248		
Saúde	0,6	0,654654	0,5	0,816497	0,654654	0	0,5	
Catete	0	0,763763	0,25	0,408248	0,327327	-0,40825	0,25	0,5

Quadro X – Matriz de Distâncias após o primeiro grupo formado

A formação do segundo grupo segue o mesmo procedimento executado quando da formação do primeiro, escolhe-se o maior valor na matriz de distâncias representada no Quadro X. A distância entre os avaliadores Centro e Barra (0,816497), apresentou este maior valor procurado, porém os avaliadores Saúde e Centro também apresentaram o mesmo valor. Como são somente duas distâncias com o mesmo valor, é seguida a convenção para escolher o primeiro par. O segundo grupo foi composto por Centro e Barra, e a matriz de distâncias terá que ser recalculada, como pode ser verificado no Quadro XI.

	Centro, Barra	**Tijuca**	**Ilha**	**Leblon, Ipanema**	**Glória**	**Urca**	**Saúde**
Tijuca	0,218218						
Ilha	0	0,327327					
Leblon, Ipanema	0,654654	0,428571	0,218218				
Glória	0,25	-0,65465	-0,10206	0,534522			
Urca	0,5	0,327327	0,375	**0,763763**	0,408248		
Saúde	0,6	0,654654	0,5	0,654654	0	0,5	
Catete	0	0,763763	0,25	0,327327	-0,40825	0,25	0,5

Quadro XI – Matriz de Distâncias após o segundo grupo formado

O recálculo da distância para os avaliadores Barra (que será substituído pelo novo grupo formado) e Glória, utilizando o Método da Ligação Completa, foi feito da seguinte forma: a distância, fornecida pelo coeficiente de Pearson, entre Barra e Glória é 0,408248, enquanto que a distância entre Centro e Glória é 0,25. Barra está mais próximo de Glória, enquanto que Centro está mais distante de Glória. Pelo método de agrupamento escolhido, a nova distância entre o par formado pelo novo grupo, (Centro e Barra) e Glória, é aquela que expressa a maior distância entre eles, representada pelo valor 0,25. A construção da distância dos outros pares acontece de modo análogo. O Apêndice B, contém o restante dos quadros, onde é possível acompanhar as mudanças ocorridas na Matriz de Distâncias, à medida que os grupos vão se formando e os valores das distâncias se alteram.

O próximo grupo formado é identificado na matriz do Quadro XI, através da distância 0,763763, correspondente aos elementos Urca e (Leblon, Ipanema). Esse novo grupo é formado pela junção de um avaliador e um outro grupo anteriormente formado, o qual é tratado como se fosse um avaliador, dentro da matriz. O ciclo então recomeça, novamente a matriz é recalculada e a formação de um novo grupo é identificada através do maior valor para as distâncias desta nova matriz.

A Tabela 3.5 mostra as etapas ou passos da formação dos grupos, até a união de todos os avaliadores em um único grupo. Os dados contidos nessa tabela são importantes pois auxiliaram na identificação dos grupos homogêneos, que surgem a partir da definição da

variação percentual máxima, utilizada para avaliar se ocorreu a junção de elementos ou grupos heterogêneos. A definição da variação percentual surge da análise que foi feita na própria Tabela 3.5.

Passo	Grupo	Nível	Diferença percentual absoluta para o nível anterior
1	Leblon, Ipanema	1	-
2	Centro, Barra	0,816497	18,35%
3	Urca, (Leblon, Ipanema)	0,763763	6,46%
4	Tijuca, Catete	0,763763	0,00%
5	Saúde, (Centro, Barra)	0,6	21,44%
6	(Saúde, Centro, Barra), (Urca, Leblon, Ipanema)	0,5	16,67%
7	(Tijuca, Catete), Ilha	0,25	50,00%
8	Glória, (Saúde, Centro, Barra, Urca, Leblon, Ipanema)	0	100,00%
9	(Glória, Saúde, Centro, Barra, Urca, Leblon, Ipanema), (Tijuca, Catete, Ilha)	-0,65465	-

Tabela 3.5: Agrupamentos gerados pela aplicação da Ligação Completa à Matriz de Distâncias

A partir dos dados da tabela anterior, foi construído o dendrograma que representa, através de uma estrutura de árvore, as junções que produziram os grupos. Esse tipo de gráfico é importante pois é através dele que podemos avaliar se a formação de um determinado grupo, juntou elementos homogêneos ou heterogêneos. A Figura 3.1 exibe esse dendrograma construído.

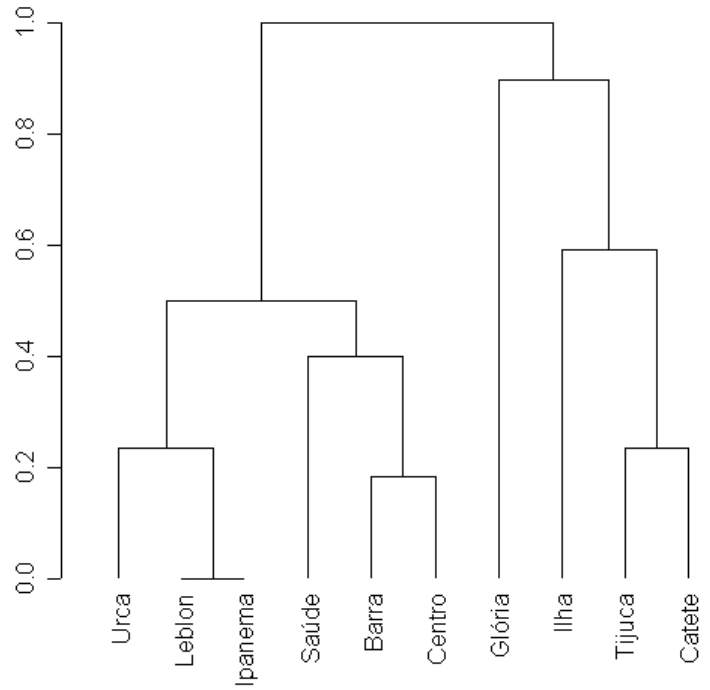


Figura 3.1 – Dendrograma referente aos Agrupamentos da Tabela 3.5[37]

Matriz Cofenética

Verificamos se o Dendrograma produzido foi uma boa simplificação da Matriz de Distâncias, e, por conseqüência, se a Análise de Agrupamento produzida atingiu os objetivos esperados, através do cálculo do Coeficiente de Correlação Cofenética, representado pela correlação entre a matriz de distâncias original, de onde os grupos surgiram, e a matriz cofenética, que surge a partir do Dendrograma.

[37] O dendrograma foi gerado pelo aplicativo estatístico R, porém esse aplicativo não ofereceu um método de agrupamento que utilize coeficientes de similaridade. A solução encontrada foi importar toda a matriz de distâncias para o aplicativo, alterar cada valor das células da matriz através da transformação Matriz = 1-abs(Matriz), aplicar o método da ligação completa na matriz e gerar o dendrograma. Por conta dessa transformação, o eixo das ordenadas inicia os agrupamentos a partir do valor zero. Entretanto, por conta dos valores negativos gerados pelo coeficiente de Pearson, o avaliador Glória se juntou ao grupo formado por Ilha, Tijuca e Catete, ao invés de se juntar ao outro grupo. Várias tentativas foram feitas para reverter o ocorrido, porém não encontramos uma solução adequada.

A matriz cofenética, como já visto na revisão da literatura do capítulo 2, informa em cada célula qual foi a distância que ocorreu à união do par de avaliadores no dendrograma. Por exemplo, vimos que os avaliadores Leblon e Centro apresentaram uma distância, gerada pelo coeficiente de Pearson, igual a 0,801784. Porém, a partir dos dados da Tabela 3.5, esses dois avaliadores se juntaram em um mesmo grupo somente no passo 6, através da distância 0,5. Portanto, na matriz cofenética, a distância entre esses dois avaliadores registrada foi 0,5 ao invés de 0,801784. Seguindo essa mesma regra, a matriz cofenética é representada no Quadro XII, a seguir.

	Barra	**Tijuca**	**Ilha**	**Centro**	**Leblon**	**Glória**	**Urca**	**Ipanema**	**Saúde**
Tijuca	-0,65465								
Ilha	-0,65465	0,25							
Centro	0,816497	-0,65465	-0,65465						
Leblon	0,5	-0,65465	-0,65465	0,5					
Glória	0	-0,65465	-0,65465	0	0				
Urca	0,5	-0,65465	-0,65465	0,5	0,763763	0			
Ipanema	0,5	-0,65465	-0,65465	0,5	1	0	0,763763		
Saúde	0,6	-0,65465	-0,65465	0,6	0,5	0	0,5	0,5	
Catete	-0,65465	0,763763	0,25	-0,65465	-0,65465	-0,65465	-0,65465	-0,65465	-0,65465

Quadro XII – Matriz Cofenética

O Coeficiente de Correlação Cofenética entre as matrizes representadas nos Quadros IX e XII foi calculado, apresentando o valor **0,67520**. Este valor encontra-se distante de 0,80 (em diante), que em Análise de Agrupamento é considerado um bom ajuste ao indicar que a matriz de distâncias original foi pouco distorcida ao final do processo (SNEATH & SOKAL, 1973 *apud* BUSSAB, 1991). Porém este valor foi aceito, para permitir a comparação com os outros coeficientes de correlação cofenética, gerados nas outras análises de agrupamento, que serão vistas mais à frente. A comparação entre diferentes coeficientes cofenéticos visa auxiliar na escolha do coeficiente final, a ser utilizado em uma única análise de agrupamento.

Foi considerado que a Análise de Agrupamento utilizando o coeficiente de Pearson e o Método da Ligação Completa obteve sucesso e, por conta disso, foram gerados os grupos a seguir.

Grupos Gerados

Analisando o dendrograma à procura de saltos que indiquem a união de grupos heterogêneos, são identificados alguns pontos que sugerem quebras de homogeneidade. Para confirmarmos esses pontos é necessário definir um patamar, representado por uma máxima variação percentual, onde estaríamos nos certificando que ocorreram quebras de homogeneidade.

Ao definirmos um patamar como sendo 15%, os grupos homogêneos estariam restritos a apenas três com mais de um elemento, apresentados na ordem do mais homogêneo para o menos[38], a saber:

(Leblon, Ipanema);

(Urca, Leblon, Ipanema);

(Tijuca, Catete).

Obs.: O restante dos avaliadores que não pertencem a um dos grupos acima representariam, cada um deles, um único grupo homogêneo: (Ilha), (Centro), (Glória), (Saúde) e (Barra)

[38] Sempre estaremos seguindo essa ordem ao apresentarmos os grupos.

Se ao invés de 15%, optássemos por 20% para a máxima variação percentual, relaxando um pouco nosso conceito de homogeneidade, os grupos formados por mais de um avaliador, em cada etapa da análise de agrupamento, seriam:

(Leblon, Ipanema);

(Urca, Leblon, Ipanema);

(Tijuca, Catete);

(Centro, Barra);

(Saúde, Centro, Barra, Urca, Leblon, Ipanema).

Obs.: O restante dos avaliadores que não pertencem a um dos grupos acima, representariam, cada um deles, um único grupo homogêneo: (Ilha) e (Glória).

3.1.4.3 Análise de Agrupamento para o Coeficiente Concordância Simples

O Coeficiente de Concordância Simples (ou Sokal e Michenner), fornece os avaliadores mais semelhantes quanto à recomendação dos artigos, independente se essa recomendação foi negativa ou positiva. Sua fórmula é expressa por $\frac{a+d}{a+b+c+d}$.

A aplicação do método da Ligação Completa na matriz de distâncias, gerou os seguintes agrupamentos, exibidos na Tabela 3.6.

Passo	Grupo	Nível	Diferença percentual absoluta para o nível anterior
1	Leblon, Ipanema	1	-
2	Centro, Barra	0,9	10%
3	Urca, (Leblon, Ipanema)	0,9	0%
4	Tijuca, Catete	0,9	0%
5	Saúde, (Centro, Barra)	0,8	11,11%
6	(Saúde, Centro, Barra), (Urca, Leblon, Ipanema)	0,7	12,5%
7	Ilha, (Saúde, Centro, Barra, Urca, Leblon, Ipanema)	0,5	40%
8	Glória, (Ilha, Saúde, Centro, Barra, Urca, Leblon, Ipanema)	0,4	20%

9	(Glória, Ilha, Saúde, Centro, Barra, Urca, Leblon, Ipanema), (Tijuca, Catete)	0,2	50%

Tabela 3.6: Agrupamentos gerados utilizando a matriz de distâncias construída através do coeficiente Sokal e Michenner

O Coeficiente Cofenético gerado para essa Análise de Agrupamento atingiu o valor **0,78234**. Por estar bem próximo de 0,80 consideramos que essa Análise de Agrupamento atingiu seus objetivos.

Os grupos homogêneos (compostos por mais de um avaliador) gerados em cada passo da análise de agrupamento, para uma variação máxima de 15%, foram os seguintes:

(Leblon, Ipanema);

(Centro, Barra);

(Urca, Leblon, Ipanema);

(Tijuca, Catete);

(Saúde, Centro, Barra);

(Saúde, Centro, Barra, Urca, Leblon, Ipanema).

3.1.4.4 Análise de Agrupamento para o Coeficiente de Jaccard

Esse coeficiente fornece a proporção de concordâncias positivas entre os elementos através da seguinte fórmula: $\frac{a}{a+b+c}$

A aplicação do método da Ligação Completa nos dados contidos na matriz do quadro anterior, gerou os seguintes agrupamentos, exibidos na Tabela 3.7.

Passo	Grupo	Nível	Diferença percentual absoluta para o nível anterior
1	Leblon, Ipanema	1	
2	Urca, (Leblon, Ipanema)	0,875	12,50%
	Centro, Barra	0,833	4,80%
4	Ilha, (Urca, Leblon, Ipanema)	0,667	19,93%
5	Tijuca, Catete	0,667	0,00%
6	Saúde, (Centro, Barra)	0,667	0,00%
7	(Ilha, Urca, Leblon, Ipanema), (Saúde, Centro, Barra)	0,444	33,43%
8	Glória, (Ilha, Urca, Leblon, Ipanema, Saúde, Centro, Barra)	0,286	35,59%
9	(Tijuca, Catete), (Glória, Ilha, Urca, Leblon, Ipanema, Saúde, Centro, Barra)	0,000	100,00%

Tabela 3.7: Agrupamentos gerados pela Ligação Completa utilizando Jaccard

O Coeficiente Cofenético gerado para essa Análise de Agrupamento atingiu **0,83709**, um valor adequado se comparado a 0,80, valor mínimo para o qual se considera que o dendrograma reproduziu, de uma forma confiável, a matriz de distâncias, o que nos permitiu assegurar que a Análise de Agrupamento aplicada atingiu seus objetivos. Veja a seguir o dendrograma gerado:

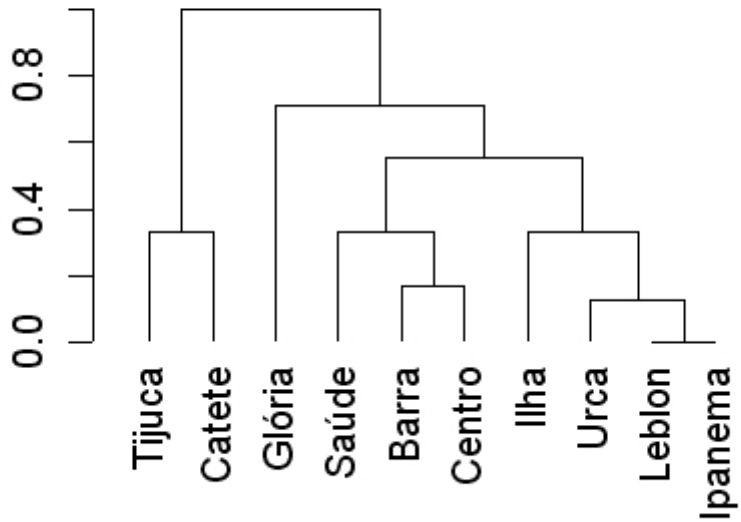


Figura 3.2 – Dendrograma referente aos Agrupamentos da Tabela 3.7

Os grupos gerados apresentam os componentes mais similares, em relação à recomendação positiva dos artigos. Para uma variação máxima de 15%, os grupos homogêneos (compostos por mais de um avaliador) gerados em cada passo da análise de agrupamento foram os seguintes:

(Leblon, Ipanema);

(Urca, Leblon, Ipanema);

(Centro, Barra);

(Tijuca, Catete);

(Saúde, Centro, Barra).

Se tivéssemos optado para uma variação máxima de 20%, o grupo a seguir seria adicionado aos anteriores:

(Ilha, Urca, Leblon, Ipanema);

3.1.4.5 Análise de Agrupamento utilizando Russel e Rao

O coeficiente de similaridade Russel e Rao fornece a proporção das concordâncias positivas, representado pelo par (1,1), em relação ao total das incidências de todos os pares. A fórmula $\frac{a}{a+b+c+d}$ expressa esse coeficiente, que varia entre 0 (os elementos não são similares) e 1 (elementos similares).

A aplicação do método da Ligação Completa gerou os agrupamentos, exibidos na Tabela 3.8.

Passo	Grupo	Nível	Diferença percentual absoluta para o nível anterior
1	Leblon, Ipanema	0,7	-
2	Ilha, Urca	0,7	0%
	Centro, Leblon, Ipanema	0,6	14,3%
4	Barra, (Centro, Leblon, Ipanema)	0,5	16,7%
5	Saúde, (Ilha, Urca)	0,5	0%
6	(Barra, Centro, Leblon, Ipanema),(Saúde, Ilha, Urca)	0,4	20%
7	Tijuca, (Barra, Centro, Leblon, Ipanema, Saúde, Ilha, Urca)	0,2	50%
8	Catete, (Tijuca, Barra, Centro, Leblon, Ipanema, Saúde, Ilha, Urca)	0,1	50%
9	Glória, (Catete, Tijuca, Barra, Centro, Leblon, Ipanema, Saúde, Ilha, Urca)	0	100%

Tabela 3.8: Agrupamentos gerados pela Ligação Completa utilizando Russel e RAO

O Coeficiente Cofenético gerado para essa Análise de Agrupamento atingiu o valor **0,82220**, um valor satisfatório, que foi aceito indicando que o dendrograma produzido (vide Apêndice B) está de acordo com a matriz de distâncias.

Os grupos homogêneos gerados pela Análise de Agrupamento, contém os avaliadores mais parecidos quando da recomendação positiva dos artigos, representada pelo par (1,1).

Para uma variação máxima de 15%, os grupos homogêneos (compostos por mais de um avaliador) gerados em cada passo da análise de agrupamento foram os seguintes:

(Leblon, Ipanema);

(Ilha, Urca);

(Centro, Leblon, Ipanema).

E para uma variação máxima de 20%, foram gerados os seguintes grupos (compostos por mais de um avaliador) em cada passo da análise de agrupamento:

(Leblon, Ipanema);

(Ilha, Urca);

(Centro, Leblon, Ipanema);

(Barra, Centro, Leblon, Ipanema);

(Saúde, Ilha, Urca);

(Barra, Centro, Leblon, Ipanema, Saúde, Ilha, Urca).

3.1.4.6 Análise de Agrupamento utilizando Hamann

O coeficiente de Hamann exprime a proporção resultante da divisão das concordâncias positivas, número de pares (1,1), descontadas as concordâncias negativas, pares (0,0), pelo total relativo à incidência de todos os pares ($a + b + c + d$). A fórmula do coeficiente de Hamann é expressa por $\frac{a}{a+b+c+d}$, gerando valores entre –1, total discordância entre os elementos, e 1, concordância perfeita entre os elementos. O coeficiente de Hamann é uma medida de similaridade.

A aplicação do método da Ligação Completa, na matriz de distâncias gerada através do coeficiente de Hamann, gerou os seguintes agrupamentos, exibidos na Tabela 3.9.

Passo	Grupo	Nível	Diferença percentual absoluta para o nível anterior
1	Leblon, Ipanema	1,0	-
2	Barra, Centro	0,8	20%
	Tijuca, Catete	0,8	0%
4	Urca, (Leblon, Ipanema)	0,8	0%
5	Saúde, (Barra, Centro)	0,6	25%
6	Ilha, (Urca, Leblon, Ipanema)	0,4	50%
7	Glória, (Saúde, Barra, Centro)	0,0	-
8	(Tijuca, Catete), (Ilha, Urca, Leblon, Ipanema)	-0,2	100%
9	(Tijuca, Catete, Ilha, Urca, Leblon, Ipanema), (Ilha, Urca, Leblon, Ipanema)	-0,6	200%

Tabela 3.9: Agrupamentos gerados pela Ligação Completa utilizando Hamann

O Coeficiente Cofenético gerado para essa Análise de Agrupamento atingiu o valor **0,43528**, longe do ideal (0,80), indicando que a matriz cofenética, gerada a partir do

dendrograma, não representou uma boa simplificação da matriz de distâncias e, portanto, a análise de agrupamento efetuada não forneceu grupos homogêneos adequados.

Os únicos grupos homogêneos (compostos por mais de um avaliador) gerados em cada passo da análise de agrupamento, para o patamar de 20%, foram os seguintes:

(Leblon, Ipanema);

(Barra, Centro);

(Tijuca, Catete);

(Urca, Leblon, Ipanema).

3.1.4.7 Análise de Agrupamento utilizando Yule

O Coeficiente de Yule, expresso pela fórmula $\frac{ad - bc}{ad + bc}$, é um coeficiente de similaridade que assume valores pertencentes ao intervalo [-1,1]. Quando b+c=0, Yule assume 1, significando total concordância entre os elementos. Assumirá –1 quando a+d=0, significando completa discordância entre os elementos.

Os grupos gerados a partir da aplicação do método da Ligação Completa, na matriz de distâncias produzida pelo coeficiente de Yule, encontram-se na Tabela 3.10.

Passo	Grupo	Nível	Diferença percentual absoluta para o nível anterior
1	Leblon, Ipanema	1,0	-
2	Barra, Centro	1,0	0%
	Urca, (Leblon, Ipanema)	1,0	0%
4	Tijuca, Ilha	1,0	0%
5	Saúde, Catete	1,0	0%
6	Glória, (Urca, Leblon, Ipanema)	1,0	0%
7	(Tijuca, Ilha), (Saúde, Catete)	1,0	0%
8	(Barra, Centro), (Glória, Urca, Leblon, Ipanema)	0,5	50%
9	(Tijuca, Ilha, Saúde, Catete), (Barra, Centro, Glória, Urca, Leblon, Ipanema)	-1,0	300%

Tabela 3.10: Agrupamentos gerados pela Ligação Completa utilizando Yule

O Coeficiente Cofenético gerado para essa Análise de Agrupamento atingiu o valor **0,42066**, longe do ideal (0,80). Esse valor indica que a matriz cofenética, gerada a partir do dendrograma, não representou uma boa simplificação da matriz de distâncias e, portanto, os grupos gerados pela análise de agrupamento têm que ser analisados com cuidado.

Dos nove agrupamentos formados, sete foram produzidos em um mesmo nível (através do valor máximo, 1), isso ocorreu devido ao fato do coeficiente de Yule ser bastante afetado quando um dos índices (*a, b, c* ou *d*) é nulo, e isso ocorre em 34 índices presentes na composição dos 45 coeficientes de Yule. Foi necessária a verificação minuciosa de cada um dos sete grupos formados, que podem vir a ser homogêneo, buscando identificar aqueles que apresentam na formação do coeficiente, um dos índices *a* ou *d* nulo, o que descaracteriza a homogeneidade de tal grupo.

Essa verificação confirmou os seguintes sete grupos homogêneos (compostos por mais de um avaliador) gerados em cada passo da análise de agrupamento:

(Leblon, Ipanema);

(Barra, Centro);

(Urca, Leblon, Ipanema);

(Tijuca, Ilha);

(Saúde, Catete);

(Glória, Urca, Leblon, Ipanema);

(Tijuca, Ilha, Saúde, Catete).

3.1.4.8 Análise de Agrupamento Utilizando o coeficiente Gower2

O Coeficiente Gower2 produz valores pertencentes ao intervalo [0,1], e é expresso pela fórmula $\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$. É um coeficiente de similaridade que considera as concordâncias positivas, pares (1,1), e negativas, pares (0,0). Quanto mais perto da unidade, o valor calculado através desse coeficiente estiver, maior será a semelhança entre os elementos (no nosso caso, mais semelhante serão em suas avaliações positivas ou negativas). Quanto mais perto de zero, mais discordantes eles se apresentarão. Esse coeficiente também é associado a Ochiai, na literatura consultada.

Os grupos gerados a partir da aplicação do método da Ligação Completa, na matriz de distâncias produzida pelo coeficiente Gower2, encontram-se na Tabela 3.11.

Passo	Grupo	Nível	Diferença percentual absoluta para o nível anterior
1	Leblon, Ipanema	1,000	-
2	Barra, Centro	0,816	18,40%
	Urca, (Leblon, Ipanema)	0,764	6,37%
4	Tijuca, Catete	0,764	0,00%
5	Saúde, (Barra, Centro)	0,64	16,23%
6	(Saúde, Barra, Centro), (Urca, Leblon, Ipanema)	0,500	21,88%
7	Ilha, (Tijuca, Catete)	0,250	50,00%
8	Glória, (Barra, Centro, Saúde, Urca, Leblon, Ipanema)	0,245	2,00%
9	(Ilha, Tijuca, Catete), (Glória, Barra, Centro, Saúde, Urca, Leblon, Ipanema)	0,000	100,00%

Tabela 3.11: Agrupamentos gerados pela Ligação Completa utilizando Gower2

O Coeficiente Cofenético calculado foi **0,80459**, valor considerado aceitável, indicando que a matriz cofenética, gerada a partir do dendrograma, produziu uma boa simplificação da matriz de distâncias e, portanto, a análise de agrupamento efetuada poderá vir a ter seus resultados fornecidos e utilizados.

Para uma variação máxima de 15%, os grupos homogêneos (compostos por mais de um avaliador) gerados em cada passo da análise de agrupamento foram os seguintes:

(Leblon, Ipanema);

(Urca, Leblon, Ipanema);

(Tijuca, Catete).

E para uma variação máxima de 20%, foram gerados em adicional aos anteriores, os seguintes grupos:

(Barra, Centro);

(Saúde, Barra, Centro).

3.1.5 Conclusões

As Análises de Agrupamento efetuadas, com diferentes coeficientes de similaridade, produziram o seguinte quadro comparativo.

Coeficiente de Similaridade	Total de Grupos Homogêneos Produzidos	Coeficiente de Correlação Cofenética
Jaccard	6	0,83709
Russel e Rao	6	0,82220
Gower2	5	0,80459
Concordância Simples	6	0,78234
Pearson	5	0,67520
Hamann	4	0,43528
Yule	7	0,42066

Quadro XIII – Comparando os dados gerados pelas Análises de Agrupamento

A partir do Quadro XIII, foram escolhidos os resultados das Análises de Agrupamento que utilizaram os coeficientes Jaccard, Russel e Rao, Gower2 e Concordância Simples, para fornecimento aos professores da disciplina. Essas foram as análises que apresentaram os

melhores coeficientes cofenéticos, juntamente com grupos homogêneos de participantes similares na recomendação (positiva ou negativa) dos artigos. Dos 14 distintos grupos formados em todas as sete aplicações, quatro deles aparecem em mais da metade dos agrupamentos gerados. O coeficiente Yule foi o que mais gerou grupos próprios, ou seja, somente aparecem no agrupamento feito através dele. Esse coeficiente produziu quatro grupos distintos, dos 14 construídos em todas as análises. Podemos concluir que os grupos gerados para os coeficientes Jaccard, Russel e Rao, Gower2 e Concordâncias Simples são semelhantes.

Quanto à viabilidade da implementação de técnicas de Análise de Agrupamento no ambiente TeamWorks, que suportou a avaliação dos artigos, este não apresentou um arquivo de Log mínimo e suficiente, necessário à implementação de rotinas destinadas a execução de análises de agrupamento. Por conta da falta desse Log, a construção das análises que foram efetuadas apresentou-se bastante complexa, sendo necessárias muitas idas e vindas, entre diferentes aplicativos, para que os dados coletados fossem alterados e transformados. As tarefas executadas, buscando solucionar a falta do arquivo Log, possibilitou a identificação de mudanças, consideradas de rápida implementação, no ambiente, que estariam sedimentando o mesmo para a construção de um arquivo de Log, e posterior programação das rotinas de Análise de Agrupamento. A lista descrevendo as mudanças detectadas, foi entregue aos professores da disciplina em questão.

Para finalizar, gostaríamos de ressaltar que as análises de agrupamento executadas com diferentes coeficientes, foram bastante proveitosas e esclarecedoras pois, além de proporcionar o acompanhamento de diferentes agrupamentos, nos ajudaram a consolidar uma sistemática de análise, que utiliza dendrograma, coeficiente de correlação, e patamares máximos de variação percentual, na avaliação de cada análise de agrupamento.

3.2 Estudo de Caso 2 – Análise de Agrupamento aplicada aos Alunos de um Curso de Física da 4ª Série

3.2.1 Introdução

Esta seção apresenta a Análise de Agrupamento aplicada nos dados de acesso de 47 alunos da quarta-série do ensino fundamental, da Escola Técnica Ferreira Viana[39] (ETEFEV), localizada na cidade do Rio de Janeiro. Essa Análise de Agrupamento objetivou a formação dos grupos homogêneos dos alunos.

A demanda que gerou essa aplicação é justificada por uma necessidade apresentada pelo professor responsável pela disciplina de Física ministrada para os alunos, que precisava gerar grupos de alunos semelhantes, baseado na freqüência destes durante a utilização da plataforma educacional que apoiava o ensino presencial da disciplina. A plataforma em questão, era a Plataforma Interativa para Internet (Pii).

O professor do curso esperava que a geração dos grupos através da Análise de Agrupamento, permitisse:

I. Fazer comparações com os grupos formados na sala de aula que desenvolveram determinadas tarefas.

II. Identificar outras características que surgiram no ambiente de ensino à distância utilizado como apoio.

III. A partir dos grupos gerados, implementar outras tarefas e outros grupos.

A execução da Análise de Agrupamento que será apresentada nessa seção, permitiu identificar os requisitos necessários para implementar uma nova interface na plataforma EAD

[39] A ETEFEV é uma unidade estadual da Fundação de Apoio à Escola Técnica (FAETEC).

utilizada, proporcionando a geração de agrupamentos para qualquer curso suportado pela mesma.

3.2.2 Os Dados Gerados pela Plataforma Pii

A seguir, é feita uma descrição suficiente da Plataforma Pii:

> A Plataforma Interativa para Internet (Pii) (Elia & Sampaio, 2001) vem sendo utilizada para apoiar cursos que necessitam de uma plataforma educacional baseada na Web e conta com vários recursos para comunicação, pesquisa e administração. Sob o ponto de vista do programa de Pós-Graduação IM-NCE/UFRJ, a plataforma Pii vem se constituindo em um laboratório para a avaliação de idéias e projetos de estudantes e pesquisadores. A referida Plataforma funciona num servidor instalado no Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro (NCE/UFRJ), pode ser acessada pela internet em http://www.nce.ufrj.br/pii, possuindo mais de 25 cursos publicados. (CHAMOVITZ *et al*, 2004),

A maioria das plataformas de EAD, registra em um arquivo físico grande parte dos eventos gerados pelos alunos que as utilizam. Esse arquivo é conhecido por Log, e a plataforma Pii também tem o seu, representado por um arquivo de banco de dados com diversas tabelas.

Nesse banco de dados, são registradas as interações de cada aluno dos cursos suportados e hospedados pela Pii com o ambiente. A Pii utiliza diversos mecanismos – registro de sessões, de avaliações, rastreamento, registro de visitas a determinadas páginas, acesso a recursos, data e hora de entrada e saída, estado de visibilidade do aluno, etc. – para registrar os eventos em geral dos participantes quando na plataforma, alimentando esse arquivo de banco de dados. Toda essa geração acontece quando o participante está interagindo na plataforma. Uma amostra do conteúdo desse Log, associado à utilização do curso de física para a quarta-série, é exibido no Quadro XIV, a seguir.

Indice	Aluno	UDidatica	HOST_Aluno	ADDR_Aluno	Data	Hora	Log	Visibilidade	IDSessao	Curso
22407	A	0	200.222.72.170	200.222.72.170	31-mai-04	00-01-1900 19:24:00	1	1	668931739	32
21993	B	0	200.151.101.216	200.151.101.216	23-mai-04	00-01-1900 17:38:16	1	1	690674800	32
21287	C	2	200.222.72.170	200.222.72.170	10-mai-04	00-01-1900 19:01:22	6	1	1056666676	32

Quadro XIV – Amostra extraída dos Log da Plataforma Pii

O Log da Pii é utilizado para diversos fins, não havendo um pré-tratamento adequado para a aplicação das técnicas de análise de agrupamento. Por conta disso foram executados tratamentos e análises nas informações, com o intuito de operacionalizar os dados, adequando-os para que as técnicas de Análise de Agrupamento pudessem ser aplicadas. Os tratamentos e análises ocorreram em aplicativos externos, Excel e Lotus Notes, que contavam com linguagens de programação ou funções adequadas.

3.2.3 Descrevendo o Curso de Física na Pii e o Log Utilizado

O Curso de Física ministrado com o auxílio da Pii, apresentou-se segmentado em cinco unidades didáticas, indicadas dentro do arquivo Log pelo campo UDidatica e associadas aos valores 1, 2, 3, 4 e 5. Os alunos quando entram na plataforma, encontram-se no ambiente da mesma, que é externo à área didática que compõe o curso. Essa condição externa está representada pelo valor 0, no campo UDidatica. Durante o pré-tratamento, esse campo foi desmembrado em variáveis UD0 a UD5, descritas mais adiante.

Os dados coletados referentes ao acesso dos estudantes pertencem ao período compreendido entre 16 de fevereiro e 23 de junho de 2004, durante o qual ocorreu a disciplina na ETEFEV.

O Quadro XIV apresenta uma amostra dos dados contidos na tabela utilizada do banco de dados, de onde se extraiu os dados para o pré-tratamento e posterior aplicação das técnicas de Análise de Agrupamento. Esses dados, que podemos chamar de brutos, estavam representados por 11 variáveis ao todo, uma breve descrição das variáveis de interesse utilizadas encontra-se a seguir:

Indice – variável seqüencial que representa o número do registro, sendo uma das chaves primárias do banco de dados do *Log*.

Aluno – contém o *e-mail* do aluno associado à variável *Indice*. Por conveniência os *e-mails* foram suprimidos do artigo e substituídos por caracteres alfanuméricos distintos.

UDidatica – representa em qual área (didática ou externa a ela) se encontra o aluno no interior da plataforma Pii. Seus valores variam de 0 a 5 no caso do curso de Física da ETEFEV.

HOST_Aluno – contém o número do IP (*internet protocol*) referente à conexão internet que o aluno está usufruindo.

Data e **Hora** – informam em que data e Hora ocorreram o acesso à plataforma Pii em relação ao valor contido pela variável Log do Aluno.

Log – variável que discrimina que ação foi tomada pelo aluno no interior da plataforma. Seus valores e significados são:

0 – Sair da plataforma.

1 – Entrar na plataforma.

3 – Presença do aluno visível para outros alunos.

4 – Presença do aluno invisível para outros alunos.

5 – Recorrer a recursos de comunicação da plataforma.

6 – Ocorreu mudança de unidade didática.

Visibilidade – informa se naquele momento o aluno encontra-se visível (1) ou invisível (0).

Curso – variável numérica que informa a que curso aquele acesso corresponde.

3.2.4 Operacionalização dos Dados Contidos no Log

Foi executado um processo de tratamento e ajuste (KOHAVI, 2001) no Log contendo os 763 registros referentes ao acesso dos alunos do curso de física, para que fosse produzido um outro formato para os dados. Esse novo formato pode ser interpretado como um Log reduzido, apresentando os dados de uma forma adequada à aplicação do agrupamento. Esse processo consistiu em compor uma análise que utilizou determinados critérios para considerar um registro válido para a análise de agrupamento que seria executada mais adiante. Esses critérios verificavam em sua essência se um aluno, que tenha tido sua entrada na plataforma registrada, teve também registrada a sua saída no mesmo dia. Quando registros de alunos que atendiam a esses critérios eram detectados, contabilizavam-se os dados nas novas variáveis produzidas para a aplicação da Análise de Agrupamento.

Esse processo reduziu os dados em exatamente 42 registros válidos de alunos, contendo cada um deles oito variáveis: *Aluno* (presente no *Log* original) e outras sete novas (*UD0, UD1, UD2, UD3, UD4* e *UD5*) e o Tempo Total de Acesso em horas (TTAH). As variáveis *Indice*, *HOST_Aluno, ADDR_Aluno, IDSessao* e *Curso* foram descartadas por não serem necessárias à aplicação das técnicas de Análise de Agrupamento. Foram descartados os poucos registros (cerca de 4,3%) que não apresentaram a variável *Aluno* preenchida de forma que pudesse ser identificada.

A variável original *UDidatica* foi utilizada para a construção de seis novas variáveis (*UD0* a *UD5*). A variável UD0, detalhada mais adiante, representa o acesso feito fora das

unidades didáticas. As variáveis *UD1* a *UD5* totalizam o número de acessos de cada a aluno em cada unidade didática do curso em questão. Era de interesse do agrupamento agrupar os alunos que tinham acessado uma ou mais unidades didáticas e, por conta desse objetivo, foram descartados quatro registros que não apresentaram acesso a alguma unidade didática, referente as variáveis UD1 a UD5.

A variável *Tempo Total* informa o tempo total gasto de cada aluno na plataforma Pii, no período coberto pelo Log (pouco mais de quatro meses). A construção da variável *Tempo Total (horas)* foi a mais complexa, pois envolveu análises e cálculos feitos nas variáveis originais *Data*, *Hora* e *Log* para detectar as saídas indevidas da plataforma, como por exemplo por "ctrl + alt + Del" ou por abandono (*timeout*) que só são controladas externamente ao arquivo de *Log* da Pii pelo arquivo Global ASA. Na Tabela 3.12, uma amostra desses dados, após o tratamento, é mostrada.

Aluno	UD0	UD1	UD2	UD3	UD4	UD5	TTAH
A	1	0	0	1	0	0	0,05
B	8	0	0	2	4	0	1,13
C	9	0	0	1	3	0	1,93
D	3	2	1	0	0	0	0,07
E	5	0	0	2	0	1	1,00

Tabela 3.12: Amostra extraída dos 42 registros compreendidos no *Log* tratado

Como forma de subsidiar o professor com os dados que surgiram desse pré-tratamento anterior à aplicação da Análise de Agrupamento, foram gerados e fornecidos ao mesmo alguns gráficos de barra e tabelas. Isso foi feito com o intuito de permitir ao professor a oportunidade da construção de alguma crítica, como por exemplo a exclusão de alguma variável UD, a ser aplicada antes da Análise de Agrupamento.

A Tabela 3.13 a seguir, exibe os totais de acessos ao ambiente inicial da plataforma e às respectivas unidades didáticas, representadas pelas variáveis *UD0* a *UD5*.

UD0	UD1	UD2	UD3	UD4	UD5
173	28	11	11	20	7

Tabela 3.13: Totais de Acesso às Unidades

É fácil notar que a variável UD0 apresenta uma incidência bem maior que as outras. Esta variável representa o acesso dos alunos do Curso de Física somente a área da plataforma Pii pertencente ao curso, mas externa às unidades didáticas. Essa área contém diversas funcionalidades inerentes a uma plataforma virtual (*chat*, *e-mail*, gerenciamento de arquivos, etc.) e é caminho obrigatório para acessar as unidades didáticas.

As Figuras 3.3 e 3.4 possibilitam uma análise gráfica da presença e da ausência da variável UD0 nos dados.

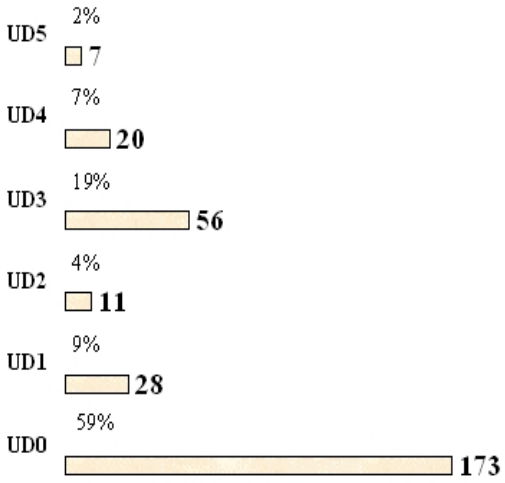


Figura 3.3 – Totais e percentuais das variáveis UD0 a UD5.

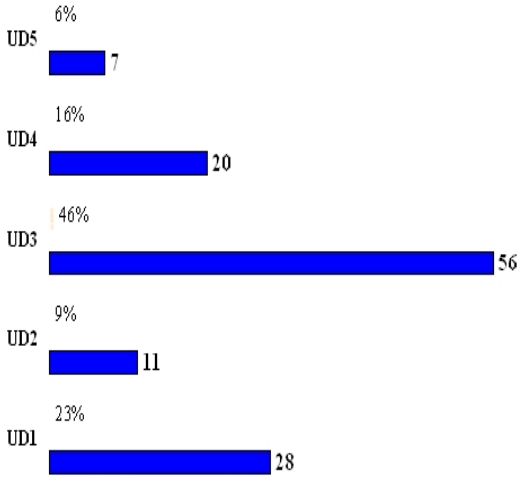


Figura 3.4 – Totais e percentuais das variáveis UD1 a UD5.

Podemos observar na Figura 3.3, que a variável UD0 representa quase 60% dos totais de acessos que foram coletados.

De acordo com a Tabela 3.13 e as Figuras 3.3 e 3.4, as unidades 2 e 5, representadas pelas variáveis UD2 e UD5, são as unidades didáticas que menos foram acessadas pelos alunos. As unidades 1, 3 e 4 (UD1, UD3 e UD4) foram as mais acessadas. Ao se inspecionar apenas o acesso às unidades didáticas, representado na Figura 3.4, observa-se que a variável UD3 é a unidade didática que sofreu mais acessos, tendo o dobro de acessos da segunda colocada, UD1, seguida de perto pela UD4. As unidades didáticas 2 e 5 (UD2 e UD5) tiveram juntas cerca de 15% de participação nos totais de acesso.

Essas informações, em conjunto com outras análises (exploratória, fatorial, análise de componentes) podem ser úteis quando da definição das variáveis que participarão do agrupamento.

O professor do curso, apesar de ter construído hipóteses devido a essas informações, resolveu seguir adiante com a aplicação da análise de agrupamento abrangendo todas as sete novas variáveis. Como ocorreu essa análise veremos a seguir.

3.2.5 Análise de Agrupamento

Em geral, os coeficientes de dissimilaridade[40] são mais adequados para as variáveis quantitativas (BUSSAB *et al*, 1990) para avaliar a proximidade entre dois objetos quaisquer. As variáveis existentes são quantitativas e como existe uma razoável incidência de valores nulos em algumas UDs, foi escolhido o coeficiente de dissimilaridade Distância Euclidiana Média – DEM, (BARROSO & ARTES, 2003) que é a Distância Euclidiana definida para um espaço de dimensão *p,* a saber:

$$d(A,B) = \left[\sum_{i=1}^{p} ((x_i(A) - x_i(B))^2 \Big/ p \right]^{1/2}$$

Ao calcularmos a distância entre os alunos E e F, d(E,F), utilizando os dados brutos temos:

$$\mathrm{d(E,F)} = \left[((5-4)^2 + (0-0)^2 + (0-0)^2 + (2-1)^2 + (0-0)^2 + (1-0)^2 + (1{,}00-1{,}14)^2) \Big/ 7 \right]^{0,5} = [0{,}4314]^{0,5} = 0{,}66$$

Notamos que a variável *Tempo Total* apresentou-se irrelevante ao cálculo da DEM (desprezando Tempo Total teríamos d(E,F) = $[0{,}4286]^{0,5}$=0,66).

Entretanto, a variável TTAH foi ressaltada pelo professor como de suma importância para as suas avaliações, sendo necessária sua análise para os objetivos do agrupamento, por medir o tempo total que cada aluno dedicou à plataforma no período analisado. Para que a magnitude das quantidades mensuradas seja comparável, permitindo assim uma contribuição equilibrada de cada uma delas no cálculo do coeficiente escolhido, foi feita uma relativização das variáveis envolvidas.

---

[40] Considera-se coeficiente de dissimilaridade aquele onde para o maior valor observado, menos parecidos serão os objetos, significando que o menor valor calculado indicará a maior semelhança entre os objetos. A Distância Euclidiana e outras derivadas dela são coeficientes de dissimilaridade.

Os valores apresentados pelas variáveis foram transformados em uma escala padrão Z, visando reduzir o efeito de escalas diferentes. Transformou-se cada valor das variáveis x_{ik} (onde *i* varia de 1 a 42, representando os valores daquela variável associada a cada aluno, e *k* varia de 1 a 7 representando cada uma das variáveis) em uma nova variável Z_{ik}, representada pela fórmula $Z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}$, onde μ_k é a média da variável *k* e σ_k seu desvio padrão. A Tabela 3.14 a seguir, exibe parte das variáveis com as respectivas transformações.

Aluno	Z_{UD0}	Z_{UD1}	Z_{UD2}	Z_{UD3}	Z_{UD4}	Z_{UD5}	$Z_{TempoTotal}$
A	-1,18	-1,09	-0,59	-0,31	-0,54	-0,44	-1,14
B	1,58	-1,09	-0,59	0,62	3,96	-0,44	0,78
L	-1,18	0,55	-0,59	-0,31	-0,54	-0,44	-0,90
M	0,00	0,55	-0,59	2,48	0,59	-0,44	-1,14
N	-1,18	-1,09	-0,59	-1,24	0,59	-0,44	-0,74
O	-0,79	-1,09	-0,59	-1,24	0,59	-0,44	-0,89
D1	-1,18	-1,09	-0,59	-0,31	-0,54	-0,44	-1,14

Tabela 3.14: Amostra das Variáveis Normalizadas

Os novos dados das variáveis foram então utilizados para construir a Matriz de Distâncias, a partir da qual surgem os grupos.

Da Tabela 3.15, onde se encontra uma amostra dessa matriz (a original tem 41 linhas por 41 colunas) é identificado o primeiro agrupamento, formado pelos elementos mais similares, alunos A e D1. Como ambos possuem os mesmos valores para suas variáveis (vide Tabela 3.14), a Distância Euclidiana Média calculada entre os dois é nula, sendo a menor distância existente na matriz de dissimilaridade. Portanto esses dois alunos podem ser considerados como o primeiro grupo. Logo após a Tabela 3.15, é mostrado o cálculo da DEM entre os alunos A e B.

	A	B	C	D	E	F	G	H
B	**2,152**							
C	**2,153**	**0,780**						
D	**1,570**	**2,588**	**2,520**					
E	**1,372**	**2,026**	**1,872**	2,054				
F	**0,855**	**1,837**	**1,571**	1,705	**1,076**			
G	**2,093**	**2,175**	**1,818**	1,837	**1,775**	1,521		
H	**1,275**	**2,221**	**2,021**	0,764	**1,632**	1,146	**1,630**	
I	**1,734**	**1,386**	**1,385**	1,508	**1,692**	1,430	**1,498**	1,173
D1	**0,000**	**2,152**	**2,153**	1,570	**1,372**	0,855	**2,093**	1,275

Tabela 3.15: Amostra da Matriz de Dissimilaridade

d(A,B) = $[((-1{,}18-1{,}58)^2+(-1{,}09+1{,}09)^2+(-0{,}59+0{,}59)^2+(-0{,}31-0{,}62)^2+(-0{,}54-3{,}96)^2+(-0{,}44+0{,}44)^2+(-1{,}14-0{,}78)^2)/7]$ = **2,152**

Os alunos **A** e **D1** que serão agrupados através do coeficiente 0,000. Após esse agrupamento a matriz de distâncias foi refeita, passando a ter uma coluna ou linha a menos (sempre que ocorrer um agrupamento, a dimensão da matriz irá diminuir), o que implicou no recálculo de todas as distâncias, exceto em casos como esse primeiro agrupamento, onde as variáveis dos elementos que formou o grupo apresentaram valores iguais.

Foi escolhido, para essa Análise de Agrupamento, o algoritmo hierárquico Método da Ligação Simples (*Single Linkage*) ou Método do Vizinho mais Próximo. Esse método é utilizado no recálculo das distâncias, e define a nova distância entre dois grupos – composto por um ou mais objetos – como sendo aquela representada pela menor distância entre todas as distâncias envolvidas. Uma alternativa gráfica à utilização do método de agrupamento, seria a análise visual do diagrama de dispersão, aplicado aos dados da Matriz de Dissimilaridade completa de onde foi retirada a amostra da Tabela 3.15, buscando identificar os grupos naturais. Através do Gráfico 3.1, é possível verificar a complexidade de um agrupamento visual das 820 distâncias que compõem a matriz.

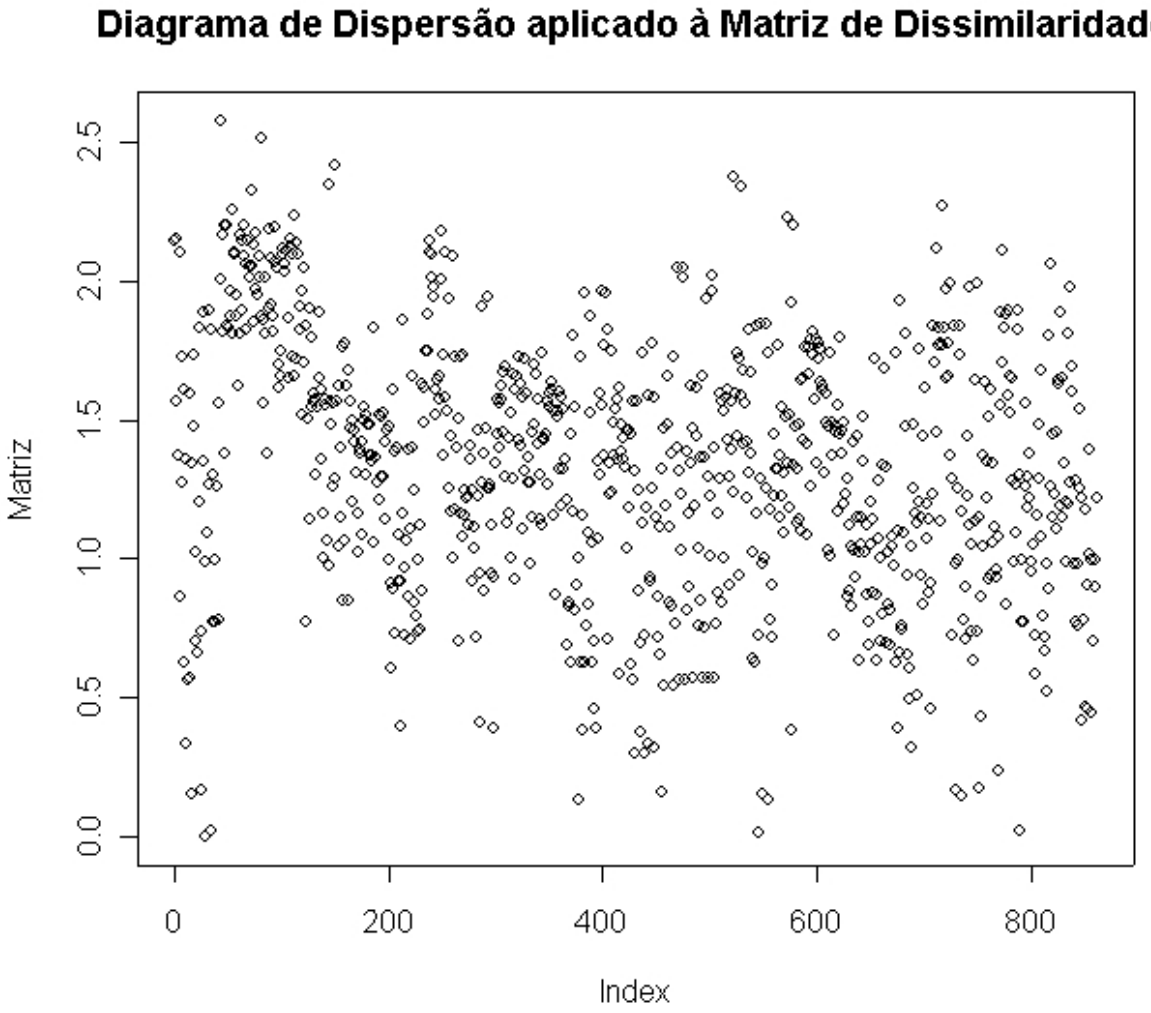


Gráfico 3.1 – Diagrama de Dispersão aplicada à Matriz de Dissimilaridade.

Por exemplo, no passo 2 do agrupamento mostrado no Quadro XV, os alunos Q e Z foram agrupados por conta de apresentarem a menor distância (0,011), para recalcular a distância entre esse novo grupo QZ e um outro grupo (que no momento pode ser composto por um único aluno, por exemplo, aluno B) qualquer da matriz fez-se:

$$d(QZ, B) = \min(d(Q, B); d(Z, B)) = \min(2{,}106; 2{,}103) = 2{,}103$$

Esse procedimento é repetido entre o grupo QZ e todos os demais elementos, e assim se repete quando um novo agrupamento se formar.

Assim, generalizando para dois grupos quaisquer α e β, contendo respectivamente i e j objetos, a distância entre eles será definida como:

$$d(Grupo_{\alpha}, Grupo_{\beta}) = \min\left(d(Grupo_{\alpha}(i), Grupo_{\beta}(j)\right): \textit{para todo i e j}$$

No Quadro XV, são mostrados os passos que determinaram a formação dos até o último agrupamento, após várias etapas intermediárias marcadas principalmente pelo recálculo das distâncias existentes na matriz, onde foram reunidos em um único grupo todos os alunos.

Passo	**Junção**	Nível	Grupos	Dif
1	A, D1	0,000	2	-
2	Q, Z	0,011	3	-
3	A, D1, J1	0,020	3	82%
4	W, K	0,132	4	560%
5	A, D1, J1, Q, Z	0,134	3	2%
6	O, N	0,160	4	19%
7	L1, A1	0,178	5	11%
8	O1, B1	0,239	6	34%
9	A, D1, J1, Q, Z, M	0,300	6	26%
10	N1, W, K	0,324	6	8%
11	V, A, D1, J1, Q, Z, M	0,376	6	16%
12	L1, A1, N1, W, K	0,382	5	2%
13	G1, R	0,385	6	1%
14	O1,B1,H	0,386	6	0%
15	V,A,D1,J1,Q,Z,M,F	0,392	6	2%
16	M1,K1	0,420	7	7%
17	M1,K1,L1,A1,N1,W,K	0,445	6	6%
18	X,O1,B1,H	0,459	6	3%
19	P1,M1,K1,L1,A1,N1,W,K	0,465	6	1%
20	F1,P1,M1,K1,L1,A1,N1,W,K	0,518	6	11%
21	O,N,V,A,D1,J1,Q,Z,M,F	0,546	5	5%
22	E1,F1,P1,M1,K1,L1,A1,N1,W,K	0,582	5	7%
23	L,E1,F1,P1,M1,K1,L1,A1,N1,W,K	0,587	5	1%
24	L,E1,F1,P1,M1,K1,L1,A1,N1,W,K,O,N,V,A,D1,J1,Q,Z,M,F	0,623	4	6%
25	U,L,E1,F1,P1,M1,K1,L1,A1,N1,W,K,O,N,V,A,D1,J1,Q,Z,M,F	0,632	4	1%
26	G,U,L,E1,F1,P1,M1,K1,L1,A1,N1,W,K,O,N,V,A,D1,J1,Q,Z,M,F	0,699	4	11%
27	T,X,O1,B1,H	0,724	4	4%
28	D,T,X,O1,B1,H	0,768	4	6%
29	C,B	0,779	5	1%
30	H1,D,T,X,O1,B1,H	0,835	4	7%
31	I1,C1	0,835	5	0%
32	G1,R,P	0,844	5	1%
33	J,E	0,852	6	1%

34	G1,R,P,J,E	0,871	5	2%
35	G,U,L,E1,F1,P1,M1,K1,L1,A1,N1,W,K,O,N,V,A,D1,J1,Q,Z,M,F,H1,D,T,X,O1,B1,H	0,875	4	0%
36	Y,G1,R,P,J,E	0,908	4	4%
37	I,G,U,L,E1,F1,P1,M1,K1,L1,A1,N1,W,K,O,N,V,A,D1,J1,Q,Z,M,F,H1,D,T,X,O1,B1,H	0,929	4	2%
38	C1,I1,I,G,U,L,E1,F1,P1,M1,K1,L1,A1,N1,W,K,O,N,V,A,D1,J1,Q,Z,M,F,H1,D,T,X,O1,B1,H	0,984	3	6%
39	S,Y,G1,R,P,J,E	1,002	3	2%
40	C1,I1,I,G,U,L,E1,F1,P1,M1,K1,L1,A1,N1,W,K,O,N,V,A,D1,J1,Q,Z,M,F,H1,D,T,X,O1,B1,H,S,Y,G1,R,P,J,E	1,066	2	6%
41	C,B,C1,I1,I,G,U,L,E1,F1,P1,M1,K1,L1,A1,N1,W,K,O,N,V,A,D1,J1,Q,Z,M,F,H1,D,T,X,O1,B1,H,S,Y,G1,R,P,J,E	1,376	1	29%

Quadro XV –Grupos formados em cada etapa da Análise de Agrupamento

Algumas características podem ser observadas através da análise do conteúdo do Quadro XV, confrontando-o com o Dendrograma exibido na Figura 3.5.

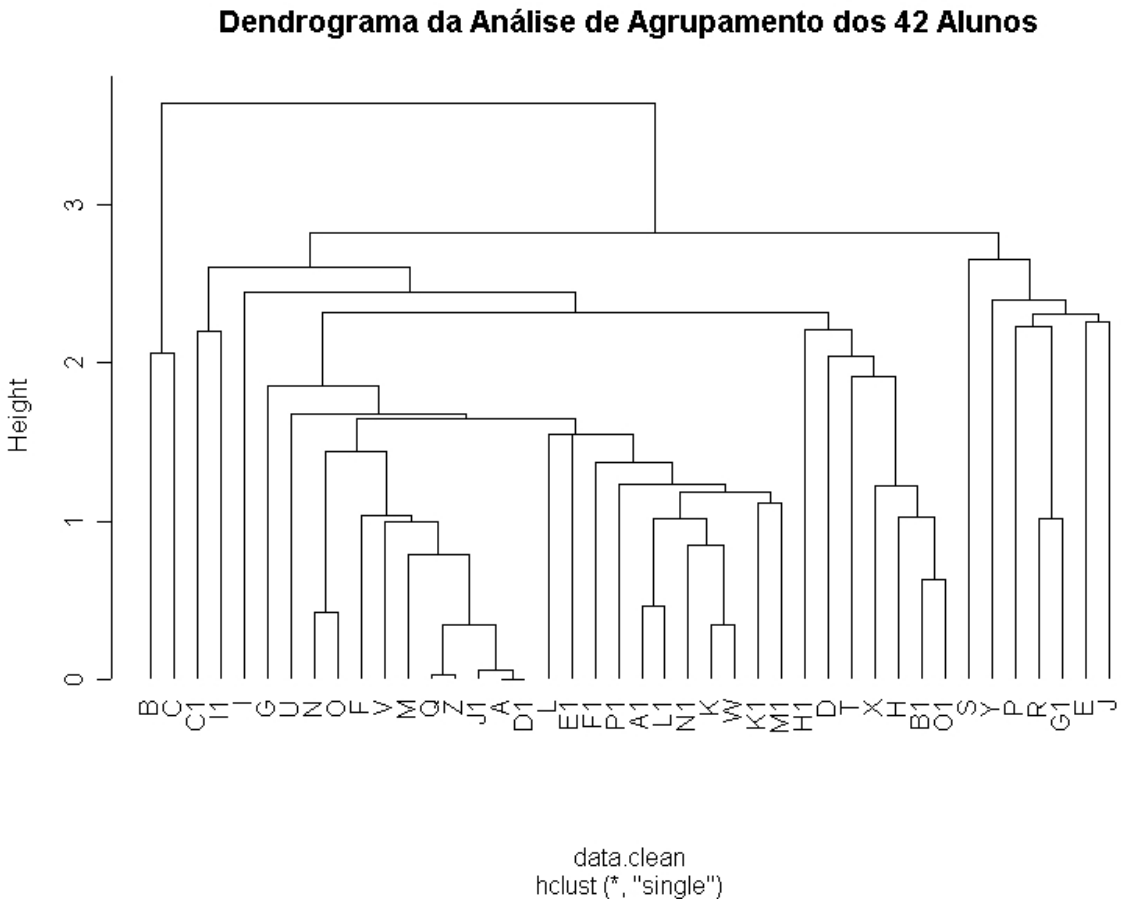


Figura 3.5 – Dendrograma dos Agrupamentos Ocorridos

O terceiro grupo, formado pelos alunos A, D1 e J1, apresentou um coeficiente 0,020 para a sua formação, foi observado que o salto (exatamente a variação percentual do coeficiente anterior - 0,011 - para o atual) para esse coeficiente foi cerca de 82%. Saltos como esses devem ser observados, pois podem indicar que no passo anterior, existiam grupos homogêneos. Isso foi feito exatamente através do acompanhamento da coluna *Dif*, existente no Quadro XV.

3.2.6 Análise dos Resultados

Foi efetuado o cálculo do Coeficiente de Correlação Cofenética, necessário para verificar se o Dendrograma e, por conseqüência, a Análise de Agrupamento efetuada, apresentaram uma boa simplificação da matriz de distâncias. O Coeficiente Cofenético alcançou o valor **0,79** que se encontra bem próximo de 0,80, valor considerado aceitável em Análise de Agrupamento. Consideramos que a Análise de Agrupamento executada nos dados, obteve êxito na formação de grupos homogêneos, os quais foram identificados a seguir.

O Gráfico 3.2, formado com os dados coletados a partir das colunas Passo (valor 5) e Dif (valor 2%) do Quadro XV, representa a relação entre a variação percentual entre dois níveis consecutivos (não foram representados os passos 1 e 2, por serem claramente homogêneos, e os passos 3 e 4 por possuírem valores discrepantes que influiriam na interpretação do gráfico).

Inspecionando esse gráfico, foram observadas algumas variações significativas (e alguns saltos), dividindo o mesmo em regiões, que são caracterizadas pela formação ou não de grupos homogêneos. É correto concluir que, em relação aos passos 6 a 9, 11, 20, 26, e 42, os grupos formados não apresentavam a homogeneidade desejada pois as variações observadas foram significativas. Entretanto, foi necessário definir um patamar para essa

variação percentual, a partir do qual se considera que o grupo formado não é homogêneo. Utilizamos um patamar de 10%, que é um patamar reduzido, com o objetivo de coletar realmente aqueles grupos mais homogêneos.

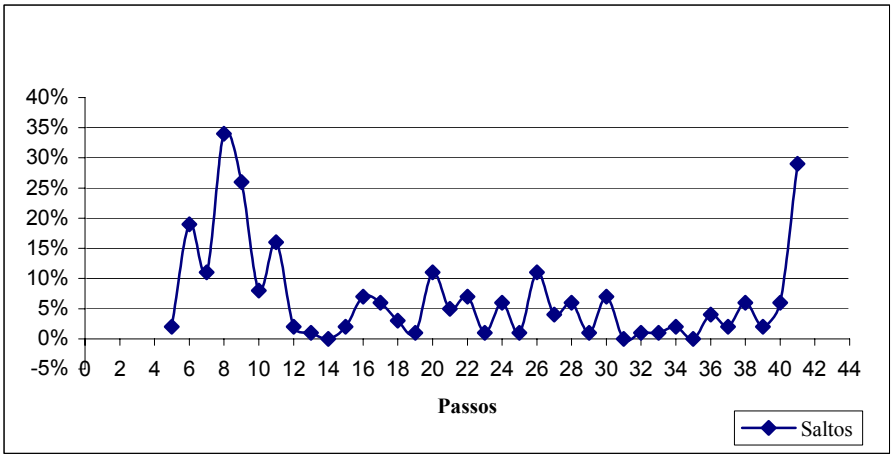


Gráfico 3.2 – Percentuais de Crescimento dos Níveis (Passos x Dif).

A partir desse patamar de 10%, analisamos o Gráfico 3.2 em conjunto com o Dendrograma da Figura 3.5, e os dados do Quadro XV, e conseguimos identificar regiões de homogeneidade, representadas pelos passos 1, 2, 12 a 19, 21 a 25, e a maior dessas, 27 a 40. Todos os grupos pertencentes a essas regiões de homogeneidade, foram fornecidos ao professor da disciplina. A ordem em que se formaram também traduz maior grau de homogeneidade, os mais homogêneos se formam antes dos menos homogêneos. Porém, é interessante identificar os saltos, que indicam a quebra de homogeneidade, que ocorrem após uma grande série de grupos homogêneos. Pelo Gráfico 3.2, esses saltos ocorreram no passo **20** (homogeneidade do passo 12 ao 19), **26** (homogeneidade do 21 ao 25) e 41 (após uma extensa série homogênea, passos 27 a 40), exatamente os limites das regiões de homogeneidade citadas anteriormente.

3.2.7 Conclusões

A identificação de grupos homogêneos, através da aplicação de técnicas de análise de agrupamento, em uma massa de dados rica em informações oriunda de um *Log* gerado por uma plataforma de EAD, com o intuito de oferecer aos professores meios para agrupar seus alunos segundo diferentes critérios e interesses didáticos, mostrou-se inteiramente viável e exeqüível por duas razões: primeiramente, porque mostrou que a técnica selecionada na literatura (BUSSAB *et al*, 1990) e as variáveis geradas pelo *Log* de acesso a uma plataforma de EAD foram suficientemente adequadas e sensíveis para a identificação de diferentes agrupamentos. Em segundo lugar, porque, mesmo utilizando um *Log* de acesso não previamente planejado para esse fim, foi capaz de produzir uma diversidade de grupos homogêneos, o que é uma situação bastante favorável para o que se pretende.

A plataforma Pii apresentou-se ideal para a implementação de uma interface para a geração de grupos homogêneos de alunos, pois além de possuir um arquivo de Log que permite a apuração dos dados relacionados aos alunos e apresentar seu desenvolvimento e manutenção associados a linguagens de programação de última geração, possui uma organização dos cursos adequada à identificação de padrões associados aos participantes que a freqüentam. Assim, como será exposto mais adiante, essa plataforma serviu de laboratório para implementarmos essa interface.

3.3 Proposta do Presente Trabalho

Após a revisão da literatura e os estudos de caso, os quais contribuíram para a consolidação dos conhecimentos apresentados, estaremos nessa seção formalizando a proposta central da dissertação.

3.3.1 Proposta

A Figura 3.6 a seguir ilustra o racional da proposta do presente trabalho que consiste em especificar um sistema de identificação de grupos homogêneos, que atendam alguns cenários pedagógicos previamente especificados de educação à distância ou de ensino presencial apoiado pela WEB, através da aplicação de técnicas de análise de agrupamento em uma massa de dados rica em informações, oriunda de um arquivo Log gerado por uma plataforma EAD.

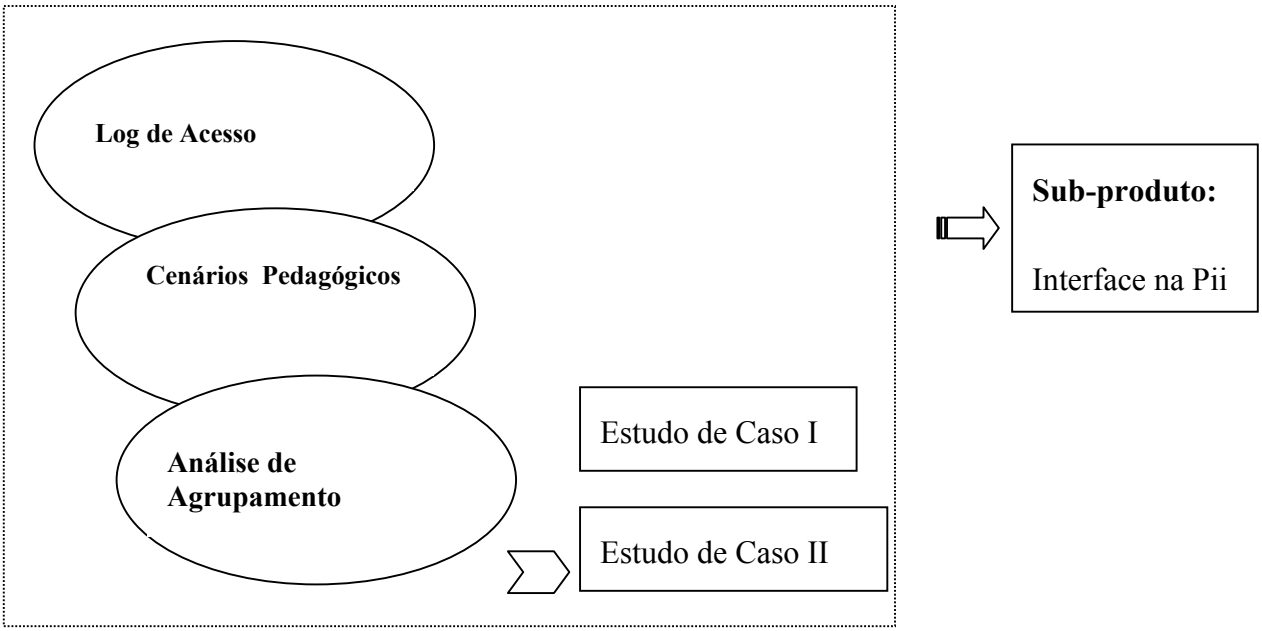


Figura 3.6 – Racional da Proposta para Identificação de Grupos Homogêneos

Na Figura 3.6, o termo Log de Acesso está associado ao arquivo de Log que subsidia com dados as técnicas de Análise de Agrupamento. A estrutura desse Log foi pesquisada na seção 2.3, onde identificamos as características necessárias para que este arquivo seja considerado um Log mínimo e suficiente para a aplicação de técnicas de agrupamento. Essa identificação permitiu concluir que o Log da plataforma Pii, utilizada no Estudo de Caso 2, atende os requisitos para coleta e apuração de dados necessária à aplicação dessas técnicas no contexto do presente trabalho.

A pesquisa efetuada nas teorias educacionais de Vygotsky e Freire, permitiu identificarmos a importância das interações nos ambientes de aprendizagem e inclusive aquelas que surgem a partir da utilização de grupos em tarefas cooperativas e colaborativas, como também a visualização de alguns cenários pedagógicos.

Na busca da reprodução daquelas interações que são possíveis (ou factíveis) de serem criadas nos ambientes virtuais de aprendizagem, utilizamos as técnicas de Análise de Agrupamento, nos moldes do modelo estatístico pesquisado na seção 2.2, nos estudos de caso (seções 3.1.e 3.2) objetivando fornecer aos professores desses estudos, grupos homogêneos de alunos para serem utilizados em tarefas, de onde surgiriam as interações.

Durante a execução desses estudos de caso, nossa proposta começou a se consolidar. O primeiro estudo nos permitiu "sentir" o processo de agrupamento de dados qualitativos através de diversos coeficientes, o que produziu uma sistemática para analisar os resultados gerados pela Análise de Agrupamento. O segundo estudo, que utilizou uma grande massa de dados quantitativos localizada em um arquivo de Log adequado, identificou os requisitos necessários para a implementação das técnicas de Análise de Agrupamento em uma plataforma EAD. Esse último estudo também permitiu que novas hipóteses para aplicação dessas técnicas surgissem, o que contribuiu para a construção de cenários pré-determinados

associados às atividades pedagógicas, identificados na Figura 3.6 como Cenários Pedagógicos.

Esses cenários auxiliam o processo de geração dos grupos, ao buscarem reproduzir as interações identificadas na representação do ambiente Sócio-Interacionista, identificado na revisão educacional (Figura 2.1). Eles devem ser formatados, preferencialmente, pelos professores, os quais possuem os conhecimentos pedagógicos suficientes para tal tarefa. Se além desses conhecimentos pedagógicos, os professores apresentarem experiências consistentes na utilização dos ambientes virtuais de aprendizagem, conhecendo suas ferramentas, características e limitações, a eficiência desses cenários identificados na reprodução de um ambiente, digamos, Sócio-Interacionista, tornar-se-á ideal.

Por fim, como um subproduto do sistema de identificação de grupos homogêneos ora proposto, criamos um protótipo de interface para a Plataforma Interativa para Internet - Pii, que aplica Análise de Agrupamento nos dados do Log dos participantes, fornecendo os grupos homogêneos de acordo com o cenário pedagógico escolhido pelo professor, de um curso suportado por esse ambiente. O cenário pedagógico poderá ser escolhido a partir de um dos 12 pré-definidos ou de um totalmente parametrizável pelo próprio professor. Em adicional à geração dos grupos que são fornecidos, esse protótipo caracteriza cada um deles através de estatísticas apuradas a partir das variáveis, utilizadas nos cenários, dos elementos que os compõem. A implementação dessa interface é o tema do próximo capítulo.

4 IMPLEMENTAÇÃO DA INTERFACE PARA ANÁLISE DE AGRUPAMENTO

4.1 Introdução

Após o estudo de caso apresentado na seção 3.2 ficou caracterizada a viabilidade da implementação de uma nova interface para a Pii, batizada de IAA. Essa nova interface, inicialmente um protótipo, estaria oferecendo ao professor responsável por qualquer curso suportado pela plataforma, a possibilidade da geração de grupos homogêneos de alunos a partir das variáveis associadas ao comportamento dos mesmos quando se encontram nesse ambiente de aprendizagem. Os grupos gerados, além de caracterizados pelos nomes dos elementos que os compõem, apresentariam também estatísticas associadas às variáveis dos componentes que foram utilizadas durante o processo de Análise de Agrupamento. Essas estatísticas teriam o objetivo de facilitar uma futura identificação dos perfis, vistos na seção 2.1 da revisão da literatura, dos grupos e conseqüentemente dos alunos pelo professor.

Entendemos que essa interface deva ser vista e criticada como um protótipo, apesar de apresentar todos os recursos para efetuar as análises de agrupamento pesquisadas para a plataforma Pii, pois se encontra ainda em desenvolvimento.

O funcionamento desse protótipo inicia ao disponibilizar uma interface, onde o professor escolhe qual tipo de cenário será utilizado para o agrupamento; segue ao apresentar resumos explicativos, acessado através da interface, contendo informações suficientes sobre Análise de Agrupamento e assuntos correlatos; e finaliza em uma página HTML onde os grupos produzidos e suas características estariam sendo fornecidos ao professor.

O protótipo construído foi produzido à parte da Plataforma Pii, ou seja, optou-se pela construção de uma programação adicional, associada a uma página HTML, que estaria sendo acessada a partir do próprio ambiente da Pii destinado ao professor do curso hospedado pela plataforma. Com isso foi possível garantir uma modificação mínima do código original da

plataforma, permitindo assim que esse módulo construído pudesse ser adicionado ou não a Pii a partir do código principal – esse fato é interessante pois o administrador da Pii pode determinar quais cursos terão ou não a ferramenta de análise de agrupamento.

Para que a construção da plataforma se tornasse realidade, foi cedido pelo professor Marcos Elia o banco de dados *log1.mdb,* contendo apenas as tabelas, duas no total, necessárias à implementação a ser desenvolvida. Esse arquivo encontra-se no formato do banco de dados *Access* da *Microsoft*, sendo que parte do mesmo já tinha sido utilizada durante a pesquisa feita no estudo de caso descrito na seção 3.2 da presente dissertação. O banco de dados teve suas tabelas analisadas para que se pudesse chegar a métodos suficientes de consulta via linguagem SQL – *Structured Query Language*. Grande parte da programação interna, composta principalmente de manipulação de vetores multidimensionais com os dados recuperados do arquivo de banco de dados citado, foi possível devido: à experiência adquirida quando da execução do estudo de caso já citado; o fato de estarmos há mais de 10 anos atuando no mercado de informática como programador e analista de sistemas; aos conhecimentos adquiridos nas disciplinas Estrutura de Dados e Banco de Dados I que fazem parte do currículo do mestrado em informática da UFRJ.

Importantes inclusões, detalhadas durante esse capítulo, foram feitas na programação para que os objetivos principais da presente dissertação fossem alcançados. Em adicional o processo possibilitou o surgimento de propostas de melhorias na plataforma Pii as quais já estão sendo aplicadas, permitindo assim que a tabela de *Log* existente no banco de dados da plataforma passasse a reter mais informações do que anteriormente, tornando-se mais rica e detalhada, as quais poderão ser incluídas em escopos de Análises de Agrupamento futuras. Com isso o *Log* da plataforma estará evoluindo em busca de um *Log* ideal.

Por fim, a programação da Interface contou com uma utilização heterogênea de linguagens de programação em ambientes distintos. Foram utilizadas principalmente *Visual*

Basic Script - *VBScript* para a manipulação dos dados, *Active Server Pages* – ASP e HTML - *hypertext markup language* na construção das páginas *web*. Consultas SQL também foram utilizadas na programação feita em *VBScript* e em *ASP*. Para suportar a programação nas linguagens escolhidas foram utilizados os ambientes de programação e edição *Visual Basic 6.0*, *Visual InterDev 6.0* e o *Frontpage*, existentes nos laboratórios do NCE/UFRJ.

4.2 Funcionamento da IAA e Visões dos Participantes

Dentro do ambiente da Pii o professor dispõe de menu de navegação formado por uma estrutura do tipo árvore de diretórios que disponibiliza os acessos às regiões que compõem a plataforma educacional, como pode ser visto na Figura 4.1, a seguir:

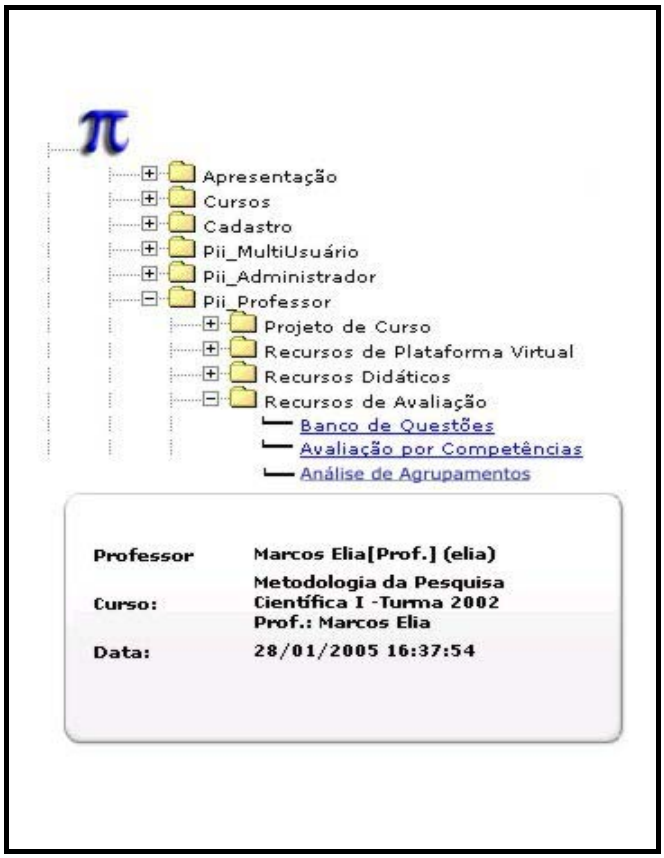


Figura 4.1 – Menu de Navegação da Pii para o Professor

Em *Recursos de Avaliação* está localizado o *link* que leva o professor até a página inicial da Interface para Análise de Agrupamento. Quando o professor clica nesse *link* a plataforma Pii abre na região à direita do menu de navegação a página de entrada da interface construída. Iniciamos a descrição do protótipo a partir desse ponto.

4.2.1 Apresentando a Interface e a Visão do Professor

Um dos objetivos durante a programação da interface foi simplificá-la para que o professor não venha a se sentir confuso, sem saber como proceder ou o quê fazer, ao visualizá-la. O fluxo representado na Figura 4.2 foi utilizado para a construção das páginas visualizadas pelo professor.

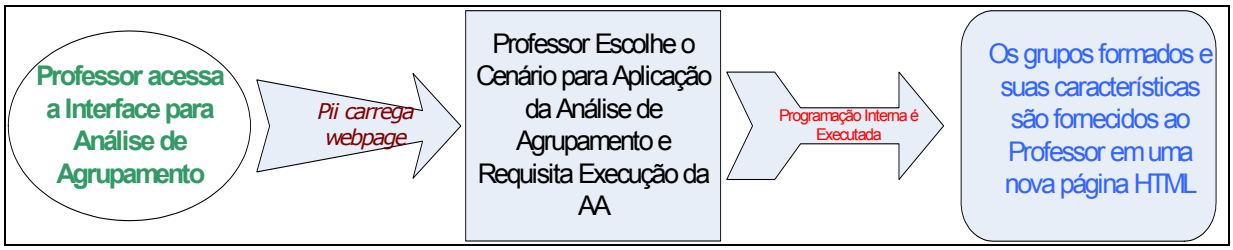


Figura 4.2 – Fluxo do Processo Macro da IAA do Ponto de Vista do Professor

A Figura 4.3 exibe a interface, que é composta de duas regiões: ***Cenários*** e ***Configure seu Cenário***.

Em ***Cenários*** é possível escolher um dos 12 cenários pré-definidos para aplicação de análise de agrupamento, o professor terá que escolher qual tipo de cenário de agrupamento será executado, e conforme o cenário escolhido, terá também que inserir um período (data inicial e final) a ser considerado para os registros que participarão da análise de agrupamento.

Também é possível acessar uma descrição de cada cenário. O cenário A1, por exemplo, foi o utilizado no estudo de caso descrito na seção 3.2. Mais adiante os cenários serão detalhados.

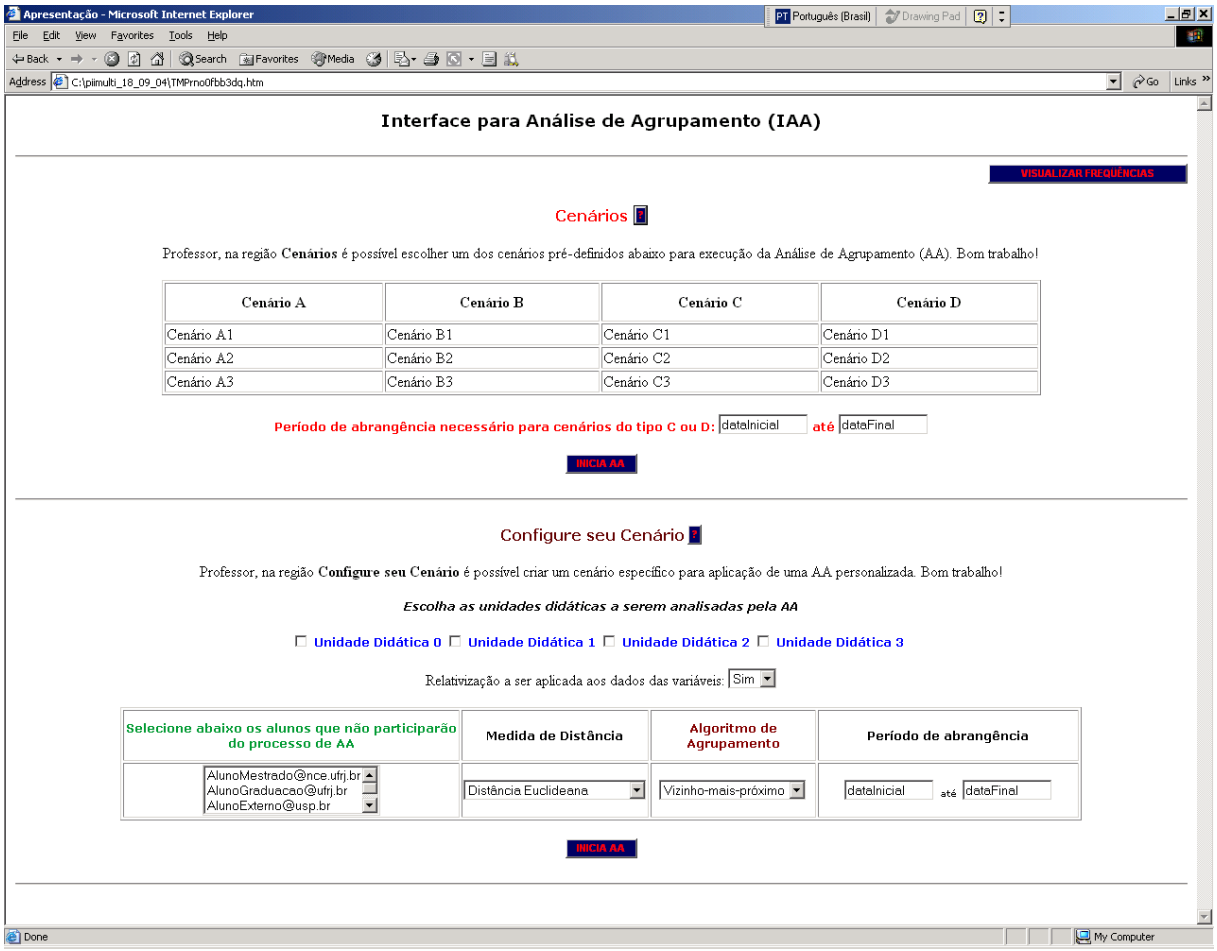


Figura 4.3 – Página inicial da Interface para Análise de Agrupamentos

Na segunda região, ***Configure seu Cenário***, encontram-se as opções para uma análise de agrupamento a ser customizada pelo professor. É possível escolher quais unidades didáticas serão utilizadas; que participantes não farão parte do agrupamento; que período de tempo o agrupamento abrangerá; se os dados coletados dos participantes serão alterados por qual tipo de relativização; qual medida de distância será utilizada na construção da matriz de distâncias; qual algoritmo de agrupamento será utilizado; qual critério utilizado para

determinar ocorrência de quebra de homogeneidade. Também se encontram disponíveis resumos explicativos para cada uma das opções.

4.2.1.1 Descrevendo os Cenários

Cenários Pré-definidos

São 12 os cenários pré-determinados localizados na região ***Cenários***. Eles são descritos da seguinte forma:

- **Cenário A** – Faz a Análise de Agrupamentos com todas as unidades didáticas (UDs) sem período de tempo específico, se desdobra em três tipos:
    - Cenário A1 – Utiliza os alunos e professores que utilizaram os cursos.
    - Cenário A2 – Utiliza apenas alunos.
    - Cenário A3 – Utiliza apenas professores.
- **Cenário B** – Semelhante ao cenário A porém excluindo a variável UD0 – que representa o acesso a qualquer ambiente da plataforma diferente das unidades didáticas – desdobra-se em:
    - Cenário B1 – Utiliza participantes alunos e professores.
    - Cenário B2 – Utiliza apenas alunos.
    - Cenário B3 – Utiliza apenas professores.
- **Cenário C** – É o cenário A com a possibilidade de definição de um intervalo de tempo, que seria o período de abrangência no qual aplica-se a Análise de Agrupamento. Divide-se em:
    - Cenário C1 – Utiliza participantes alunos e professores.

- Cenário C2 – Utiliza apenas alunos.
- Cenário C3 – Utiliza apenas professores.

- **Cenário D** - É o cenário B com a possibilidade de definição de um intervalo de tempo.
  - Cenário D1 – Utiliza participantes alunos e professores.
  - Cenário D2 – Utiliza apenas alunos.
  - Cenário D3 – Utiliza apenas professores.

Todos esses 12 cenários acima apresentam as seguintes configurações referentes a Análise de Agrupamentos que estará sendo aplicada:

Relativização das variáveis, onde foi utilizada a padronização estatística conhecida também por normalização padrão;

Distância Euclidiana Média como medida de distância utilizada na construção da matriz de distâncias;

Algoritmo hierárquico Ligação Simples (vizinho-mais-próximo)*;* acima de 10% foi o critério escolhido para quebra de homogeneidade – esse valor representa o crescimento percentual da distância que determina a atual associação de dois elementos (grupos ou participantes) em um novo grupo em relação a distância utilizada no agrupamento anterior.

Cenário Configurável

Na região ***Configure seu Cenário*** a interface permite ao professor definir a configuração de um ambiente para aplicação de uma Análise de Agrupamento personalizada. Será possível escolher quais variáveis participarão da Análise de Agrupamento, quais

participantes não farão parte da análise, qual o período (data de início e/ou data de fim) de abrangência da análise a ser executada e quais serão os parâmetros definidos referentes a aplicação da análise desejada, os parâmetros são: que relativização será executada nos dados; que tipo de distância será utilizada e se ela será de similaridade ou dissimilaridade; qual algoritmo será aplicado para a formação dos grupos.

Resumos explicativos acessíveis através de *links* foram disponibilizados nessa região para elucidar as dúvidas que o professor venha a apresentar durante o acesso à mesma.

4.2.2 Visões dos Participantes

Uma ferramenta adicional que foi inserida nessa interface pode ser acessada através do botão *Visualizar Freqüências*. Esse botão abre uma página HTML contendo as freqüências nas unidades didáticas de cada participante associado ao atual curso. Essa consulta estruturada e isolada é uma facilidade aproveitada da programação que tem o objetivo de permitir ao professor um maior conhecimento do acesso feito às unidades didáticas, o que pode vir a ser útil quando o professor venha a decidir que tipo de cenário utilizará para efetuar o agrupamento. *Visualizar Freqüências* também se encontra disponível para o administrador do curso, se este porventura estiver interessado em verificar a assiduidade dos participantes nas unidades didáticas. O administrador encontrará esse *link* na seção *Estatísticas* pertencente à seção *Pii_Administrador*, localizada na árvore de diretórios da Pii. A seguir, na Figura 4.4, é exibido um exemplo da página gerada por *Visualizar Freqüências* – os e-mails dos participantes foram alterados:

Participantes	UD0	UD1	UD2	UD3	UD4	UD5	UD6	UD7	UD8	UD9	UD10
a@ufrj.br	2	0	0	0	0	0	0	0	1	0	0
b@ufrj.br	44	2	1	0	2	1	3	3	3	2	2
c@ufrj.br	17	1	0	0	0	3	3	1	1	2	3
d@ufrj.br	6	0	0	0	0	0	1	2	1	1	1
e@ufrj.br	14	3	1	2	2	1	2	1	2	1	4
f@ufrj.br	73	8	10	4	9	10	12	7	9	1	3
g@ufrj.br	8	1	0	1	0	0	0	3	1	0	3
h@ufrj.br	36	0	2	4	6	2	6	6	3	0	1
i@ufrj.br	24	5	1	5	1	2	0	1	1	4	2
j@ufrj.br	12	0	0	0	0	0	1	1	3	1	1
k@ufrj.br	59	6	6	6	6	2	2	4	2	7	1
l@ufrj.br	11	2	1	1	4	1	1	1	1	2	1
m@ufrj.br	27	11	7	7	8	1	0	0	1	0	1
Total	333	39	29	30	38	23	31	30	29	21	23

Figura 4.4 – Exemplo das Freqüências totais por participantes

O participante Aluno, por definição, não possui acesso a *Visualizar Freqüências*, porém ele poderá verificar nos grupos publicados pelo professor, em que grupo ele está situado para cumprir determinada tarefa do curso em andamento. Essa funcionalidade é acessada através da subseção *Verificar Grupos Publicados*, localizada na seção *Grupos de Estudo*, pertencente a árvore de diretórios da Pii, representada na Figura 4.1.

4.3 Programação Interna da IAA e Diagramas UML

A programação interna executa três processos principais a partir das definições feitas na interface apresentada na seção anterior: 1) Recuperação dos dados existentes no arquivo *log1.mdb;* 2) transformação desses dados gerando novos dados; 3) Exibição desses novos dados através da página final da interface contendo os grupos formados pela Análise de Agrupamento. A Figura 4.5 a seguir, exibe o Diagrama de Contexto que representa o funcionamento macro da IAA.

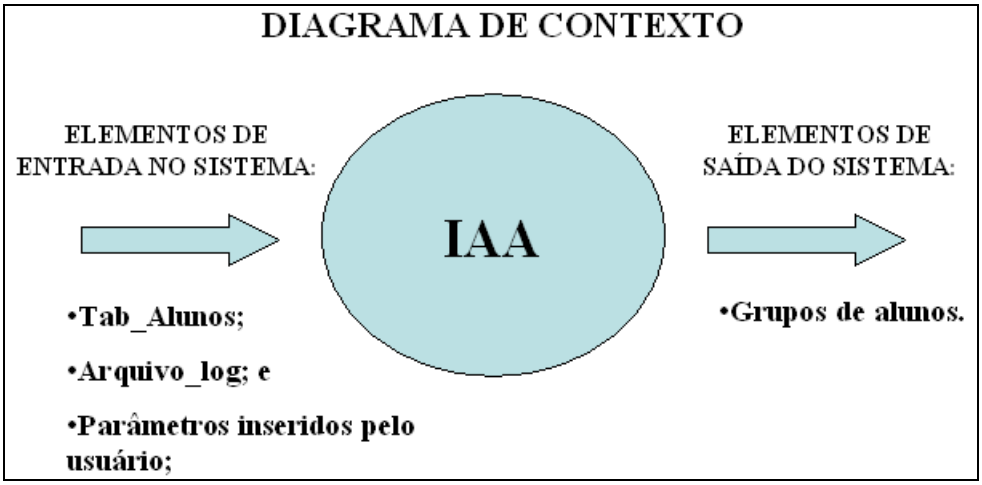


Figura 4.5 – Diagrama de Contexto da IAA

Essa programação interna é composta por mais de mil linhas de código, e seu funcionamento passou por várias etapas de homologação, as quais permitiram um acompanhamento detalhado e apurado do que o código em questão estava executando, esse acompanhamento proporcionou diversos melhoramentos os quais contribuíram para que o código apresentasse uma maior eficiência e eficácia. Podemos assegurar, baseados nessas homologações[41], que a programação executa rigorosamente o que se espera dela. Antes de comentarmos as principais rotinas dessa programação, apresentamos a seguir amostras das tabelas oriundas do arquivo de banco de dados citado anteriormente, contendo apenas os campos que, efetivamente, fizeram parte do escopo da programação. Também são apresentadas, da Tabela 4.1 a 4.4, as estruturas das tabelas do arquivo de banco de dados, fornecidas pelo programa *Access*.

[41] As homologações utilizaram os aplicativos Excel e Access para executar os mesmos procedimentos existentes na programação construída.

Aluno	Udidatica	Data	Hora	Log	Curso
A	0	31-mai-04	00-01-1900 19:24:00	1	32
B	0	23-mai-04	00-01-1900 17:38:16	1	32
C	2	10-mai-04	00-01-1900 19:01:22	6	32
D	3	17-mai-04	00-01-1900 22:54:17	6	32

Tabela 4.1: Amostra da tabela Log_pii existente no arquivo log1.mdb

Nome do Campo	Tipo de Dados
Aluno	Texto
Udidatica	Numero
Data	Data/Hora
Hora	Data/Hora
Log	Texto
Curso	Numero

Tabela 4.2: Estrutura da tabela **Log_pii**

Curso	Tipousuario	E_Mail
3	4	arthurantunes@aol.com
3	4	edsonarantes@ccead.puc-rio.br
3	4	diegoarmando@terra.com.br
3	4	zinedinezidane@rjnet.com.br

Tabela 4.3: Amostra da tabela **tabalunos** existente no arquivo log1.mdb

Nome do Campo	Tipo de Dados
Curso	Numero
Tipousuario	Texto
E_Mail	Texto

Tabela 4.4: Estrutura da tabela **tabalunos**

Os diagramas de classes a seguir, figuras 4.6 a 4.8, representam as relações das tabelas que serviram de estrutura para o modelo utilizado na IAA.

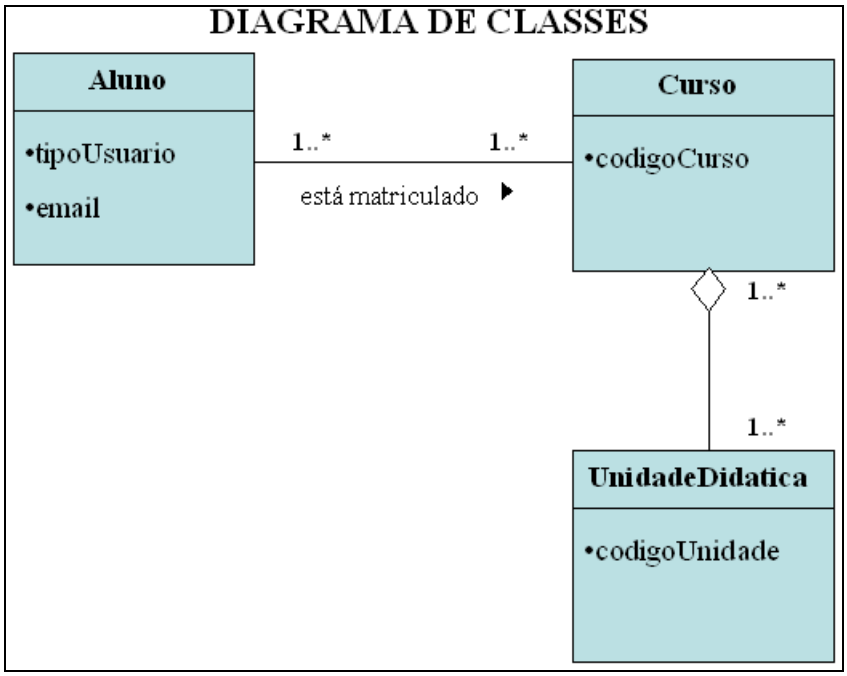


Figura 4.6 – Primeira parte do Diagrama de Classes

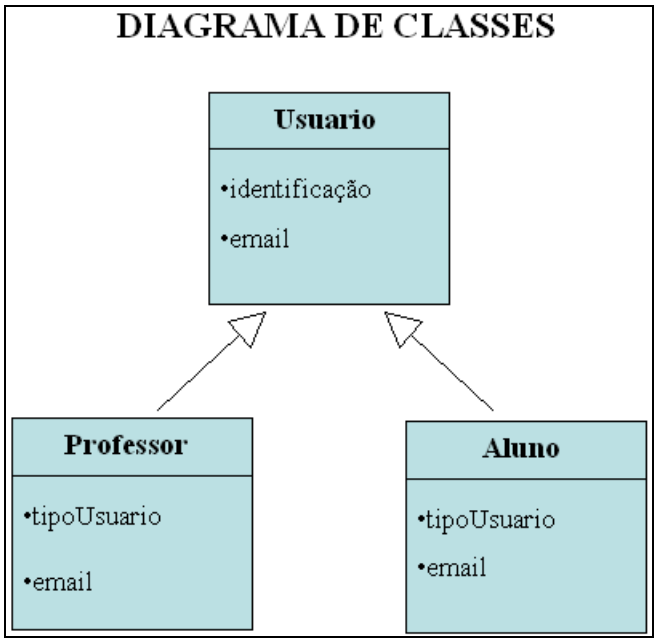


Figura 4.7 – Segunda parte do Diagrama de Classes

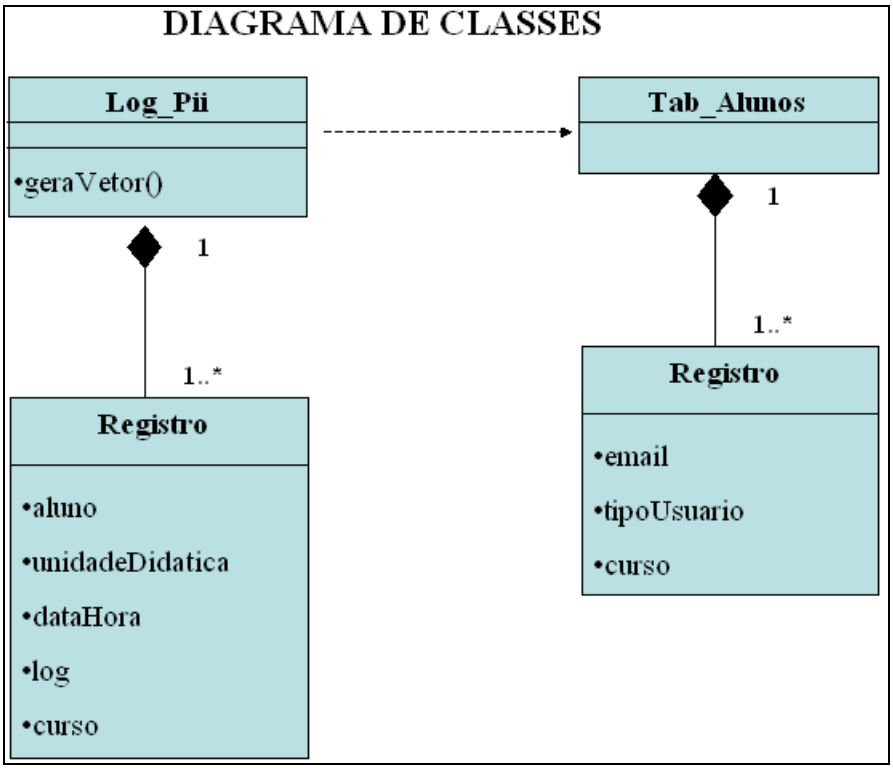


Figura 4.8 – Terceira Parte do Diagrama de Classes

4.3.1 Rotinas do Escopo Principal

Quando o professor clica no botão *Aplicar AA* na página onde se encontram os cenários, o escopo principal da programação que foi feita é acionado. Esse escopo é composto das seguintes rotinas e sub-rotinas, localizadas no Quadro XVI, associadas aos tópicos da revisão da literatura apresentados no capítulo 2:

Rotinas	Tópicos da Revisão
Call GeraVetor(NumeroCurso, TipoCenario) *Call HomologacaoUDs(VetTemp, IndiceVetTemp - 1)* *Call Homologacao(IndiceAlunoApurado - 1, QuantidadesUD + 1)*	**2.3 Análise sobre LOG**
Call CalculaMedias(QuantidadesUD + 1) *Call CalculaDP(QuantidadesUD + 1)* *Call Normaliza(QuantidadesUD + 1)*	**2.2.3 Agrupamentos Simples e Relativização de Variáveis**
Call CriaMatrizDistancias(IndiceAlunoApurado)	**2.2.4 Definindo Alguns dos Principais Componentes da Análise de Agrupamento** **2.2.7.1 Coeficientes Utilizados para Variáveis Quantitativas**
Call QUICKSORT *ExecutaQuickSort VetDistancias(), 0, (Distancias - 1)*	
Call ConstruindoGrupos(Distancias)	**2.2.8.3 Técnicas de Partição para Análise de Agrupamento**
Call GerandoPerfis(IndiceGrupo) *RecuperaPosicaoAluno (Elemento)*	**3 METODOLOGIA**

Quadro XVI – Rotinas do Escopo Principal e Tópicos Correspondentes da Revisão da Literatura

Antes de resumirmos os principais eventos das rotinas e sub-rotinas anteriores, apresentamos o Diagrama de Atividades, na figura a seguir, composto pelas principais rotinas.

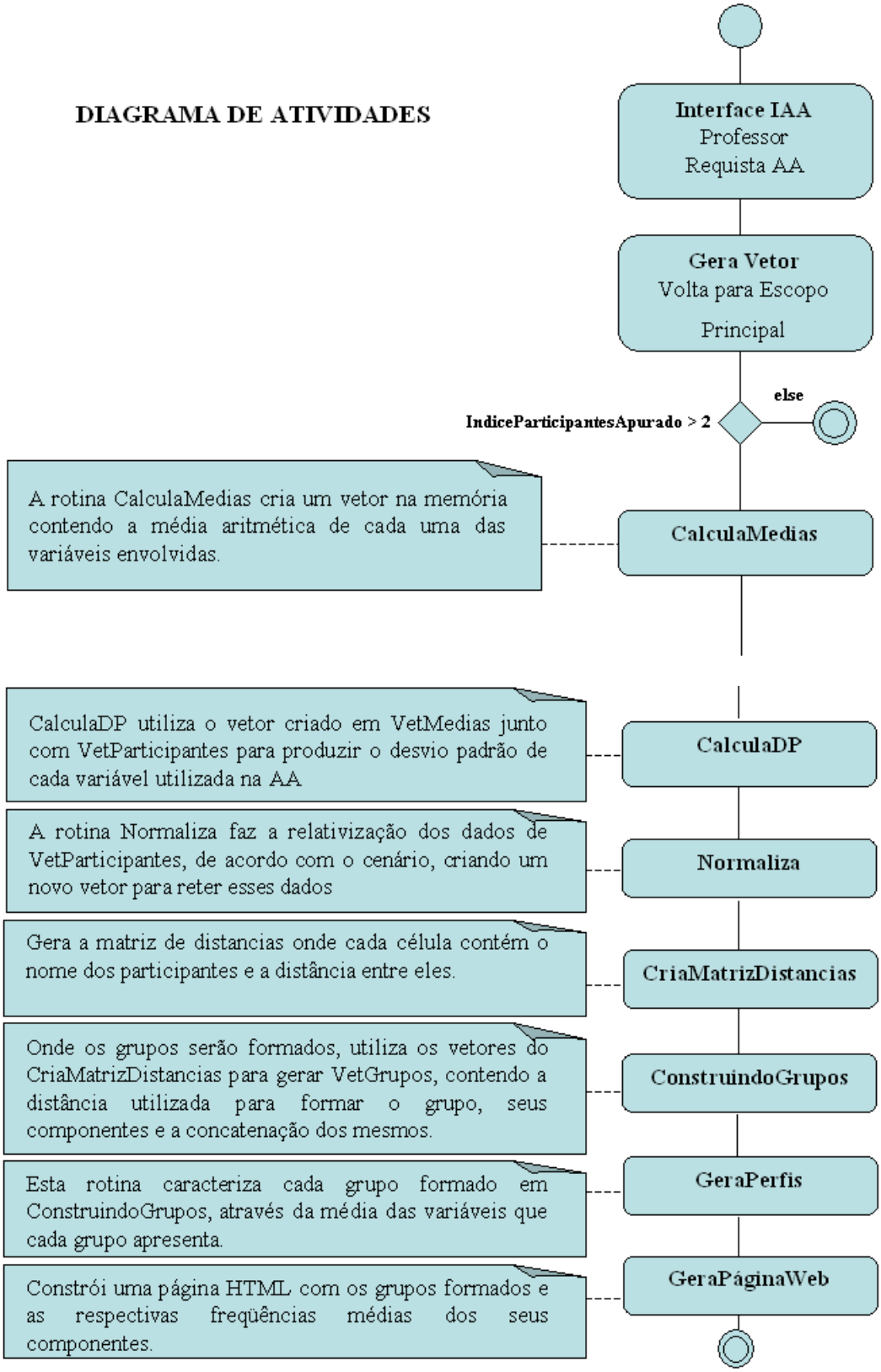


Figura 4.9 – Diagrama de Atividades da IAA

4.3.1.1 GeraVetor(NumeroCurso, TipoCenario)

O principal objetivo dessa rotina é recuperar os dados existentes na Tabela 4.1, verificar a consistência desses dados e caso sejam consistentes, transformá-los numa estrutura semelhante a existente na Figura 4.4. Essa nova estrutura é composta pelos dados contidos em um vetor que através da sub-rotina *Homologacao*, é exportado para o arquivo texto *dados.txt.* Uma amostra desse arquivo, relativa a um curso hospedado pela plataforma Pii, pode ser observada na Figura 4.6, a seguir:

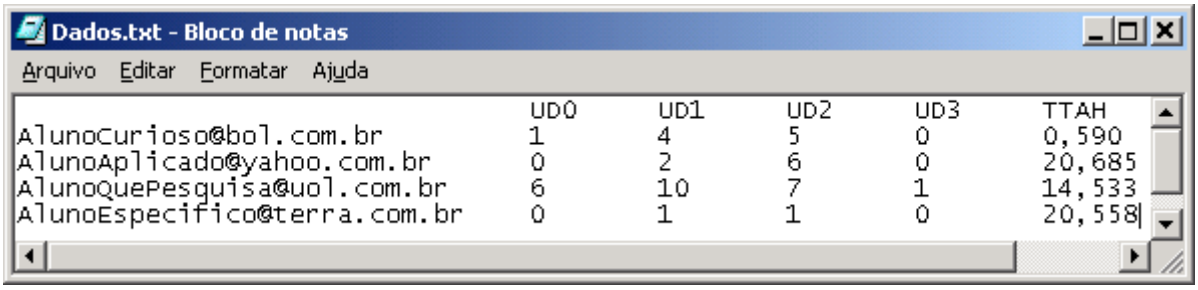


Figura 4.10 – Arquivo contendo a tabela de freqüências por Aluno nas UDs

Na figura anterior, o arquivo *dados.txt* é composto por uma matriz de quatro linhas por seis colunas. Cada linha representa um aluno que teve seus dados utilizados na Análise de Agrupamento. A primeira coluna é o nome do aluno, a segunda representa o total de visitas que esse aluno fez na UD0 (já detalhada em Cenário B), da terceira a quinta colunas temos os totais referentes às visitas que cada aluno fez a cada unidade didática (esse curso teve 3 unidades didáticas), e a última coluna é o tempo total em horas que o aluno permaneceu na plataforma Pii.

Portanto, os dados lidos da Tabela 4.1 tiveram que ser transformados nesse tipo de estrutura para que fosse possível a aplicação da análise de agrupamento. Uma outra solução seria que a plataforma gerasse uma nova tabela com uma estrutura semelhante disponibilizando os dados para a leitura e execução da Análise de Agrupamento, porém fazendo desse modo estaríamos sobrecarregando a plataforma, pois a mesma teria que gravar

dados na tabela *Log_pii*, a qual preenche outras funções da plataforma, e gravar novamente os mesmos dados, porém de uma outra forma, em uma nova tabela.

Optamos então em construir uma rotina que recuperasse diretamente os dados na tabela *Log_pii*, utilizando para isso consultas SQL bem estruturadas.

A rotina inicia recuperando os dados referentes ao curso, conhecido através da manipulação de um *cookie* (uma espécie de variável representada pelo conteúdo de um determinado arquivo armazenado no servidor). Essa recuperação inicial serve para conhecer os participantes existentes nesses cursos e quantas UDs existem nesse curso, esses totais estarão sendo utilizados na definição dos vetores utilizados durante a execução da rotina, otimizando o código já que está ocorrendo alocação de memória. O núcleo da rotina é composto por dois *loopings*, um dentro do outro. Também de acordo com o cenário, participantes de determinado tipo poderão ser excluídos ou não dessa contabilização, desempenhando um importante papel nessa tarefa a tabela *tabalunos.* No protótipo construído estamos atendendo somente os cenários cujos participantes sejam alunos e professores ou somente alunos.[42]

A rotina considera como dados válidos para serem guardados no vetor principal aqueles onde foi possível identificar os momentos nos quais o participante entrou na plataforma, gerando um evento no *log,* e saiu da mesma, gerando um outro evento, mesmo que essa saída tenha ocorrido no dia seguinte (o critério adotado nessa específica situação foi que o primeiro evento desse participante fosse uma saída da plataforma). Quaisquer outras situações foram descartadas nas apurações efetuadas.

---

[42] No estudo de caso da seção 3.2 o registro referente ao professor da disciplina foi excluído dos dados que participaram da AA através da sua deleção simples e pura, pois conhecíamos o e-mail do mesmo.

As variáveis relativas às unidades didáticas exibem o total de vezes que um determinado participante visitou o ambiente representado pela unidade didática. Foi contabilizada também, como no estudo de caso da seção 3.2, o tempo total de acesso desse participante na plataforma Pii durante todas as visitas que efetuou.

Ao finalizar com o último participante, a variável *IndiceAlunoApurado* retém o total de participantes que tiveram suas informações coletadas, e o vetor *VetParticipantes* contendo todos os dados totais das variáveis por participante válido. O arquivo *dados.txt* gerado pela sub-rotina *Homologacao*, serviu para efetuarmos grande parte das homologações – testes de consistência dos dados, das rotinas seguintes existentes no protótipo. Porém a verificação da consistência dos dados existentes nesse arquivo teve que ser feita manualmente, através da impressão de uma amostra de um participante e contabilização a partir dos dados identificados nessa impressão, dos totais referentes às variáveis UDs e TTAH. Esse arquivo, o *dados.txt*, permitiu a inclusão na seção **Estatísticas** do tópico **Pii_administrador** localizado na árvore de diretórios da Pii da Figura 4.1, uma rotina modificada e oriunda de *GeraVetor* que apresenta o conteúdo visto em *dados.txt* adicionado dos totais por variáveis e dos títulos das colunas.

O arquivo *dados.txt* pode ainda ser utilizado para análises em pacotes estatísticos, como por exemplo Análise Discriminante, de Componentes ou mesmo uma Análise Fatorial nas variáveis, o que seria importante, como frisou Bussab em nossa pesquisa, ao exemplificar que dados pertencentes a muitas variáveis poderiam ser analisados com o intuito de redução de mesmas, descartando aquelas que contribuem muito pouco com as observações. O *Splus* apresenta uma ferramenta que facilita a leitura desses tipos de dados no formato texto gerado, ao oferecer uma ponte direta de submissão e recuperação de informações.

Após o término da rotina *GeraVetor*, o escopo principal continua a execução das outras rotinas se os dados coletados forem suficientes para a execução da análise de

agrupamento, isso somente ocorrerá se tivermos dados coletados associados a mais de dois participantes (esse foi o critério utilizado pois na nossa pesquisa não encontramos nenhuma referência ao número mínimo de elementos, tampouco de variáveis associadas aos mesmos, necessários à aplicação de uma análise de agrupamento). Essa verificação é feita pelo comando *If IndiceAlunoApurado > 2 Then*, representado na Figura 4.5 pelo *shape* condicional *IndiceAlunoApurado > 2*. Se tivermos mais que dois participantes, as rotinas a seguir são executadas, caso contrário, o programa termina sua execução enviando uma mensagem para o professor explicando o que ocorreu.

4.3.1.2 CalculaMedias, CalculaDP e Normaliza e Segunda Parte do Diagrama

A rotina *CaculaMedias* constrói o vetor *VetMedias* contendo as médias de todas as variáveis numéricas e exporta esses dados para o arquivo *medias.txt*, que foi utilizado na homologação da programação.

A rotina *CalculaDP* cria o vetor *VetDP* preenchendo cada posição dele com o Desvio Padrão[43] de cada variável numérica, utilizando para isso a média de cada uma, existente em *VetMedias*. Essa rotina cria uma saída para esses dados no arquivo texto um arquivo texto, *DPs.txt*, que foi utilizado durante a homologação.

A rotina chamada a seguir, *normaliza*, executa a relativização das variáveis apuradas em *GeraVetor*, atualizando o vetor *VetParticipantes.* Essa rotina pode executar, de acordo como cenário escolhido, um dos seguintes três tipos de relativização: Padronização Estatística – Normal (0,1); Padronização pela média (divide cada observação das variáveis pela média) e

a padronização que utiliza os máximos e mínimos das variáveis. Para fins de homologação, essa rotina exporta esses novos dados para o arquivo *normalizadas.txt*.

Antes de descrevermos as rotinas restantes do escopo principal, se faz necessário apresentar a segunda parte do diagrama, na Figura 4.7, que resume o funcionamento dessas rotinas:

4.3.1.3 CriaMatrizDistancias e QUICKSORT

A rotina CriaMatrizDistancias contém a programação necessária para a criação da matriz de distâncias. Essa rotina gera a matriz de distâncias utilizando a distância associada ao cenário escolhido. Caso seja um dos 12 cenários pré-definidos, essa distância é a Distância Euclidiana Média, a mesma utilizada no estudo de caso da seção 3.2, caso contrário, utiliza a distância escolhida pelo professor.

A matriz de distâncias construída é formada por um vetor, *VetDistancias*, de três dimensões, as duas primeiras correspondendo aos dois participantes de onde será calculada a distância a partir de seus dados relativizados, que calculada será a terceira dimensão. Ressaltamos que o tamanho desse vetor foi otimizado pois essa pesquisa concluiu que esse tamanho não ultrapassa $n(n-1)/2$. Essa rotina também exporta os dados calculados, para o arquivo *Distancias.txt*, para ser utilizado na homologação.

Construída a Matriz de Distâncias é necessário ordená-la, pois estando o vetor referente a ela ordenado, a geração dos grupos será mais rápida pois as menores distâncias estarão localizadas no princípio do vetor (se a distância associada ao cenário escolhida for de

---

[43] O desvio padrão calculado refere-se a $\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$

dissimilaridade) e as maiores estarão no fim do vetor (distância de similaridade). Para ordená-la foi adicionada à programação a conhecida rotina de ordenação *Quicksort*[44], considerada um das mais rápidas e eficientes rotinas para ordenação de dados que se encontram desordenados. Essa rotina irá indexar o vetor *VetDistancias* do menor para o maior valor. Essa rotina aciona a função *ExecutaQuicksort* onde é feita realmente a ordenação e a re-arrumação do vetor *VetDistancias* utilizando para isso bastante recursividade. A rotina *Quicksort* ainda gera o arquivo *DistOrdenadas.txt*, que foi bastante útil na comparação com o arquivo *Distancias.txt* durante a fase da homologação da programação.

4.3.1.4 ConstruindoGrupos

É nessa rotina que os grupos são formados de acordo como cenário escolhido inicialmente na interface. Ela preenche o vetor *VetGrupos* à medida que os grupos vão se formando. Cada elemento desse vetor corresponde a um grupo formado, sendo que dependendo do cenário escolhido, é comum um elemento conter outros elementos desse vetor.

A rotina inicia seu funcionamento ao definir o tamanho e o número de dimensões de *VetGrupos*, que possuirá um tamanho igual ao número de distâncias calculadas em *CriaMatrizDistancias* subtraída de uma unidade, e um total de (Nº de unidades didáticas + 6) dimensões. Em *ConstruindoGrupos* somente serão preenchidas as quatro primeiras dimensões, referentes à distância (ou nível) na qual o grupo foi formado, primeiro elemento desse grupo, segundo elemento desse grupo e a concatenação desses dois elementos separados

---

[44] O Quicksort é um algoritmo de ordenação freqüentemente utilizado devido a sua rapidez e eficiência na ordenação de dados que não estejam ordenados. É considerado um dos mais rápidos que existe, sua rapidez se justifica pois além de possuir uma complexidade O(*n log n*) para o caso médio e O(n^2) em seu pior caso (cuja ocorrência é rara), funciona baseado na estratégia da divisão e conquista. Ele particiona uma lista de *n* elementos em duas listas menores que serão ordenadas através da divisão em outras duas listas menores, através de chamadas recursivas ao próprio *Quicksort*. O algoritmo foi originalmente publicado por C.A.R. Hoare em 1962 (FLENSBURG , 2005).

por ponto-e-vírgula. Tanto o primeiro quanto o segundo elementos podem ser grupos. As outras dimensões serão preenchidas na rotina *GerandoPerfis*.

A rotina utiliza vetores auxiliares para executar o procedimento necessário a cada criação de um grupo, que significa percorrer o vetor *VetDistanciasTemp* (cópia de *VetDistancias*) para saber em que posição terá que substituir o elemento que acabou de se juntar a um outro na formação de um novo grupo, recalculando a distância entre esses novos elementos de acordo com o cenário escolhido na interface.

A rotina termina sua execução gerando o arquivo texto *GruposFinais.txt* com o conteúdo do vetor *VetGrupos*, colocando em cada linha a distância (ou o nível) utilizada na formação do grupo e seus componentes. A seguir, a Figura 4.8 que representa essa amostra:

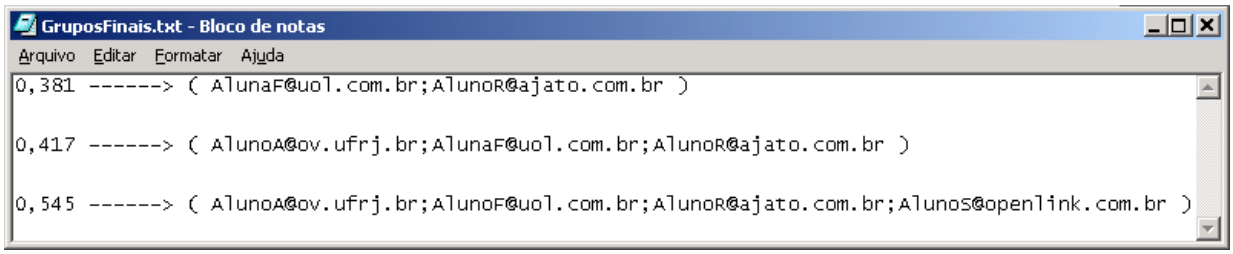


Figura 4.11 – Amostra de *GruposFinais.txt* (os e-mails foram alterados).

4.3.1.5 GerandoPerfis

O principal objetivo dessa rotina é caracterizar um pouco mais os grupos formados em *ConstruindoGrupos*, para que se criem condições para a aplicação do que foi visto quando revisamos na seção 2.1 o educador Paulo Freire e suas teorias referentes à necessidade de se conhecer mais o educando e o meio em que ele vive.

Essa caracterização ocorre através do preenchimento das dimensões vazias de cada elemento de *VetGrupos*. Essas dimensões não preenchidas anteriormente referem-se às variáveis que medem a freqüência de visitas feitas às unidades didáticas e ao tempo total

despendido na plataforma porém, diferente das variáveis de mesmo objetivo existentes em *VetParticipante* associadas apenas a um determinado participante, em VetGrupos elas representam a freqüência total gerada pela soma da freqüência individual dos elementos dos elementos que compõem o grupo. Ou seja, supondo um grupo composto somente pelos participantes A e B, sendo sete o total de visitas que o participante A fez a UD5, e 13 o total de visitas que B fez a mesma UD5, o total de visitas do grupo, formado pelos dois, à UD5 será 20.

A rotina implementa essa totalização utilizando a quarta dimensão de *VetGrupos* que apresenta os elementos concatenados e separados por ponto-e-vírgula, auxiliada pela função *RecuperaPosicaoAluno*, que busca em *VetParticipantes* a posição de cada um dos elementos que formam o grupo.

Após efetuar todas as totalizações necessárias em *VetGrupos*, a rotina exporta para o arquivo texto *GruposFinaisComPerfis.txt* os grupos formados, a distância que formou cada um deles e as médias das freqüências de cada um nas variáveis que participaram do agrupamento. A utilização da média busca representar cada grupo como se fosse um único participante, porém outras estatísticas poderiam ter sido utilizadas como mediana, moda, etc. A figura a seguir, exibe uma amostra desses grupos já com os perfis gerados:

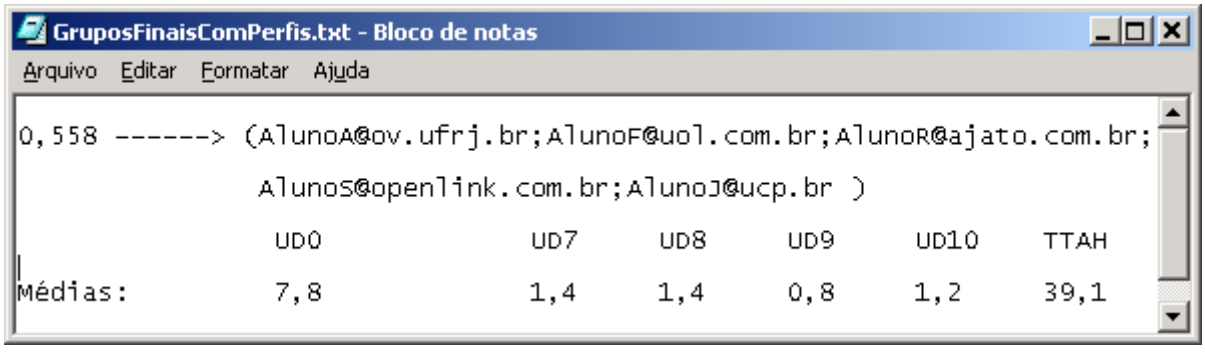


Figura 4.12 – Amostra de *GruposFinaisComPerfis.txt* – os e-mails foram alterados

Finalizando a apresentação das rotinas e sub-rotinas, as Figura 4.13 e 4.14 compõem o Diagrama de Componentes, onde a organização dos serviços gerados pela IAA são apresentados.

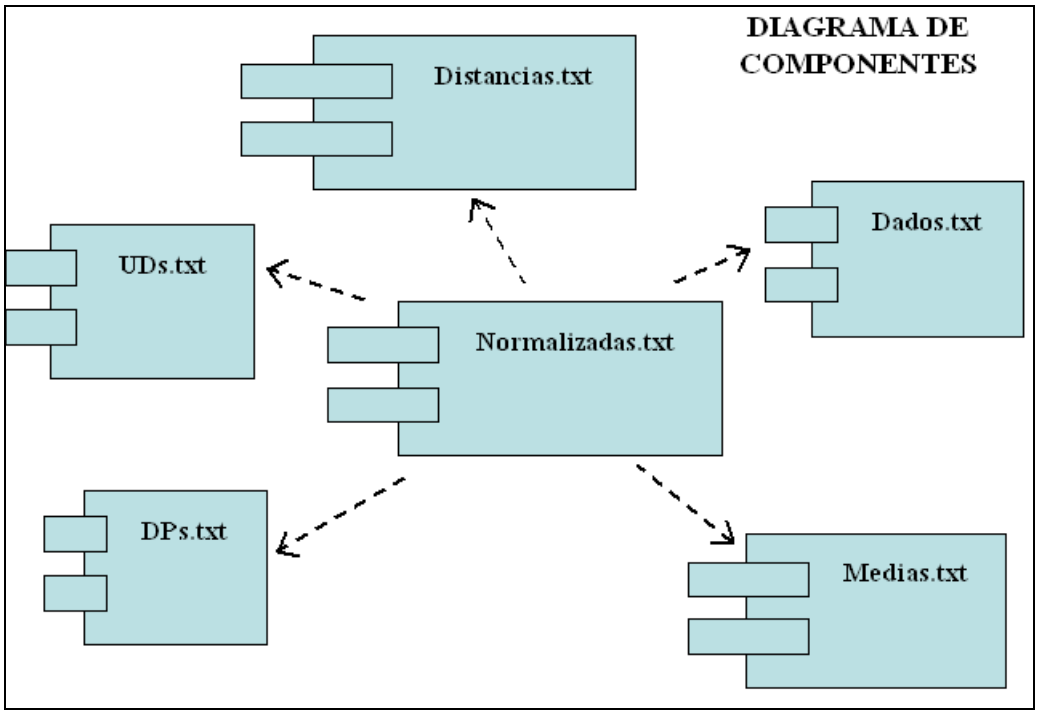


Figura 4.13 – Diagrama de Componentes (primeira parte)

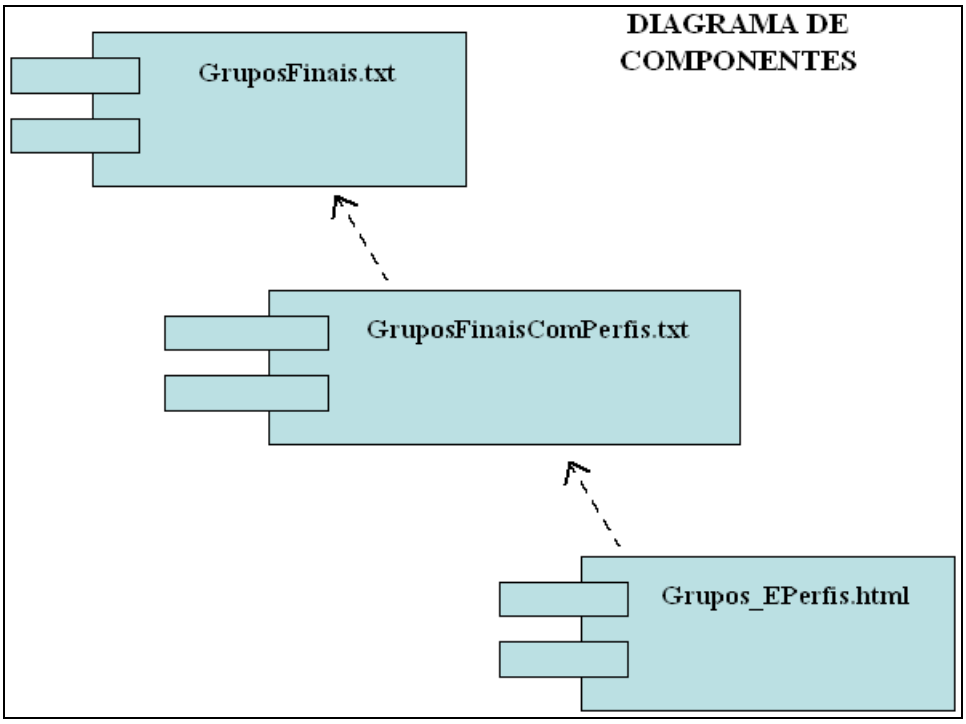


Figura 4.14 – Diagrama de Componentes (segunda parte)

4.4 Novas Implementações Em Andamento no Log e na Pii

Durante a programação da interface, foram detectadas e analisadas algumas novas possibilidades de implementação, cujo esforço computacional apresentou-se devido ao fato de grande parte das rotinas construídas poderem ser reutilizadas. Dentre essas possibilidades, duas estão sendo implementadas:

Análise de Agrupamento somente para os professores dos cursos hospedados pela Plataforma Pii.

Adição ao *Log* da Pii e posterior inclusão na interface para análise de agrupamento, do rastreamento dos cliques que são gerados na árvore de diretórios da plataforma, como pode ser revista na Figura 4.1.

A primeira implementação, refere-se à aplicação de uma análise de agrupamentos para todos os professores, classificados como tal na tabela *tabalunos*, independente do curso. Ou seja, a plataforma estaria fornecendo uma forma de verificar quais professores apresentam comportamentos semelhantes durante suas interações com o ambiente da Pii. Essa implementação encontra-se na fase final de programação e homologação.

A segunda implementação está envolvendo alterações na programação da Pii necessárias à inserção de novas informações no *Log* da plataforma. Essas alterações possibilitam que se registre a *string* que identifica o clique em cada ramo e galho da árvore de diretórios da Pii, o que permitiria um desdobramento da variável UD0 em várias outras. Estaríamos abrindo o escopo da análise de agrupamento, criando mais cenários pré-definidos e permitindo ao professor uma análise mais profunda do comportamento do participante do seu curso. Tal implementação está em andamento.

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 Conclusões

Durante o mestrado, participamos da formação de grupos em algumas disciplinas. Na maioria dessas experiências, os critérios utilizados na construção dos grupos estavam associados às características dos alunos, principalmente sobre a afinidade entre os colegas, a qual contribuía quase totalmente para essas construções. Em algumas dessas experiências, a disciplina contava com algum tipo de ambiente EAD apoiando a aula, porém, as informações dos alunos produzidas nesse ambiente não foram determinantes para a formação dos grupos. Isso ocorreu porque as interações inerentes a um ambiente da sala de aula são, na maioria das vezes, mais que suficientes para permitir a construção de agrupamentos de alunos, pois são ricas em informações e geração de conhecimentos tácitos para os participantes. Seria então um contra-senso, até mesmo uma perda de tempo, a utilização de Análise de Agrupamento em substituição a esse *modus operandis*. Entretanto, no contexto dos ambientes virtuais de aprendizagem, a utilização de Análise de Agrupamento para formar grupos se justifica, pois a perda de grande parte das interações presenciais diretas prejudica a aquisição e formação, por parte do professor, dos conhecimentos tácitos sobre os alunos que seriam utilizados, como na sala de aula, para a formação dos grupos.

Ao pensarmos, inicialmente, na utilização da Análise de Agrupamento nos ambientes de EAD, possuíamos como objetivo a identificação dos perfis dos alunos que estivessem reunidos em um grupo, porém, percebemos durante a pesquisa que essa propriedade não ocorreria de imediato, pois é um produto que surge (ou não) de uma observação criteriosa nas tarefas que um determinado grupo desempenha, independente se esse grupo foi construído com auxílio das técnicas de Análise de Agrupamento.

Quando pesquisamos as teorias educacionais, entendemos que a nossa proposta poderia ser utilizada com a finalidade de oferecer condições para se alcançar os objetivos que lá estavam: provocar e utilizar as interações entre os participantes para a concretização da aprendizagem. Essas interações que ocorrem mais facilmente no ambiente clássico de ensino, apresentam-se diferenciadas e mais enigmáticas em ambientes EAD. As dificuldades que ocorrem podem ser resumidas, principalmente, pela falta das situações *tête-à-tête* entre os participantes.

A revisão da literatura e os estudos de caso proporcionaram um aprendizado consistente das técnicas de Análise de Agrupamento, sobretudo quando da prática ocorrida nos estudos de caso, onde percebemos que estávamos diante de uma ferramenta que além de gerar grupos naturais de participantes, permitia levantar hipóteses e identificar características sobre esses grupos. Se antes não existiam informações suficientes sobre os participantes que pudessem ser utilizadas, a partir da construção dos grupos essas informações começaram a aparecer, possibilitando até mesmo o surgimento de taxonomias.

Os resultados gerados nos estudos de caso também foram determinantes no sentido de: esclarecer e consolidar os conceitos de Análise de Agrupamento; facilitar a identificação de caminhos necessários ao andamento da dissertação.

No primeiro estudo de caso foi possível, através da metodologia utilizada para avaliar as diversas análises de agrupamento efetuadas, identificar os coeficientes de similaridade mais indicados para um contexto semelhante com variáveis qualitativas envolvidas. Também foi possível identificar para a plataforma CSCL, de onde foram recuperados os dados qualitativos, os requisitos necessários para a construção de um arquivo Log suficiente para a aplicação das técnicas de análise de agrupamento. Além disso, surgiu nesse estudo de caso, por conta da prática efetuada, uma sistemática voltada para a análise dos resultados de uma Análise de Agrupamento, permitindo verificar se os resultados produzidos

foram consistentes. Essa sistemática foi adicionada, em parte, na implementação da interface produzida na plataforma Pii.

O segundo estudo foi caracterizado pela aplicação da Análise de Agrupamento nos dados dos alunos de um curso, Física para a 4ª Série do ensino fundamental, apoiado por uma plataforma clássica de EAD, a Pii. Esses dados foram identificados de acordo com o cenário pedagógico desejado pelo professor da disciplina, e era constituído por variáveis indicando a visita em determinadas unidades didáticas e o tempo total despendido por cada aluno no ambiente. Os resultados desse estudo propiciaram a identificação dos cenários pedagógicos mais comuns que poderiam ser utilizados em análises de agrupamento, os quais foram consolidados e implementados na interface produzida para a Pii. Esse estudo também permitiu identificar melhorias, baseadas na revisão que fizemos sobre esse assunto, no Log dessa plataforma o que possibilitará ao mesmo tornar-se um conjunto de dados mais rico do que já é, atualmente.

A interface construída para a Pii laureou toda a pesquisa desenvolvida, pois possibilitou ao professor de algum curso apoiado pela plataforma, a geração de grupos homogêneos através das técnicas de Análise de Agrupamento aplicadas nos dados dos alunos, de acordo com um cenário pedagógico escolhido. A interface também apresenta a caracterização dos grupos, através da exibição de estatísticas associadas aos grupos, e uma avaliação dos agrupamentos produzidos utilizando para isso o coeficiente de correlação cofenética.

Podemos considerar que, por conta dos resultados obtidos durante a dissertação, o objetivo central do trabalho, a utilização de Análise de Agrupamento em ambientes virtuais de aprendizagem para a obtenção de grupos de alunos homogêneos, traçado inicialmente na introdução e consolidado na seção 3.3, mostrou-se promissora.

Por conta disso, nos sentimos a confirmar o que o educador Paulo Freire explana em sua última entrevista (FREIRE, 1997), ao considerar que o ser humano detém a capacidade de se inserir no mundo, e que essa inserção é representada pela tomada de decisão no sentido da intervenção a ser feita no mundo em que vive. Acreditamos que foi isso que de certa maneira fizemos ao buscar no ambiente de educação a distância (um mundo particular) as inserções registradas dos alunos, utilizando-as para darmos condições (e aí se situou a nossa inserção) para que acontecessem as interações produtivas, como as que ocorrem nas boas salas de aula. Essas interações ocorrerão mais facilmente, no nosso entendimento, através da utilização de grupos (formados a partir dos grupos homogêneos gerados) para a realização das tarefas educacionais.

5.2 Trabalhos Futuros

A interface construída para a Pii continua sendo trabalhada para disponibilizar outras técnicas e algoritmos de Análise de Agrupamento como, por exemplo, técnicas de partição que utilizam o algoritmo k-Means. A implementação de uma ferramenta gráfica que produza o Dendrograma relacionado aos grupos gerados através da interface, com o intuito de facilitar a interpretação dos resultados pelo professor, é um dos trabalhos que podem vir a ser desenvolvido, pois alguns pacotes estatísticos oferecem esse tipo de facilidade ao receber os dados e, a partir deles, gerar um arquivo contendo o dendrograma. O desenvolvimento das rotinas que produziram a interface gerou outras possibilidades para a plataforma, através da reutilização de algumas dessas rotinas em regiões externas a da interface da Análise de Agrupamento. Uma dessas rotinas fornece o total de acessos, por unidade didática, de cada aluno em um determinado período de tempo.

O Log da plataforma Pii, bastante analisado no estudo de caso 2, possibilita o acréscimo de outras funcionalidades que buscariam registrar outros eventos da plataforma. A implementação dessas funcionalidades aumentará o escopo da análise de agrupamento aplicada à plataforma Pii. O desenvolvimento e a construção de um Log para o ambiente de CSCL TeamWorks, por conta da identificação dos requisitos necessários, ocorrida na seção 3.1, é um outro tópico a ser considerado.

O trabalho desenvolvido permitiu identificar a falta de informações na literatura acerca dos coeficientes de similaridade destinados a variáveis qualitativas. A solução para suprir essa falta de informações, em nosso entendimento, passaria pela execução de pesquisas centradas na identificação das principais características desses coeficientes, assim como a adequabilidade dos mesmos em situações de análise de agrupamento que envolva variáveis qualitativas.

A produção dos dendrogramas através do aplicativo estatístico utilizado, permitiu identificar lacunas referentes à programação destinada a Análise de Agrupamento neste aplicativo. Essas lacunas poderão ser solucionadas pois o aplicativo estatístico R é desenvolvido por um grupo de programadores, ligados ao projeto GNU, receptíveis a pacotes, construídos por terceiros, desenvolvidos e que apresentem funcionalidades inéditas que, se úteis, poderão ser disponibilizadas para os usuários. Portanto, são reais as possibilidades para a construção de um pacote com o intuito de suprir essas lacunas encontradas.

Acreditamos que as informações contidas no presente trabalho em conjunto com o conteúdo do Apêndice C, onde se encontram rotinas que executam uma Análise de Agrupamento específica, poderão servir de base para o desenvolvimento de um componente (ou um adaptador) que possa vir a ser adicionado e utilizado em qualquer ambiente virtual de aprendizagem.

Por fim, é nossa intenção contribuir para a disseminação da cultura da coleta das informações contidas nos arquivos de Log de ambientes virtuais de aprendizagem, alertando que essas coletas tendem a identificar e produzir padrões, os quais, normalmente, geram conhecimentos significativos acerca dos participantes desses ambientes. Esse tipo de cultura, muito comum em ambientes *e-commerce*, ainda se encontra na fase de crescimento moderado no ambiente *e-learning*.

REFERÊNCIAS

BARROSO, L.P.; ARTES, R. **Análise Multivariada: minicurso do 10º Simpósio de Estatística Aplicada a Experimentação Agronômica,** Lavras: – Departamento de Ciências Exatas, Universidade Federal de Lavras – MG. 151p., 2003

BATISTA, G.E.A.P. **Pre-processamento de Dados em Aprendizado de Máquina Supervisionado**. Tese (Doutorado) - ICMC – USP, São Paulo, 2003.

BERGMAN, E. M.; FESER, E. J. **Industrial and Regional Clusters: Concepts and Comparative Applications**. Virginia: West Virginia University - Regional Research Institute, 1998. Disponível em http://www.rri.wvu.edu/WebBook/Bergman-Feser/contents.htm. Acesso em: março 2005

BUSSAB, W. de O.; MIAZAKI, E.S.; ANDRADE, D.F. **Introdução à Análise de Agrupamentos.** In: 9º Simpósio Nacional de Probabilidade e Estatística, São Paulo. Associação Brasileira de Estatística, 105p.,1990.

BUSSAB, W. DE O.; MORETTIN, P.A. **Estatística Básica**. 5ª ed. São Paulo, 526p., 2003.

CENTRO PAULO FREIRE, de Estudos e Pesquisas. **Biografia de Paulo Freire**. Disponível em: http://www.paulofreire.org.br/asp/Index.asp. Acesso em: jan. 2005.

CHAMOVITZ, I. *et al*. Processo Cooperativo de Elaboração de Um Projeto de Pesquisa: A Contribuição do Uso de Uma Plataforma de Ensino a Distância. In: **X WIE - Workshop sobre Informática na Escola**, Salvador-BA. Anais do X WIE (SBC), 2004. Disponível em http://www.api.adm.br/GRS/publicados/GRS_wie_sbc2004_3939.pdf. Acesso em: junho 2004.

COOLEY, R.; MOBASHER, B.; SRIVASTAVA, J. **Web Usage Minning: discovery and application of interesting paterns from web data**. Ph.d.dissertation, Department of computer science, University of Minnesota, Mineapolis, 2000.

CLIFFORD, H. T.; STEPHENSON, W. **An Introduction to numerical taxonomy**. London: Academic Press, 229p., 1975.

DEMO, P. É Errando Que a Gente Aprende [Entrevista a Ricardo Prado]. **REVISTA NOVA ESCOLA**, Brasília: v. 16, n. 144, p. 49-51, agosto de 2001. Disponível em http://novaescola.abril.com.br/index.htm?ed/144_ago01/html/fala_mestre. Acesso em: fev. 2005.

_____. **Pesquisa: Princípio Científico e Educativo**. 3ª ed. São Paulo, Cortez, 1992.

ELIA, M.F.; SAMPAIO, F.F. Plataforma Interativa para Internet(PII): Uma proposta de Pesquisa-Ação a Distância para Professores, **Anais SBIE 2001 - XII Simpósio Brasileiro de Informática na Educação**, 102-109, 2001.

EVERITT, B. S. **Cluster Analysis: A Survey**, Springer-Verlog, Berlin, 1974.

FLENSBURG, FH, **Sorting algorithms – Quicksort**. Disponível em http://www.iti.fh-flensburg.de/lang/algorithmen/sortieren/quick/quicken.htm. Acesso em: jan. 2005.

FREIRE, P. **A Importância do Ato de Ler: em três artigos que se completam**. São Paulo: Cortez, 1997.

FREIRE, P. *et al*. **O Processo Educativo Segundo Paulo Freire e Pichon-Rivière**, 2ª ed. Petrópolis, Vozes, 1989.

GOWER, J. C. **Measures of similarity, dissimilarity, and distance.** In: Encyclopedia of Statistical Sciences, Vol. 5, ed. S. Kotz, N. L. Johnson, and C. B. Read, 397–405. New York: John Wiley & Sons, 1985.

HAIR, Jr., J. F., ANDERSON, R. E., TATHAM, R. L., BLACK, W. C. **Multivariate Data Analysis** (5th ed.). New York: Macmillan Publishing Company, 1998.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis.** Third Edition, Prentice-Hall, 1988.

KOHAVI, R. Mining E-Commerce Data: The Good, The Bad, and the Ugly. In: **The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining** - KDD 2001, San Francisco, Estados Unidos, 2001.

KOSALA, R.; BLOCKELL, H. **Web mining research: A survey**. ACM SIGKDD Explorations, 2(1):1-15, 2000. Disponível em http://www.acm.org/sigs/sigkdd/explorations/issue2-1/kosala.pdf. Acesso em: jan. 2004

KRZANOWSKI, W. J.; MARRIOTT, F. H. C. **Multivariate Analysis Part 2**. John Wiley 1a Edição, 1995

LIKERT, R. **A Technique for the Measurement of Attitudes**. Archives of Psychology; 140:1-50, 1932.

MARTINS, J.C. **Vygotsky e o Papel das Interações Sociais na Sala de Aula: Reconhecer e Desvendar o Mundo**, 1997. Disponível em http://cre.edunet.sp.gov.br/pdf/ideias_28_p111-122_c.pdf . Acesso em: jan. 2005.

MICROSOFT Corporation, **Internet Information Services (IIS) 6.0 Resource Kit** - A Resource Kit Publication, 2003.

MOTTA, C.L.R.; BORGES, M.R.S. TeamWorks: teamwork collaborative environment. In: **Proceedings of Sixth Brazilian Symposium of Multimedia and Hypermedia**, 2000, p.259-272.

MULTIEDUCAÇÃO, C. I**. O significado de Zona de Desenvolvimento Proximal na Teoria Histórico Cultural**. Disponível em: http://www.rio.rj.gov.br/multirio/cime/ME03/ME03_008.html. Acesso em: jan. 2005.

NICOL, D. M. **WIMPE : Web Interface for Managing Programs Electronically**, 2001. Disponível em http://www.cs.dartmouth.edu/~nicol/wimpe/wimpe6.1.html. Acesso em: dez. de 2004.

OLIVEIRA, M. K. ; **VYGOTSKY - Aprendizado e desenvolvimento um processo sóciohistórico**. 4 ed., São Paulo, Scipione, 1997.

PNUD Brasil (Programa Nacional das Nações Unidas para o Desenvolvimento). **Desenvolvimento Humano e IDH.** Disponível em: http://www.pnud.org.br/idh/. Acesso em: março de 2005.

ROQUE, G. O. B.; CHAMOVITZ, I.; CAMPOS, J.A.S.; ARAUJO, J.F.S., GOUVEA, M.T.ª; CARDOSO, R.P.; AZAMBUJA, S.; MOURA, S.A. Aspectos Relevantes para o Desenvolvimento de Ambientes Educacionais para a WEB. In: **SIECI 2004: Simpósium Iberoamericano de Educación, Cibernética e Informática**, Orlando - Florida: Instituto Internacional de Informática y Sistémica (IIIS), 2004. Disponível em http://www.api.adm.br/GRS/publicados Acesso em: fev. de 2005.

SILVA, D. R., SENO, W. P., VIEIRA, M. T. P. Acompanhamento do Aprendizado em Educação a Distância com Uso de Data Mining. In: **Proceedings of XXVII Conferência Latinoamericana de Informática**, Mérida, Venezuela, 2001.

SILVA, D. R.; VIEIRA, M. T. P.; Modelo para acompanhamento do aprendizado em educação a distância. In: **Anais do VII Workshop de Informática na Escola**, Fortaleza, Brasil, 2001.

SOUTO, M. A. M. **Diagnóstico *On-Line* do Estilo Cognitivo de Aprendizagem do Aluno em Um Ambiente Adaptativo de Ensino e Aprendizagem na Web: uma Abordagem Empírica baseada na sua Trajetória de Aprendizagem**. Tese de Doutorado: Programa de Pós-Graduação em Computação, Porto Alegre, 2003.

TAROUCO, L. M. R., HACK, L., GELLER, M. O Processo de avaliação na educação a distância. In: **IV Workshop Informática na Educação**, Porto Alegre: Gráfica da UFRGS. p.71 – 90, 2000.

VYGOTSKY, L. S. **Pensamento e Linguagem**, 3ed. São Paulo, Martins Fontes, 1991.

ZAIANE, O. R. WEB Mining: Concepts, Practices and Research. In: **SIMPÓSIO BRASILEIRO DE BANCO DE DADOS**, João Pessoa, 2000. Disponível em http://www.cs.ualberta.ca/~zaiane/courses/sbbd2000/sbbdtuto.pdf. Acesso em: junho 2004.

APÊNDICES

APÊNDICE A – Exemplos de Classificações

Na Classificação de Harvard, as estrelas são classificadas em tipos ou classes espectrais, arranjada de acordo com as intensidades das linhas espectrais de certos elementos químicos dominantes. Essa classificação é uma seqüência representada por letras (O, B, A, F, G, K e M), ordenadas da maior para a menor temperatura de superfície, com isso as estrelas mais próximas de O (brancas e azuis) são aquelas que apresentam as maiores temperaturas de superfície e são conhecidas por estrelas do tipo *early* (sendo as mais novas do universo), enquanto que as estrelas mais próximas de M (vermelhas) são conhecidas por serem do tipo *late*, apresentando temperatura em torno dos 3.000 K. Cada um dos sete tipos espectrais (O até M) é subdividido em 10 grupos, de 0 (*early*) a 9 (*late*), possibilitando vários tipos diferentes de classificação – O5, O7, F4, K3, M1, etc. Uma estrela do tipo O2 possivelmente apresenta temperatura de superfície em torno de 28.000 K. Essa classificação possibilitou referir-se a uma estrela como sendo, por exemplo, "estrela tipo O" ficando subentendido que se trata de uma estrela cujo tipo espectral é O, segundo a Classificação de Harvard[45]. O quadro a seguir, apresenta características e exemplos dessa classificação de estrelas.

Tipo Esp.	Cor	Tsup (K)	Linhas proeminentes de absorção	Estrelas
O	Branco-Azulada e Azul	30.000	He ionizado (fortes), elementos pesados ionizados (OIII, NIII, SiIV), fracas linhas de H	Puppis
B	Azulada	20.000	He neutro (moderadas), elementos pesados 1 vez ionizados	Rigel (B8), Lyrae
A	Branca	10.000	He neutro (muito fracas), ionizados, H (fortes)	Vega (A0) Sirius (A1)
F	Amarelada	7.000	elementos pesados 1 vez ionizados, metais neutros (Fe I, Ca I), H (moderadas)	Canopus (F0), Procyon

[45]Uma outra forma de classificação das estrelas refere-se à dimensão das mesmas, sendo possível dividi-las em Supergigantes, Gigantes brilhantes, Gigantes, Subgigantes, Anãs ou Normais e Sub-anãs.

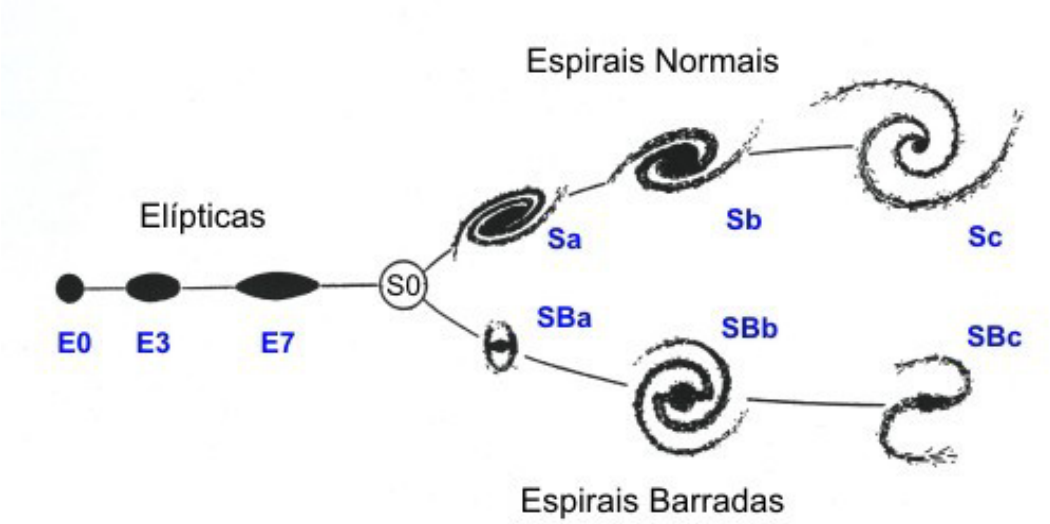
G	Amarela	6.000	elementos pesados 1 vez ionizados, metais neutros, H (relativamente fracas)	Sol (G2) Alfa Centauri (G2)
K	Laranja	4.000	elementos pesados 1 vez ionizados, metais neutros, H (fracas)	Arcturus (K2) Aldebaran (K5)
M	Vermelha	3.000	Átomos neutros (fortes), moleculares (moderadas), H (muito fracas)	Betelgeuse (M2), Antares

Hubble, por sua vez, em sua classificação das galáxias, utilizou o critério aspecto como elemento classificador. Por isso sua classificação ficou conhecida como morfológica. Ele classificou as galáxias dividindo-as em três grupos principais: Espirais, Elípticas e Irregulares.

As espirais são as mais observadas e constituem 65% das galáxias conhecidas, caracterizando-se por apresentarem braços espiralados com origem no centro, dividem-se ainda em espirais normais (quando os braços espirais partem do núcleo) e barradas (os braços espirais se desenvolvem a partir de uma barra luminosa que atravessa o núcleo). São representadas por tipos ou classes, S (espirais normais) e SB (espirais barradas), os quais possuem ainda os subtipos a, b ou c referindo-se a diferentes graus de enrolamento dos braços e diferentes proporções de tamanho do bojo em relação à galáxia. As galáxias Via Láctea (entre o tipo Sb e Sc) e Andrômeda são classificadas como espirais.

As galáxias elípticas caracterizam-se por possuírem estrelas mais antigas e frias, sendo escassas as estrelas do tipo O e B além, é claro, de apresentarem uma forma elíptica, como a galáxia Centaurus A. São representadas pela letra E, podendo ser seguida ainda por um número entre 0 e 7, correspondente aos seus eixos maior e menor, assim as galáxias quase esféricas são as do tipo E0, e as bem achatadas são as do tipo E7.

As irregulares não possuem uma forma bem definida, dividindo-se em galáxias onde é possível distinguir suas estrelas e naquelas onde não é possível. São exemplos de Irregulares a Grande e a Pequena Nuvem de Magalhães. A figura a seguir, mostra galáxias elípticas e espirais conforme a Classificação Morfológica de Hubble.



APÊNDICE B – Complemento das Etapas das Análises de Agrupamento da Seção 3.1

1 – Etapas da Análise de Agrupamento utilizando o Coeficiente de Pearson

- Segundo grupo gerado e identificação do 3º

	Centro, Barra	**Tijuca**	**Ilha**	**Leblon, Ipanema**	**Glória**	**Urca**	**Saúde**
Tijuca	0,218218						
Ilha	0	0,327327					
Leblon, Ipanema	0,654654	0,428571	0,218218				
Glória	0,25	-0,65465	-0,10206	0,534522			
Urca	0,5	0,327327	0,375	**0,763763**	0,408248		
Saúde	0,6	0,654654	0,5	0,654654	0	0,5	
Catete	0	0,763763	0,25	0,327327	-0,40825	0,25	0,5

- Terceiro grupo gerado e identificação do 4º

	Centro, Barra	**Tijuca**	**Ilha**	**Glória**	**Urca, (Leblon, Ipanema)**	**Saúde**
Tijuca	0,218218					
Ilha	0	0,327327				
Glória	0,25	-0,65465	-0,10206			
Urca, (Leblon, Ipanema)	0,5	0,327327	0,218218	0,408248		
Saúde	0,6	0,654654	0,5	0	0,5	
Catete	0	**0,763763**	0,25	-0,40825	0,25	0,5

- Quarto grupo gerado e identificação do 5º

	Centro, Barra	**Ilha**	**Glória**	**Urca, (Leblon, Ipanema)**	**Saúde**
Ilha	0				

Glória	0,25	-0,10206			
Urca, (Leblon, Ipanema)	0,5	0,218218	0,408248		
Saúde	**0,6**	0,5	0	0,5	
Catete, Tijuca	0	0,25	-0,65465	0,25	0,5

- Quinto grupo gerado e identificação do 6º

	(Centro, Barra), Saúde	**Ilha**	**Glória**	**Urca, (Leblon, Ipanema)**
Ilha	0			
Glória	0	-0,10206		
Urca, (Leblon, Ipanema)	**0,5**	0,218218	0,408248	
Catete, Tijuca	0	0,25	-0,65465	0,25

- Sexto grupo gerado e identificação do 7º

	(Centro, Barra, Saúde), (Urca, Leblon, Ipanema)	**Ilha**	**Glória**
Ilha	0		
Glória	0	-0,10206	
(Catete, Tijuca)	0	**0,25**	-0,65465

- Sétimo grupo gerado e identificação do 8º

	(Centro, Barra, Saúde), (Urca, Leblon, Ipanema)	**Glória**
Glória	0	
(Catete, Tijuca), Ilha	0	-0,65465

- Oitavo grupo gerado e identificação do 9º

	(Centro, Barra, Saúde), (Urca, Leblon, Ipanema), Glória
(Catete, Tijuca), Ilha	-0,65465

2 – Etapas da Análise de Agrupamento utilizando o Coeficiente de Concordâncias Simples

- Identificação do 1º a ser gerado

	Barra	**Tijuca**	**Ilha**	**Centro**	**Leblon**	**Glória**	**Urca**	**Ipanema**	**Saúde**
Tijuca	0,6								
Ilha	0,5	0,5							
Centro	0,9	0,7	0,6						
Leblon	0,8	0,6	0,7	0,9					
Glória	0,7	0,2	0,4	0,6	0,7				
Urca	0,7	0,5	0,8	0,8	0,9	0,6			
Ipanema	0,8	0,6	0,7	0,9	**1**	0,7	0,9		
Saúde	0,8	0,8	0,7	0,9	0,8	0,5	0,7	0,8	
Catete	0,5	0,9	0,4	0,6	0,5	0,4	0,4	0,5	0,7

- Primeiro grupo gerado e identificação do 2º

	Barra	Tijuca	Ilha	Centro	Leblon, Ipanema	Glória	Urca	Saúde
Tijuca	0,6							
Ilha	0,5	0,5						
Centro	**0,9**	0,7	0,6					
Leblon, Ipanema	0,8	0,6	0,7	0,9				
Glória	0,7	0,2	0,4	0,6	0,7			
Urca	0,7	0,5	0,8	0,8	0,9	0,6		
Saúde	0,8	0,8	0,7	0,9	0,8	0,5	0,7	
Catete	0,5	0,9	0,4	0,6	0,5	0,4	0,4	0,7

- Segundo grupo gerado e identificação do 3º

	Barra, Centro	Tijuca	Ilha	Leblon, Ipanema	Glória	Urca	Saúde
Tijuca	0,6						
Ilha	0,5	0,5					
Leblon, Ipanema	0,8	0,6	0,7				
Glória	0,6	0,2	0,4	0,7			
Urca	0,7	0,5	0,8	**0,9**	0,6		
Saúde	0,8	0,8	0,7	0,8	0,5	0,7	
Catete	0,5	0,9	0,4	0,5	0,4	0,4	0,7

- Terceiro grupo gerado e identificação do 4º

	Barra, Centro	Tijuca	Ilha	(Leblon, Ipanema), Urca	Glória	Saúde
Tijuca	0,6					
Ilha	0,5	0,5				
Leblon, Ipanema, Urca	0,7	0,5	0,7			
Glória	0,6	0,2	0,4	0,6		
Saúde	0,8	0,8	0,7	0,7	0,5	
Catete	0,5	**0,9**	0,4	0,4	0,4	0,7

- Quarto grupo gerado e identificação do 5º

	Barra, Centro	Ilha	(Leblon, Ipanema), Urca	Glória	Saúde
Ilha	0,5				
Leblon, Ipanema, Urca	0,7	0,7			
Glória	0,6	0,4	0,6		
Saúde	**0,8**	0,7	0,7	0,5	
Catete, Tijuca	0,5	0,4	0,4	0,2	0,7

- Quinto grupo gerado e identificação do 6º

	Barra, Centro, Saúde	Ilha	(Leblon, Ipanema), Urca	Glória
Ilha	0,5			
Leblon, Ipanema, Urca	**0,7**	0,7		

Glória	0,5	0,4	0,6	
Catete, Tijuca	0,5	0,4	0,4	0,2

- Sexto grupo gerado e identificação do 7º

	Barra, Centro, Saúde, Leblon, Ipanema, Urca	**Ilha**	**Glória**
Ilha	**0,5**		
Glória	0,5	0,4	
Catete, Tijuca	0,4	0,4	0,2

- Sétimo grupo gerado e identificação do 8º

	Barra, Centro, Saúde, Leblon, Ipanema, Urca, Ilha	**Glória**
Glória	**0,4**	
Catete, Tijuca	0,4	0,2

- Oitavo grupo gerado e identificação do 9º

	Barra, Centro, Saúde, Leblon, Ipanema, Urca, Ilha, Glória
Catete, Tijuca	**0,2**

Matriz Cofenética referente ao Coeficiente de Concordâncias Simples

	Barra	**Tijuca**	**Ilha**	**Centro**	**Leblon**	**Glória**	**Urca**	**Ipanema**	**Saúde**
Tijuca	0,2								
Ilha	0,5	0,2							
Centro	0,9	0,2	0,5						
Leblon	0,7	0,2	0,5	0,7					
Glória	0,4	0,2	0,5	0,4	0,4				
Urca	0,7	0,2	0,5	0,7	0,9	0,4			
Ipanema	0,7	0,2	0,5	0,7	1	0,4	0,9		
Saúde	0,8	0,2	0,5	0,8	0,7	0,4	0,7	0,7	
Catete	0,2	0,9	0,2	0,2	0,2	0,2	0,2	0,2	0,2

3 – Etapas da Análise de Agrupamento utilizando o Coeficiente de Jaccard

Os quadros a seguir, exibem a evolução da Matriz de Distâncias calculada para os pares de avaliadores, através do coeficiente de Jaccard, já inicializada com o primeiro grupo formado.

- Primeiro grupo gerado e identificação do 2º

	Barra	**Tijuca**	**Ilha**	**Centro**	**Leblon, Ipanema**	**Glória**	**Urca**	**Saúde**
Tijuca	0,333							
Ilha	0,444	0,375						
Centro	0,833	0,500	0,556					
Leblon, Ipanema	0,714	0,429	0,667	0,857				
Glória	0,500	0,000	0,333	0,429	0,571			
Urca	0,625	0,375	0,778	0,750	**0,875**	0,500		
Saúde	0,667	0,600	0,625	0,833	0,714	0,286	0,625	
Catete	0,167	0,667	0,250	0,333	0,286	0,000	0,250	0,400

- Segundo grupo gerado e identificação do 3º

	Barra	**Tijuca**	**Ilha**	**Centro**	**Glória**	**Urca, (Leblon, Ipanema)**	**Saúde**
Tijuca	0,333						
Ilha	0,444	0,375					
Centro	**0,833**	0,500	0,556				
Glória	0,500	0,000	0,333	0,429			
Urca, (Leblon, Ipanema)	0,625	0,375	0,667	0,750	0,500		
Saúde	0,667	0,600	0,625	0,833	0,286	0,625	
Catete	0,167	0,667	0,250	0,333	0,000	0,250	0,400

- Terceiro grupo gerado e identificação do 4º

	Barra, Centro	**Tijuca**	**Ilha**	**Glória**	**Urca, (Leblon, Ipanema)**	**Saúde**
Tijuca	0,333					
Ilha	0,444	0,375				
Glória	0,429	0,000	0,333			
Urca, (Leblon, Ipanema)	0,625	0,375	**0,667**	0,500		
Saúde	0,667	0,600	0,625	0,286	0,625	
Catete	0,167	0,667	0,250	0,000	0,250	0,400

- Quarto grupo gerado e identificação do 5º

	Barra, Centro	**Tijuca**	**Glória**	**Ilha, (Urca, Leblon, Ipanema)**	**Saúde**
Tijuca	0,333				
Glória	0,429	0,000			
Ilha , (Urca, Leblon, Ipanema)	0,444	0,375	0,333		
Saúde	0,667	0,600	0,286	0,625	
Catete	0,167	**0,667**	0,000	0,250	0,400

- Quinto grupo gerado e identificação do 6º

	Barra, Centro	Glória	Ilha, (Urca, Leblon, Ipanema)	Saúde
Glória	0,429			
Ilha , (Urca, Leblon, Ipanema)	0,444	0,333		
Saúde	**0,667**	0,286	0,625	
Catete, Tijuca	0,167	0,000	0,250	0,400

- Sexto grupo gerado e identificação do 7º

	(Barra, Centro), Saúde	Glória	Ilha, (Urca, Leblon, Ipanema)
Glória	0,286		
Ilha , (Urca, Leblon, Ipanema)	**0,444**	0,333	
Catete, Tijuca	0,167	0,000	0,250

- Sétimo grupo gerado e identificação do 8º

	(Barra, Centro, Saúde), (Ilha , Urca, Leblon, Ipanema)	Glória
Glória	**0,286**	
Catete, Tijuca	0,167	0,000

- Oitavo grupo gerado e identificação do 9º

	(Barra, Centro, Saúde, Ilha , Urca, Leblon, Ipanema), Glória
Catete, Tijuca	**0,000**

Matriz Cofenética referente ao Coeficiente de Jaccard.

	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde
Tijuca	0,000								
Ilha	0,444	0,000							
Centro	0,833	0,000	0,444						
Leblon	0,444	0,000	0,667	0,444					
Glória	0,286	0,000	0,286	0,286	0,286				
Urca	0,444	0,000	0,667	0,444	0,875	0,286			
Ipanema	0,444	0,000	0,667	0,444	1	0,286	0,875		
Saúde	0,667	0,000	0,444	0,667	0,444	0,286	0,444	0,444	
Catete	0,000	0,667	0,000	0,000	0,000	0,000	0,000	0,000	0,000

4 – Etapas da Análise de Agrupamento utilizando o Coeficiente de Russel e Rao

Os Quadros a seguir, exibem a evolução da Matriz de Distâncias calculada para os pares de avaliadores, através do coeficiente de Russel e Rao.

- Identificação do 1º grupo

	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde
Tijuca	0,2								
Ilha	**0,4**	**0,3**							
Centro	0,5	0,3	**0,5**						
Leblon	0,5	0,3	**0,6**	0,6					
Glória	0,3	0,0	**0,3**	0,3	0,4				
Urca	**0,5**	**0,3**	**0,7**	**0,6**	**0,7**	**0,4**			
Ipanema	0,5	0,3	**0,6**	0,6	0,7	0,4	**0,7**		
Saúde	0,4	0,3	**0,5**	0,5	0,5	0,2	**0,5**	0,5	
Catete	0,1	0,2	**0,2**	0,2	0,2	0,0	**0,2**	0,2	0,2

- Primeiro grupo gerado e identificação do 2º

	Barra	Tijuca	Centro	Leblon	Glória	Ilha, Urca	Ipanema	Saúde
Tijuca	0,2							
Centro	0,5	0,3						
Leblon	**0,5**	**0,3**	**0,6**					
Glória	0,3	0	0,3	**0,4**				
Ilha, Urca	0,4	0,3	0,5	**0,6**	0,3			
Ipanema	**0,5**	**0,3**	**0,6**	**0,7**	**0,4**	0,6		
Saúde	0,4	0,3	0,5	**0,5**	0,2	0,5	**0,5**	
Catete	0,1	0,2	0,2	**0,2**	0	0,2	**0,2**	0,2

- Segundo grupo gerado e identificação do 3º

	Barra	Tijuca	Centro	Glória	Ilha, Urca	Ipanema, Leblon	Saúde
Tijuca	0,2						
Centro							
Glória	0,3	0	**0,3**				
Ilha, Urca	0,4	0,3	**0,5**	0,3			
Ipanema, Leblon	**0,5**	**0,3**	**0,6**	**0,4**	**0,6**		
Saúde	0,4	0,3	**0,5**	0,2	0,5	**0,5**	
Catete	0,1	0,2	**0,2**	0	0,2	**0,2**	0,2

- Terceiro grupo gerado e identificação do 4º

	Barra	Tijuca	Glória	Ilha, Urca	(Ipanema, Leblon), Centro	Saúde
Tijuca	**0,2**					
Glória	**0,3**	0				
Ilha, Urca	**0,4**	0,3	0,3			
(Ipanema, Leblon), Centro	**0,5**	**0,3**	**0,3**	**0,5**		
Saúde	**0,4**	0,3	0,2	0,5	**0,5**	
Catete	**0,1**	0,2	0	0,2	**0,2**	0,2

- Quarto grupo gerado e identificação do 5º

	Barra, (Ipanema, Leblon, Centro)	Tijuca	Glória	Ilha, Urca	Saúde
Tijuca	0,2				
Glória	0,3	0			
Ilha, Urca	**0,4**	**0,3**	**0,3**		
Saúde	**0,4**	**0,3**	**0,2**	**0,5**	
Catete	0,1	0,2	0	**0,2**	**0,2**

- Quinto grupo gerado e identificação do 6º

	Barra, (Ipanema, Leblon, Centro)	Tijuca	Glória	Saúde, (Ilha, Urca)
Tijuca	**0,2**			
Glória	**0,3**	0		
Saúde, (Ilha, Urca)	**0,4**	**0,3**	**0,2**	
Catete	**0,1**	0,2	0	**0,2**

- Sexto grupo gerado e identificação do 7º

	(Barra, Ipanema, Leblon, Centro), (Saúde, Ilha, Urca)	Tijuca	Glória
Tijuca	**0,2**		
Glória	**0,2**	**0**	
Catete	**0,1**	**0,2**	0

- Sétimo grupo gerado e identificação do 8º

	Tijuca, (Barra, Ipanema, Leblon, Centro, Saúde, Ilha, Urca)	Glória
Glória	0	
Catete	**0,1**	0

- Oitavo grupo gerado e identificação do 9º

	Catete, (Tijuca, Barra, Ipanema, Leblon, Centro, Saúde, Ilha, Urca)
Glória	**0**

Matriz Cofenética referente ao Coeficiente de Russel e RAO

	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde
Tijuca	**0,2**								
Ilha	0,4	**0,2**							
Centro	0,5	**0,2**	0,4						
Leblon	0,5	**0,2**	0,4	0,6					
Glória	0	0	**0**	0	0				
Urca	0,4	**0,2**	0,7	0,4	0,4	**0**			
Ipanema	0,5	**0,2**	0,4	0,6	0,7	0	0,4		
Saúde	0,4	**0,2**	0,5	0,4	0,4	0	0,5	0,4	
Catete	0,1	0,1	0,1	0,1	0,1	0	0,1	0,1	0,1

5 – Etapas da Análise de Agrupamento utilizando o Coeficiente de Hamann

Os quadros a seguir, exibem a evolução da Matriz de Distâncias calculada para os pares de avaliadores, através do coeficiente de Hamann, já inicializada com o primeiro grupo formado.

- Identificação do 1º grupo

	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde
Tijuca	0,2								
Ilha	0,0	0,0							
Centro	0,8	0,4	0,2						
Leblon	0,6	0,2	0,4	0,8					
Glória	0,4	-0,6	-0,2	0,2	0,4				
Urca	0,4	0,0	0,6	0,6	0,8	0,2			
Ipanema	0,6	0,2	0,4	0,8	**1,0**	0,4	0,8		
Saúde	0,6	0,6	0,4	0,8	0,6	0,0	0,4	0,6	
Catete	0,0	0,8	-0,2	0,2	0,0	-0,2	-0,2	0,0	0,4

- Primeiro grupo gerado e identificação do 2º

	Barra	Tijuca	Ilha	Centro	Glória	Urca	Leblon, Ipanema	Saúde
Tijuca	0,2							
Ilha	0,0	0,0						
Centro	**0,8**	0,4	0,2					
Glória	0,4	-0,6	-0,2	0,2				
Urca	0,4	0,0	0,6	0,6	0,2			
Leblon, Ipanema	0,6	0,2	0,4	0,8	0,4	0,8		
Saúde	0,6	0,6	0,4	0,8	0,0	0,4	0,6	
Catete	0,0	0,8	-0,2	0,2	-0,2	-0,2	0,0	0,4

- Segundo grupo gerado e identificação do 3º

	Barra, Centro	Tijuca	Ilha	Glória	Urca	Leblon, Ipanema	Saúde
Tijuca	0,2						
Ilha	0,0	0,0					
Glória	0,2	-0,6	-0,2				
Urca	0,4	0,0	0,6	0,2			
Leblon, Ipanema	0,6	0,2	0,4	0,4	0,8		
Saúde	0,6	0,6	0,4	0,0	0,4	0,6	
Catete	0,0	**0,8**	-0,2	-0,2	-0,2	0,0	0,4

- Terceiro grupo gerado e identificação do 4º

	Barra, Centro	Ilha	Glória	Urca	Leblon, Ipanema	Saúde
Ilha	0,0					
Glória	0,2	-0,2				
Urca	0,4	0,6	0,2			
Leblon, Ipanema	0,6	0,4	0,4	**0,8**		
Saúde	0,6	0,4	0,0	0,4	0,6	
Tijuca, Catete	0,0	-0,2	-0,6	-0,2	0,0	0,4

- Quarto grupo gerado e identificação do 5º

	Barra, Centro	Ilha	Glória	Urca, (Leblon, Ipanema)	Saúde
Ilha	0,0				
Glória	0,2	-0,2			
Urca, (Leblon, Ipanema)	0,4	0,4	0,2		
Saúde	**0,6**	0,4	0,0	0,4	
Tijuca, Catete	0,0	-0,2	-0,6	-0,2	0,4

- Quinto grupo gerado e identificação do 6º

	Saúde, (Barra, Centro)	Ilha	Glória	Urca, (Leblon, Ipanema)
Ilha	0,0			
Glória	0,0	-0,2		
Urca, (Leblon, Ipanema)	**0,4**	0,4	0,2	
Tijuca, Catete	0,0	-0,2	-0,6	-0,2

- Sexto grupo gerado e identificação do 7º

	Saúde, (Barra, Centro)	Glória	Ilha, (Urca, Leblon, Ipanema)
Glória	**0,0**		
Ilha, (Urca, Leblon, Ipanema)	0,0	-0,2	
Tijuca, Catete	0,0	-0,6	-0,2

- Sétimo grupo gerado e identificação do 8º

	Glória, (Saúde, Barra, Centro)	Ilha, (Urca, Leblon, Ipanema)
Ilha, (Urca, Leblon, Ipanema)	-0,2	
Tijuca, Catete	-0,6	**-0,2**

- Oitavo grupo gerado e identificação do 9º

	Glória, (Saúde, Barra, Centro)
(Tijuca, Catete), (Ilha, Urca, Leblon, Ipanema)	-0,6

Matriz Cofenética referente ao Coeficiente de Hamann

	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde
Tijuca	-0,6								
Ilha	-0,6	-0,2							
Centro	0,8	-0,6	-0,6						
Leblon	-0,6	-0,2	0,4	-0,6					
Glória	0,0	-0,6	-0,6	0,0	-0,6				
Urca	-0,6	-0,2	0,4	-0,6	0,8	-0,6			
Ipanema	-0,6	-0,2	0,4	-0,6	**1,0**	-0,6	0,8		
Saúde	0,6	-0,6	-0,6	0,6	-0,6	0,0	-0,6	-0,6	
Catete	-0,6	0,8	-0,2	-0,6	-0,2	-0,6	-0,2	-0,2	-0,6

6 – Etapas da Análise de Agrupamento utilizando o Coeficiente de Yule

Os Quadros a seguir, exibe a Matriz de Distâncias calculada para os pares de avaliadores, através do coeficiente de Yule.

- Identificação do 1º grupo

	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde
Tijuca	0,45								
Ilha	0,00	1,00							
Centro	1,00	1,00	0,25						
Leblon	1,00	1,00	0,50	1,00					
Glória	0,71	-1,00	-0,25	0,50	1,00				
Urca	1,00	1,00	0,75	1,00	1,00	1,00			
Ipanema	1,00	1,00	0,50	1,00	**1,00**	1,00	1,00		
Saúde	0,88	1,00	1,00	1,00	1,00	0,00	1,00	1,00	
Catete	0,00	1,00	1,00	1,00	1,00	-1,00	1,00	1,00	1,00

- Primeiro grupo gerado e identificação do 2º

	Barra	Tijuca	Ilha	Centro	Glória	Urca	Leblon , Ipanema	Saúde
Tijuca	0,45							
Ilha	0,00	1,00						
Centro	**1,00**	1,00	0,25					
Glória	0,71	-1,00	-0,25	0,50				
Urca	1,00	1,00	0,75	1,00	1,00			
Leblon , Ipanema	1,00	1,00	0,50	1,00	1,00	1,00		

Saúde	0,88	1,00	1,00	1,00	0,00	1,00	1,00	
Catete	0,00	1,00	1,00	1,00	-1,00	1,00	1,00	1,00

- Segundo grupo gerado e identificação do 3º

	Barra, Centro	Tijuca	Ilha	Glória	Urca	Leblon , Ipanema	Saúde
Tijuca	0,45						
Ilha	0,00	1,00					
Glória	0,50	-1,00	-0,25				
Urca	1,00	1,00	0,75	1,00			
Leblon , Ipanema	1,00	1,00	0,50	1,00	**1,00**		
Saúde	0,88	1,00	1,00	0,00	1,00	1,00	
Catete	0,00	1,00	1,00	-1,00	1,00	1,00	1,00

- Terceiro grupo gerado e identificação do 4º

	Barra, Centro	Tijuca	Ilha	Glória	Urca, (Leblon , Ipanema)	Saúde
Tijuca	0,45					
Ilha	0,00	**1,00**				
Glória	0,50	-1,00	-0,25			
Urca, (Leblon , Ipanema)	1,00	1,00	0,50	1,00		
Saúde	0,88	1,00	1,00	0,00	1,00	
Catete	0,00	1,00	1,00	-1,00	1,00	1,00

- Quarto grupo gerado e identificação do 5º

	Barra, Centro	Tijuca, Ilha	Glória	Urca, (Leblon , Ipanema)	Saúde
Tijuca, Ilha	0,00				
Glória	0,50	-1,00			
Urca, (Leblon , Ipanema)	1,00	0,50	1,00		
Saúde	0,88	1,00	0,00	1,00	
Catete	0,00	1,00	-1,00	1,00	**1,00**

- Quinto grupo gerado e identificação do 6º

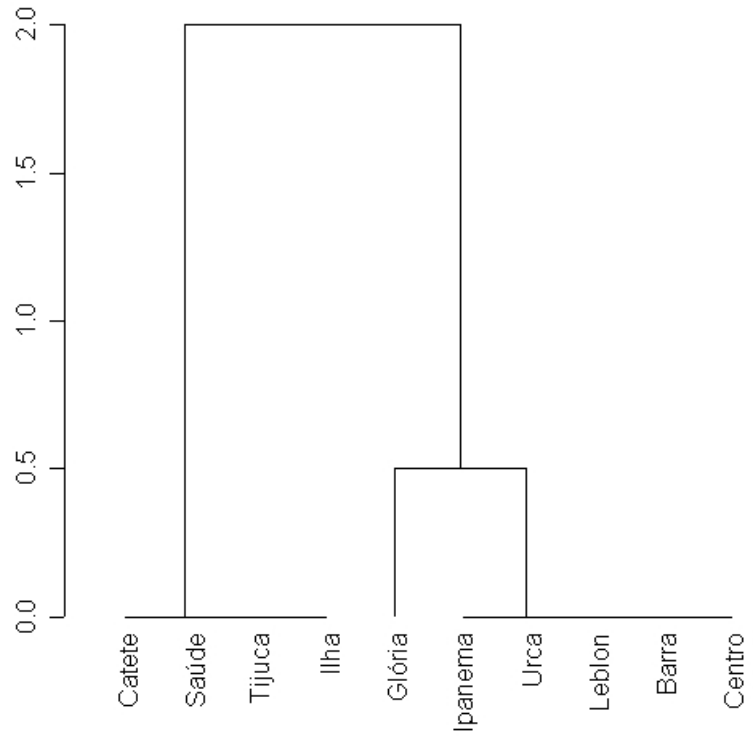
	Barra, Centro	Tijuca, Ilha	Glória, (Urca, Leblon, Ipanema)
Tijuca, Ilha	0,00		
Glória, (Urca, Leblon, Ipanema)	0,50	-1,00	
Saúde, Catete	0,00	**1,00**	-1,00

- Sexto grupo gerado e identificação do 7º

	Barra, Centro	Glória, (Urca, Leblon, Ipanema)
Glória, (Urca, Leblon, Ipanema)	**0,50**	
(Tijuca, Ilha), (Saúde, Catete)	0,00	-1,00

- Sétimo grupo gerado e identificação do 8º

	(Barra, Centro), (Glória, Urca, Leblon, Ipanema)
(Tijuca, Ilha), (Saúde, Catete)	-1,00



Dendrograma da Análise de Agrupamento utilizando Coeficiente de Yule

Matriz Cofenética referente ao Coeficiente de Yule

	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde
Tijuca	-1,00								
Ilha	-1,00	1,0							
Centro	**1,00**	-1,00	-1,00						
Leblon	0,50	-1,00	-1,00	0,50					
Glória	0,50	-1,00	-1,00	0,50	1,0				
Urca	0,50	-1,00	-1,00	0,50	1,0	1,0			
Ipanema	0,50	-1,00	-1,00	0,50	1,0	1,0	1,0		
Saúde	-1,00	1,0	1,0	-1,00	-1,00	-1,00	-1,00	-1,00	
Catete	-1,00	1,0	1,0	-1,00	-1,00	-1,00	-1,00	-1,00	1,0

7 – Etapas da Análise de Agrupamento utilizando o Coeficiente de Gower2

Os quadros a seguir, exibem a Matriz de Distâncias calculada para os pares de avaliadores, através do coeficiente Gower2.

- Identificação do 1º grupo

	Barra	**Tijuca**	**Ilha**	**Centro**	**Leblon**	**Glória**	**Urca**	**Ipanema**	**Saúde**
Tijuca	0,349								
Ilha	0,200	0,327							
Centro	0,816	0,535	0,255						
Leblon	0,655	0,429	0,327	0,802					
Glória	0,490	0,000	0,153	0,375	0,535				
Urca	0,500	0,327	0,438	0,612	0,764	0,408			
Ipanema	0,655	0,429	0,327	0,802	**1,000**	0,535	0,764		
Saúde	0,640	0,655	0,500	0,816	0,655	0,245	0,500	0,655	
Catete	0,200	0,764	0,250	0,408	0,327	0,000	0,250	0,327	0,500

- Primeiro grupo gerado e identificação do 2º

	Barra	**Tijuca**	**Ilha**	**Centro**	**Glória**	**Urca**	**Leblon, Ipanema**	**Saúde**
Tijuca	0,349							
Ilha	0,200	0,327						
Centro	**0,816**	0,535	0,255					
Glória	0,490	0,000	0,153	0,375				
Urca	0,500	0,327	0,438	0,612	0,408			
Leblon, Ipanema	0,655	0,429	0,327	0,802	0,535	0,764		
Saúde	0,640	0,655	0,500	0,816	0,245	0,500	0,655	
Catete	0,200	0,764	0,250	0,408	0,000	0,250	0,327	0,500

- Segundo grupo gerado e identificação do 3º

	Barra, Centro	**Tijuca**	**Ilha**	**Glória**	**Urca**	**Leblon, Ipanema**	**Saúde**
Tijuca	0,349						
Ilha	0,200	0,327					
Glória	0,375	0,000	0,153				
Urca	0,500	0,327	0,438	0,408			
Leblon, Ipanema	0,655	0,429	0,327	0,535	**0,764**		
Saúde	0,640	0,655	0,500	0,245	0,500	0,655	
Catete	0,200	0,764	0,250	0,000	0,250	0,327	0,500

- Terceiro grupo gerado e identificação do 4º

	Barra, Centro	**Tijuca**	**Ilha**	**Glória**	**Urca, (Leblon, Ipanema)**	**Saúde**
Tijuca	0,349					
Ilha	0,200	0,327				
Glória	0,375	0,000	0,153			

Urca, (Leblon, Ipanema)	0,500	0,327	0,327	0,408		
Saúde	0,640	0,655	0,500	0,245	0,500	
Catete	0,200	0,764	0,250	0,000	0,250	0,500

- Quarto grupo gerado e identificação do 5º

	Barra, Centro	**Ilha**	**Glória**	**Urca, (Leblon, Ipanema)**	**Saúde**
Ilha	0,200				
Glória	0,375	0,153			
Urca, (Leblon, Ipanema)	0,500	0,327	0,408		
Saúde	**0,640**	0,500	0,245	0,500	
Tijuca, Catete	0,200	0,250	0,000	0,250	0,500

- Quinto grupo gerado e identificação do 6º

	Barra, Centro, Saúde	**Ilha**	**Glória**	**Urca, (Leblon, Ipanema)**
Ilha	0,200			
Glória	0,245	0,153		
Urca, (Leblon, Ipanema)	**0,500**	0,327	0,408	
Tijuca, Catete	0,200	0,250	0,000	0,250

- Sexto grupo gerado e identificação do 7º

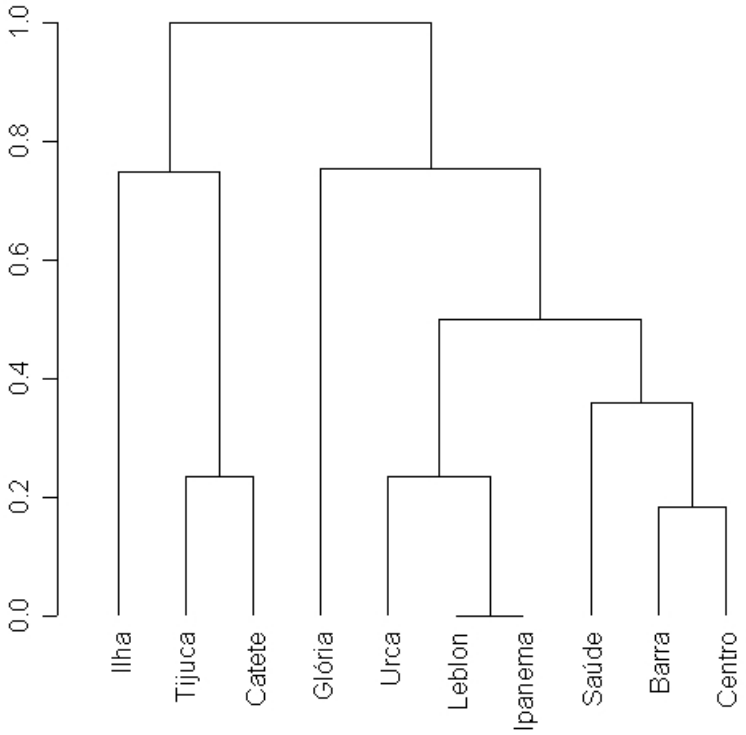
	(Barra, Centro, Saúde), (Urca, Leblon, Ipanema)	**Ilha**	**Glória**
Ilha	0,200		
Glória	0,245	0,153	
Tijuca, Catete	0,200	**0,250**	0,000

- Sétimo grupo gerado e identificação do 8º

	(Barra, Centro, Saúde), (Urca, Leblon, Ipanema)	**Glória**
Glória	**0,245**	
Ilha, (Tijuca, Catete)	0,200	0,000

- Oitavo grupo gerado e identificação do 9º

	Glória, (Barra, Centro, Saúde, Urca, Leblon, Ipanema)
Ilha, (Tijuca, Catete)	**0,000**



Dendrograma da Análise de Agrupamento utilizando Coeficiente Gower2

Matriz Cofenética referente ao Coeficiente Gower2

	Barra	Tijuca	Ilha	Centro	Leblon	Glória	Urca	Ipanema	Saúde
Tijuca	0,000								
Ilha	0,000	0,250							
Centro	0,816	0,000	0,000						
Leblon	0,500	0,000	0,000	0,500					
Glória	0,245	0,000	0,000	0,245	0,245				
Urca	0,500	0,000	0,000	0,500	0,764	0,245			
Ipanema	0,500	0,000	0,000	0,500	1,000	0,245	0,764		
Saúde	0,64	0,000	0,000	0,64	0,500	0,245	0,500	0,500	
Catete	0,000	0,764	0,250	0,000	0,000	0,000	0,000	0,000	0,000

8 – Códigos Utilizados para a Geração dos Dendrogramas no pacote Estatístico R

- **Dendrograma para o Coeficiente de Pearson – Figura 11**

```
MMM <- structure(c(0.218218, 0, 0.816497, 0.654654, 0.408248, 0.5, 0.654654, 0.6, 0, 0.327327, 0.534522, 0.428571, -0.65465, 0.327327, 0.428571, 0.654654, 0.763763, 0.102062, 0.218218, -0.10206, 0.375, 0.218218, 0.5, 0.25, 0.801784, 0.25, 0.612372, 0.801784, 0.816497, 0.408248, 0.534522, 0.763763, 1, 0.654654, 0.327327, 0.408248, 0.534522, 0, -0.40825, 0.763763, 0.5, 0.25, 0.654654, 0.327327, 0.5), Size = as.integer(10), Labels = c("Barra", "Tijuca", "Ilha", "Centro", "Leblon", "Glória", "Urca", "Ipanema", "Saúde", "Catete"), Diag = FALSE, Upper = FALSE, method = "euclidean", call = quote(dist(x = x)), class = "dist")
MMM <- (1-MMM)
GG  <- hclust(MMM, method="complete")
dend1 <- as.dendrogram(GG)
plot(dend1)
```

- **Dendrograma para o Coeficiente de YULE**

```
MMM <- structure(c(0.45, 0, 1, 1, 0.71, 1, 1, 0.88, 0, 1, 1, 1, -1, 1, 1, 1, 1, 0.25, 0.5, -0.25, 0.75, 0.5, 1, 1, 1, 0.5, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, -1, 1, 1, 1, 1, 1, 1), Size = as.integer(10), Labels = c("Barra", "Tijuca", "Ilha", "Centro", "Leblon", "Glória", "Urca", "Ipanema", "Saúde", "Catete"), Diag = FALSE, Upper = FALSE, method = "euclidean", call = quote(dist(x = x)), class = "dist")
MMM <- (1-MMM)
GG  <- hclust(MMM, method="complete")
dend1 <- as.dendrogram(GG)
plot(dend1)
```

- **Dendrograma para o Coeficiente de JACCARD - Figura 12**

```
MMM <- structure(c(0.333, 0.444, 0.833, 0.714, 0.5, 0.625, 0.714, 0.667, 0.167, 0.375, 0.5, 0.429, 0, 0.375, 0.429, 0.6, 0.667, 0.556, 0.667, 0.333, 0.778, 0.667, 0.625, 0.25, 0.857, 0.429, 0.75, 0.857, 0.833, 0.333, 0.571, 0.875, 1, 0.714, 0.286, 0.5, 0.571, 0.286, 0, 0.875, 0.625, 0.25, 0.714, 0.286, 0.4), Size = as.integer(10), Labels = c("Barra", "Tijuca", "Ilha", "Centro", "Leblon", "Glória", "Urca", "Ipanema", "Saúde", "Catete"), Diag = FALSE, Upper = FALSE, method = "euclidean", call = quote(dist(x = x)), class = "dist")
MMM <- (1-MMM)
GG  <- hclust(MMM, method="complete")
dend1 <- as.dendrogram(GG)
plot(dend1)
```

- **Dendrograma para o Coeficiente de Gower2**

```
MMM <- structure(c(0.349148624, 0.2, 0.816496581, 0.654653671, 0.489897949, 0.5, 0.654653671, 0.64, 0.2, 0.327326835, 0.534522484, 0.428571429, 0, 0.327326835, 0.428571429, 0.654653671, 0.763762616, 0.255155182, 0.327326835, 0.153093109, 0.4375, 0.327326835, 0.5, 0.25, 0.801783726, 0.375, 0.612372436, 0.801783726, 0.816496581, 0.40824829, 0.534522484, 0.763762616, 1, 0.654653671, 0.327326835, 0.40824829, 0.534522484, 0.244948974, 0, 0.763762616, 0.5, 0.25, 0.654653671, 0.327326835, 0.5), Size = as.integer(10), Labels = c("Barra", "Tijuca", "Ilha", "Centro", "Leblon", "Glória", "Urca", "Ipanema", "Saúde", "Catete"), Diag = FALSE, Upper = FALSE, method = "euclidean", call = quote(dist(x = x)), class = "dist")
MMM <- (1-MMM)
GG  <- hclust(MMM, method="complete")
dend1 <- as.dendrogram(GG)
plot(dend1)
```

APÊNDICE C

As rotinas contidas nesse apêndice atendem a um modelo de Análise de Agrupamento que objetiva agrupar alunos, a partir das variáveis que contabilizaram visitas às unidades didáticas (inclusive a UD0) e o tempo total gasto na plataforma Pii.

A Análise de Agrupamento utilizou valores relativizados (através da padronização estatística) para o cálculo da Distância Euclidiana Média entre os alunos, gerando a Matriz de Distâncias onde foi aplicada o Método da Ligação Completa (Complete Linkage) para geração dos grupos, os quais foram exportados para um arquivo texto já caracterizados através das médias das variáveis dos alunos existentes em cada grupo formado.

O programa funciona *stand alone*, sendo necessário apenas um simples *Form* contendo um botão para iniciar sua execução e um arquivo de banco de dados contendo as tabelas do capítulo 4.. O código, contendo as rotinas e sub-rotinas utilizadas, encontra-se a seguir[46].

CÓDIGO:

```
Public Conn As ADODB.Connection
Private VetAlunos() As Variant
Private VetAlunosNormalizados() As Variant
Private VetMedias() As Variant
Private VetDP() As Variant
Private VetDistancias() As Variant
Private VetGrupos() As Variant
Private QuantidadesUD As Integer
Private IndiceAlunoApurado As Integer
Private Distancias As Integer
Private IndiceGrupo As Integer


Public Sub conectar()
```

[46] Foi utilizado tamanho 10 para a fonte.

```
   'Cria o objeto RecordSet e atribui a variável
   Set Conn = New ADODB.Connection
   'Abre a conexão com o banco de dados
   Conn.Open "DBQ=C:\DIRETORIO\BANCODEDADOS.MDB;Driver={Microsoft Access Driver (*.mdb)}",
"USUÁRIOADMINISTRADOR", "SENHA"

End Sub
```

Sub desconectar()

```
   Conn.Close
   Set Conn = Nothing

End Sub
```

Private Sub Command1_Click()

```
Dim NumeroCurso As Integer
Dim TipoCenario As String

Conectar

'Abaixo são atribuídos o número do curso e o cenário

NumeroCurso = 3 'Tem mais de 10 alunos
TipoCenario = "Cenário A2"

'Rotina núcleo do programa, é onde acontece a coleta e apresentação dos dados para homologação
Call GeraVetor(NumeroCurso, TipoCenario)

'Tem que haver pelo menos dois alunos para aplicação de agrupamento
If IndiceAlunoApurado > 2 Then
   Call CalculaMedias(QuantidadesUD + 1)
   Call CalculaDP(QuantidadesUD + 1)
   Call Normaliza(QuantidadesUD + 1)
   Call CriaMatrizDistancias(IndiceAlunoApurado)
   Call QUICKSORT
   Call ConstruindoGruposLigacaoCompleta(Distancias)
   Call GerandoPerfis(IndiceGrupo)
End If
desconectar
```

End Sub

Public Function ExecutaQuickSort(VetDistanciasParaOrdenar As Variant, LimiteInferior As Long, LimiteSuperior As Long)

```
Dim LimiteInferiorTemp As Long
Dim LimiteSuperiorTemp As Long
Dim ElementoCentral As Long
Dim ValorTemp As Variant
Dim elementoAux As Variant

LimiteInferiorTemp = LimiteInferior
LimiteSuperiorTemp = LimiteSuperior
If LimiteSuperior <= LimiteInferior Then Exit Function
ElementoCentral = (LimiteInferior + LimiteSuperior) \ 2
ValorTemp = VetDistanciasParaOrdenar(ElementoCentral, 2)
Do While (LimiteInferiorTemp <= LimiteSuperiorTemp)
```

```
    Do While (VetDistanciasParaOrdenar(LimiteInferiorTemp, 2) < ValorTemp And LimiteInferiorTemp <
LimiteSuperior)
        LimiteInferiorTemp = LimiteInferiorTemp + 1
    Loop
    Do While (ValorTemp < VetDistanciasParaOrdenar(LimiteSuperiorTemp, 2) And LimiteSuperiorTemp >
LimiteInferior)
        LimiteSuperiorTemp = LimiteSuperiorTemp - 1
    Loop
    If (LimiteInferiorTemp <= LimiteSuperiorTemp) Then
      elementoAux = VetDistanciasParaOrdenar(LimiteInferiorTemp, 2)
      VetDistanciasParaOrdenar(LimiteInferiorTemp, 2) = VetDistanciasParaOrdenar(LimiteSuperiorTemp, 2)
      VetDistanciasParaOrdenar(LimiteSuperiorTemp, 2) = elementoAux
      elementoAux = VetDistanciasParaOrdenar(LimiteInferiorTemp, 0)
      VetDistanciasParaOrdenar(LimiteInferiorTemp, 0) = VetDistanciasParaOrdenar(LimiteSuperiorTemp, 0)
      VetDistanciasParaOrdenar(LimiteSuperiorTemp, 0) = elementoAux
      elementoAux = VetDistanciasParaOrdenar(LimiteInferiorTemp, 1)
      VetDistanciasParaOrdenar(LimiteInferiorTemp, 1) = VetDistanciasParaOrdenar(LimiteSuperiorTemp, 1)
      VetDistanciasParaOrdenar(LimiteSuperiorTemp, 1) = elementoAux
      LimiteInferiorTemp = LimiteInferiorTemp + 1
      LimiteSuperiorTemp = LimiteSuperiorTemp - 1
    End If
  Loop

  'Recursividade
  If (LimiteInferior < LimiteSuperiorTemp) Then
     ExecutaQuickSort VetDistanciasParaOrdenar, LimiteInferior, LimiteSuperiorTemp
  End If
  If (LimiteInferiorTemp < LimiteSuperior) Then
      ExecutaQuickSort VetDistanciasParaOrdenar, LimiteInferiorTemp, LimiteSuperior
  End If
```

End Function

Private Sub QUICKSORT()

```
'Essa procedure tem o objetivo de executar o algoritmo do QuickSort para ordenar o vetor VetDistancias
'através da sua terceira coluna, que é o valor da DEM. Feito isso, a montagem dos grupos será mais rápida
'pois teremos VetDistancias ordenado da menor para a maior distância entre os participantes

    Dim Indice As Integer
    Dim Exportar As String

    ExecutaQuickSort VetDistancias(), 0, (Distancias - 1)

    Open "C:\DIRETORIO\DistOrdenadas.txt" For Output As #1
    For Indice = 0 To (Distancias - 1)
       Print #1, VetDistancias(Indice, 0) & ", " & VetDistancias(Indice, 1) & ", " & VetDistancias(Indice, 2) &
Chr(13)
    Next
    Close #1
```

End Sub

Private Sub Normaliza(TotVarQuant As Integer)

```
    Dim i As Integer
    Dim j As Integer
    Dim Exportar As String
    ReDim VetAlunosNormalizados(IndiceAlunoApurado, TotVarQuant)

    For i = 0 To (IndiceAlunoApurado - 1)
       VetAlunosNormalizados(i, 0) = VetAlunos(i, 0)
```

```
    For j = 1 To TotVarQuant
      If VetDP(j - 1) <> 0 Then
        VetAlunosNormalizados(i, j) = ((VetAlunos(i, j) - VetMedias(j - 1)) / VetDP(j - 1))
      End If
    Next
  Next

  'Abre o aquivo Normalizadas.txt para ser utilizado na homologação dos cálculos
  Open "C:\DIRETORIO\normalizadas.txt" For Output As #1

  For i = 0 To (IndiceAlunoApurado - 1)
    Exportar = ""
    For j = 0 To TotVarQuant
      Exportar = Exportar & CStr(VetAlunosNormalizados(i, j)) & Chr(9)
    Next
    Exportar = Exportar & Chr(13)
    Print #1, Exportar
  Next
  Close #1

End Sub


Private Sub CalculaDP(TotVarQuant As Integer)

  Dim i As Integer
  Dim j As Integer

  'Cria o vetor de Desvios Padrões com o tamanho exato da quantidade de variáveis numéricas e o inicializa
  ReDim VetDP(TotVarQuant)
  For j = 0 To (TotVarQuant - 1)
    VetDP(j) = 0
  Next
  For j = 0 To (TotVarQuant - 1)
    For i = 0 To (IndiceAlunoApurado - 1)
      VetDP(j) = VetDP(j) + ((VetAlunos(i, j + 1) - VetMedias(j)) ^ 2)
    Next
    VetDP(j) = VetDP(j) / (IndiceAlunoApurado - 1)
    VetDP(j) = Sqr(Format(VetDP(j), "##0.000"))
  Next

  'Abre o aquivo DPs.txt para ser utilizado na homologação dos cálculos
  Open "C:\DIRETORIO\DPs.txt" For Output As #1
  Exportar = ""
  For j = 0 To (TotVarQuant - 1)
    If j <> (TotVarQuant - 1) Then 'Incluido apenas para evitar um tab adicional ao fim do registro
      Exportar = Exportar & CStr(VetDP(j)) & Chr(9)
    Else
      Exportar = Exportar & CStr(VetDP(j))
    End If
  Next
  Exportar = Exportar & Chr(13)
  Print #1, Exportar
  Close #1

End Sub

Private Sub CalculaMedias(TotVarQuant As Integer)

  Dim i As Integer
```

```
    Dim j As Integer

    Dim Exportar As String

    'Cria o vetor de médias com o tamanho exato da quantidade de variáveis numéricas
    ReDim VetMedias(TotVarQuant)
    For j = 0 To (TotVarQuant - 1)
       VetMedias(j) = 0
    Next

    For j = 0 To (TotVarQuant - 1)
       For i = 0 To (IndiceAlunoApurado - 1)
          VetMedias(j) = VetMedias(j) + VetAlunos(i, j + 1)
       Next
       VetMedias(j) = (VetMedias(j) / IndiceAlunoApurado)
       VetMedias(j) = Format(VetMedias(j), "##0.000")
    Next

    'Abre o aquivo Medias.txt para ser utilizado na homologação dos cálculos

    Open "C:\DIRETORIO\medias.txt" For Output As #1
    Exportar = ""
    For j = 0 To (TotVarQuant - 1)
       If j <> (TotVarQuant - 1) Then 'Incluido apenas para evitar um tab adicional ao fim do registro
          Exportar = Exportar & CStr(VetMedias(j)) & Chr(9)
       Else
          Exportar = Exportar & CStr(VetMedias(j))
       End If
    Next
    Exportar = Exportar & Chr(13)
    Print #1, Exportar
    Close #1
```

End Sub

Private Sub Homologacao(Indice As Integer, Campos As Integer)

```
Dim Exportar As String

Open "C:\DIRETORIO\dados.txt" For Output As #1

'Exporta para o arquivo dados.txt todos os registros e seus respectivos campos
For i = 0 To Indice
   Exportar = ""
   For j = 0 To Campos
      If j <> (Campos) Then 'Incluido apenas para evitar um tab adicional ao fim do registro
         Exportar = Exportar & CStr(VetAlunos(i, j)) & Chr(9)
      Else
         Exportar = Exportar & CStr(VetAlunos(i, j))
      End If
   Next
   Exportar = Exportar & Chr(13)
   VetAlunos(i, 4); Tab; VetAlunos(i, 5); Tab; VetAlunos(i, 6); Tab; VetAlunos(i, 7);
   Print #1, Exportar
Next
Close #1
```

End Sub

Private Sub HomologacaoUDs(UDs As Variant, Indice As Integer)

```
Dim Exportar As String

Open "C:\DIRETORIO\UDs.txt" For Output As #1
'Exporta para o arquivo UDs.txt todas as UDs encontradas
For i = 0 To Indice
   Exportar = "UD" & CStr(UDs(i, 1)) & Chr(13)
   Print #1, Exportar
Next
Close #1
```

End Sub

Private Sub ConstruindoGruposLigacaoCompleta(TotDistancias As Integer)

```
Dim i As Integer
Dim Campos As Integer
Dim IndiceOriginal As Integer
Dim Um, Dois As String
Dim IndicePrimeiro As Integer

'O vetor VetGrupos tem quatro dimensoes, a primeira contendo o nível representado
'pela distância que foi utilizada para a formação daquele grupo, a segunda dimensão
'contendo o primeiro componente (que pode ser outro grupo) do grupo que acaba de ser
'formado, a terceira dimensão contendo o outro componente, e a quarta contendo todos
'os dois componentes juntos separados por virgula. As outras dimensões, representadas por
'(QuantidadesUD + 1), serão utilizadas somente na rotina CriandoPerfis

ReDim VetGrupos(TotDistancias - 1, 4 + (QuantidadesUD + 2))
'Como VetDistancias está ordenada, a menor distância é o seu primeiro elemento
'e consequentemente esse é o primeiro grupo a ser formado
IndiceGrupo = 0
IndiceOriginal = 0
VetGrupos(IndiceGrupo, 0) = VetDistancias(IndiceOriginal, 2) 'Valor da Distancia
VetGrupos(IndiceGrupo, 1) = VetDistancias(IndiceOriginal, 0) 'Primeiro elemento
VetGrupos(IndiceGrupo, 2) = VetDistancias(IndiceOriginal, 1) 'Segundo elemento
VetGrupos(IndiceGrupo, 3) = VetDistancias(IndiceOriginal, 0) & ";" & VetDistancias(IndiceOriginal, 1)

'Preenche VetDistanciasTemp a partir de VetDistancias já retirando o primeiro grupo formado
TotDistanciasTemp = TotDistancias - 2

ReDim VetDistanciasTemp(TotDistancias - 2, 3)

For i = 0 To TotDistanciasTemp
   VetDistanciasTemp(i, 0) = VetDistancias(i + 1, 0) 'Primeiro participante
   VetDistanciasTemp(i, 1) = VetDistancias(i + 1, 1) 'Segundo participante
   VetDistanciasTemp(i, 2) = VetDistancias(i + 1, 2) 'DEM
Next
TotDistanciasTemp = TotDistanciasTemp + 1
Um = VetGrupos(IndiceGrupo, 1)
Dois = VetGrupos(IndiceGrupo, 2)
Do While TotDistanciasTemp > 1
   cont = 0
   Do While cont < TotDistanciasTemp
      If (VetDistanciasTemp(cont, 0) = Um Xor VetDistanciasTemp(cont, 1) = Dois) Or
(VetDistanciasTemp(cont, 0) = Dois Xor VetDistanciasTemp(cont, 1) = Um) Then
      Select Case VetDistanciasTemp(cont, 0)
         Case Um
            'Apura quem é o outro participante
            Outro = VetDistanciasTemp(cont, 1)
```

```
        'Armazena a distância entre eles
        DistanciaDaPrimeiraDupla = VetDistanciasTemp(cont, 2)
        VetDistanciasTemp(cont, 0) = VetGrupos(IndiceGrupo, 3)

        'Busca um outro par semelhante
        For Busca = (cont + 1) To (TotDistanciasTemp - 1)
          'Achou um outro par
          If VetDistanciasTemp(Busca, 0) = Dois And VetDistanciasTemp(Busca, 1) = Outro Then
            If VetDistanciasTemp(Busca, 2) > DistanciaDaPrimeiraDupla Then 'Método Complete Linkage
                VetDistanciasTemp(cont, 2) = VetDistanciasTemp(Busca, 2)
            End If
            'Zera o conteúdo da outra dupla pois o mesmo não é mais necessário
            VetDistanciasTemp(Busca, 0) = ""
            VetDistanciasTemp(Busca, 1) = ""
            VetDistanciasTemp(Busca, 2) = ""
          End If
          If VetDistanciasTemp(Busca, 0) = Outro And VetDistanciasTemp(Busca, 1) = Dois Then
            If VetDistanciasTemp(Busca, 2) > DistanciaDaPrimeiraDupla Then 'Método Complete Linkage
                VetDistanciasTemp(cont, 2) = VetDistanciasTemp(Busca, 2)
            End If
            'Zera o conteúdo da outra dupla pois o mesmo não é mais necessário
            VetDistanciasTemp(Busca, 0) = ""
            VetDistanciasTemp(Busca, 1) = ""
            VetDistanciasTemp(Busca, 2) = ""
          End If
        Next
      Case Dois
        'Apura quem é o outro participante
        Outro = VetDistanciasTemp(cont, 1)
        'Armazena a distância entre eles
        DistanciaDaPrimeiraDupla = VetDistanciasTemp(cont, 2)
        VetDistanciasTemp(cont, 0) = VetGrupos(IndiceGrupo, 3)

        'Busca um outro par semelhante
        For Busca = (cont + 1) To (TotDistanciasTemp - 1)

          'Achou um outro par
          If VetDistanciasTemp(Busca, 0) = Um And VetDistanciasTemp(Busca, 1) = Outro Then
            If VetDistanciasTemp(Busca, 2) > DistanciaDaPrimeiraDupla Then 'Método Complete Linkage
                VetDistanciasTemp(cont, 2) = VetDistanciasTemp(Busca, 2)
            End If
            'Zera o conteúdo da outra dupla pois o mesmo não é mais necessário
            VetDistanciasTemp(Busca, 0) = ""
            VetDistanciasTemp(Busca, 1) = ""
            VetDistanciasTemp(Busca, 2) = ""
          End If

          If VetDistanciasTemp(Busca, 0) = Outro And VetDistanciasTemp(Busca, 1) = Um Then
            If VetDistanciasTemp(Busca, 2) > DistanciaDaPrimeiraDupla Then 'Método Complete Linkage
                VetDistanciasTemp(cont, 2) = VetDistanciasTemp(Busca, 2)
            End If
            'Zera o conteúdo da outra dupla pois o mesmo não é mais necessário
            VetDistanciasTemp(Busca, 0) = ""
            VetDistanciasTemp(Busca, 1) = ""
            VetDistanciasTemp(Busca, 2) = ""
          End If
        Next
    End Select

    Select Case VetDistanciasTemp(cont, 1)
```

```
Case Um
   'Apura quem é o outro participante
   Outro = VetDistanciasTemp(cont, 0)
   'Armazena a distância entre eles
   DistanciaDaPrimeiraDupla = VetDistanciasTemp(cont, 2)
   VetDistanciasTemp(cont, 1) = VetGrupos(IndiceGrupo, 3)

   'Busca um outro par semelhante
   For Busca = (cont + 1) To (TotDistanciasTemp - 1)
      'Achou um outro par
      If VetDistanciasTemp(Busca, 0) = Dois And VetDistanciasTemp(Busca, 1) = Outro Then
         If VetDistanciasTemp(Busca, 2) > DistanciaDaPrimeiraDupla Then 'Método Complete Linkage
               VetDistanciasTemp(cont, 2) = VetDistanciasTemp(Busca, 2)
         End If
         'Zera o conteúdo da outra dupla pois o mesmo não é mais necessário
         VetDistanciasTemp(Busca, 0) = ""
         VetDistanciasTemp(Busca, 1) = ""
         VetDistanciasTemp(Busca, 2) = ""
      End If
      If VetDistanciasTemp(Busca, 0) = Outro And VetDistanciasTemp(Busca, 1) = Dois Then
         If VetDistanciasTemp(Busca, 2) > DistanciaDaPrimeiraDupla Then 'Método Complete Linkage
               VetDistanciasTemp(cont, 2) = VetDistanciasTemp(Busca, 2)
         End If
         'Zera o conteúdo da outra dupla pois o mesmo não é mais necessário
         VetDistanciasTemp(Busca, 0) = ""
         VetDistanciasTemp(Busca, 1) = ""
         VetDistanciasTemp(Busca, 2) = ""
      End If
   Next
Case Dois
   'Apura quem é o outro participante
   Outro = VetDistanciasTemp(cont, 0)
   'Armazena a distância entre eles
   DistanciaDaPrimeiraDupla = VetDistanciasTemp(cont, 2)
   VetDistanciasTemp(cont, 1) = VetGrupos(IndiceGrupo, 3)

   'Busca um outro par semelhante
   For Busca = (cont + 1) To (TotDistanciasTemp - 1)
      'Achou um outro par
      If VetDistanciasTemp(Busca, 0) = Um And VetDistanciasTemp(Busca, 1) = Outro Then
         If VetDistanciasTemp(Busca, 2) > DistanciaDaPrimeiraDupla Then 'Método Complete Linkage
               VetDistanciasTemp(cont, 2) = VetDistanciasTemp(Busca, 2)
         End If
         'Zera o conteúdo da outra dupla pois o mesmo não é mais necessário
         VetDistanciasTemp(Busca, 0) = ""
         VetDistanciasTemp(Busca, 1) = ""
         VetDistanciasTemp(Busca, 2) = ""

      End If
      If VetDistanciasTemp(Busca, 0) = Outro And VetDistanciasTemp(Busca, 1) = Um Then

         If VetDistanciasTemp(Busca, 2) > DistanciaDaPrimeiraDupla Then 'Método Complete Linkage
               VetDistanciasTemp(cont, 2) = VetDistanciasTemp(Busca, 2)
         End If
         'Zera o conteúdo da outra dupla pois o mesmo não é mais necessário
         VetDistanciasTemp(Busca, 0) = ""
         VetDistanciasTemp(Busca, 1) = ""
         VetDistanciasTemp(Busca, 2) = ""
      End If
   Next
```

```
        End Select
        'Tem que localizar a outra distancia
        End If
        cont = cont + 1
    Loop
    'O Do While abaixo limpa VetDistancias dos elementos vazios
    Dim LinhasEmBranco As Integer
    LinhasEmBranco = 0
    cont = 0
    Do While cont < (TotDistanciasTemp - LinhasEmBranco - 1)
        If VetDistanciasTemp(cont, 0) = "" Then
            For ContAux = (cont + 1) To (TotDistanciasTemp - LinhasEmBranco - 1)
                VetDistanciasTemp(ContAux - 1, 0) = VetDistanciasTemp(ContAux, 0)
                VetDistanciasTemp(ContAux - 1, 1) = VetDistanciasTemp(ContAux, 1)
                VetDistanciasTemp(ContAux - 1, 2) = VetDistanciasTemp(ContAux, 2)
            Next
            VetDistanciasTemp(TotDistanciasTemp - LinhasEmBranco - 1, 0) = ""
            VetDistanciasTemp(TotDistanciasTemp - LinhasEmBranco - 1, 1) = ""
            VetDistanciasTemp(TotDistanciasTemp - LinhasEmBranco - 1, 2) = ""
            LinhasEmBranco = LinhasEmBranco + 1
        End If
        'Garante a eficiência do algoritmo mesmo se o próximo elemento que substituiu o atual
        'estiver vazio
        If VetDistanciasTemp(cont, 0) <> "" Then cont = cont + 1

    Loop
    TotDistanciasTemp = TotDistanciasTemp - LinhasEmBranco

    'Exportando para homologar
    Open "C:\DIRETORIO\DistSegundoEmDiante.txt" For Output As #1
    For IndicePrimeiro = 0 To (TotDistanciasTemp - 1)
        Print #1, VetDistanciasTemp(IndicePrimeiro, 0) & ";" & VetDistanciasTemp(IndicePrimeiro, 1) & "----> "
& VetDistanciasTemp(IndicePrimeiro, 2) & Chr(13)
    Next
    Close #1

    IndiceGrupo = IndiceGrupo + 1
    VetGrupos(IndiceGrupo, 0) = VetDistanciasTemp(IndiceOriginal, 2) 'Distância entre os dois elememtos
    VetGrupos(IndiceGrupo, 1) = VetDistanciasTemp(IndiceOriginal, 0)
    VetGrupos(IndiceGrupo, 2) = VetDistanciasTemp(IndiceOriginal, 1)
    VetGrupos(IndiceGrupo, 3) = VetGrupos(IndiceGrupo, 1) & ";" & VetGrupos(IndiceGrupo, 2)

    'Guarda os dois componentes do novo grupo
    Um = VetGrupos(IndiceGrupo, 1)
    Dois = VetGrupos(IndiceGrupo, 2)

    'Atualiza o vetor VetDistanciasTemp tirando a primeira posição que acaba de se tornar um grupo
    For ContAux = 1 To TotDistanciasTemp
        VetDistanciasTemp(ContAux - 1, 0) = VetDistanciasTemp(ContAux, 0)
        VetDistanciasTemp(ContAux - 1, 1) = VetDistanciasTemp(ContAux, 1)
        VetDistanciasTemp(ContAux - 1, 2) = VetDistanciasTemp(ContAux, 2)

    Next
    VetDistanciasTemp(TotDistanciasTemp, 0) = ""
    VetDistanciasTemp(TotDistanciasTemp, 1) = ""
    VetDistanciasTemp(TotDistanciasTemp, 2) = ""
    'Atualiza o tamanho do vetor
    TotDistanciasTemp = TotDistanciasTemp - 1
Loop
```

```
Open "C:\DIRETORIO\GruposFinaisLigCompleta.txt" For Output As #1
For i = 0 To (IndiceGrupo)
   Print #1, VetGrupos(i, 0) & " ------> " & "( " & VetGrupos(i, 3) & " )" & Chr(13)
Next
Close #1
```

End Sub

Private Sub CriaMatrizDistancias(TotDistancias As Integer)

```
   Dim i As Integer
   Dim j As Integer
   Dim Campos As Integer
   Dim TotalCelulas As Integer
   Dim DEM As Double

   'O total de células da matriz é dado por n(n-1)/2
   TotalCelulas = (TotDistancias * (TotDistancias - 1)) / 2

   ReDim VetDistancias(TotalCelulas, 3)
   'VetDistancias armazena os nomes dos alunos, a distancia DEM
   'entre eles, o total de alunos, e a soma das frequências nas UDs e a soma dos tempos
   'dos dois participantes que produziram aquela distância

   Distancias = 0
   For i = 0 To (IndiceAlunoApurado - 2)
      For j = (i + 1) To (IndiceAlunoApurado - 1)
         VetDistancias(Distancias, 0) = VetAlunos(i, 0) 'Retem o nome de um dos dois alunos
         VetDistancias(Distancias, 1) = VetAlunos(j, 0) 'Retem o nome do outro dos dois alunos
         'Calcula e atribui ao vetor a DEM - Distância Euclidiana Media
         DEM = 0
         For Campos = 1 To (QuantidadesUD + 1)
            DEM = DEM + ((VetAlunosNormalizados(i, Campos) - VetAlunosNormalizados(j, Campos)) ^ 2)
         Next
         DEM = Sqr(DEM / (QuantidadesUD + 1))
         VetDistancias(Distancias, 2) = Format(DEM, "##0.000")
         Distancias = Distancias + 1
      Next
   Next
   Open "C:\DIRETORIO\Distancias.txt" For Output As #1
   For i = 0 To (Distancias - 1)
      Print #1, VetDistancias(i, 0) & ";" & VetDistancias(i, 1) & ";" & VetDistancias(i, 2) & Chr(13)
   Next
   Close #1
```

End Sub

Private Function GeraVetor(sCurso As Integer, Cenario As String)

```
Dim SQL As String
Dim strSQL As String
Dim RS As ADODB.Recordset
Dim rsTemp As ADODB.Recordset
Dim RS2 As ADODB.Recordset
Dim DataAnterior As String
Dim DataAtual As String
Dim AlunoLog As String
Dim Vet() As Variant
Dim HoraAnterior As Date
```

```
Dim VetTemp() As Variant
Dim IndiceVetTemp As Integer
Dim i As Integer
Dim j As Integer
Dim IndiceAluno As Integer
Dim iTotal As Integer
Dim MaxParticipantes As Integer
Dim TempoAtual As Date
Dim TempoDecorrido As Date
Dim TempoPraMeiaNoite As Date
Dim PrimeiroRegistro As Boolean
Dim HorasViradas
Dim UDAnterior
Dim cont, contAlunos As Integer
Dim AlunosIniciais As Integer
Dim VetAlunosParticipantes() As Variant

SQL = "Select Log_pii.Aluno, tabalunos.Tipousuario From Log_pii INNER JOIN tabalunos ON Log_pii.Aluno
="
SQL = SQL & " tabalunos.E_Mail where Log_pii.Aluno <> " And Log_pii.Curso = " & sCurso & " And "
SQL = SQL & " tabalunos.Curso = " & sCurso & " group by Log_pii.aluno, tabalunos.Tipousuario"

Set RS = Conn.Execute(SQL)

MaxParticipantes = 100
ReDim VetAlunosParticipantes(MaxParticipantes, 2)

IndiceAluno = 0
AlunosIniciais = 0
RS.MoveFirst
'O Do abaixo monitora os alunos identificados inicialmente no Log, os registros deles ainda serão criticados
Open "C:\DIRETORIO\AlunosIniciais.txt" For Output As #1
Do While Not RS.EOF
   Print #1, RS!aluno & "---->" & RS!tipousuario & Chr(13)
   VetAlunosParticipantes(AlunosIniciais, 0) = RS!aluno
   VetAlunosParticipantes(AlunosIniciais, 1) = RS!tipousuario
   AlunosIniciais = AlunosIniciais + 1
   RS.MoveNext
Loop
RS.MoveFirst
Close #1

'O código abaixo recupera a quantidade de unidades didáticas existentes no curso
strSQL = "SELECT UDidatica FROM Log_pii where Curso = " & sCurso & " Group By UDidatica Order By
UDidatica"
Set rsTemp = Conn.Execute(strSQL)
rsTemp.MoveFirst
QuantidadesUD = 0
Do While Not rsTemp.EOF
   If rsTemp("UDidatica") <> "" And CInt(rsTemp("UDidatica")) < 11 Then QuantidadesUD = QuantidadesUD
+ 1
   rsTemp.MoveNext
Loop

ReDim VetTemp(QuantidadesUD, 2) 'Esse vetor guarda apenas a quantidade de vezes que cada UD foi visitada,
guardando o índice da UD
'O loop abaixo serve para pegar o índice de cada UD, pois existem cursos com UDs: UD1, UD2, UD4, UD10,...
não são sequenciais
rsTemp.MoveFirst
IndiceVetTemp = 0
```

```
Do While Not rsTemp.EOF
   If rsTemp("UDidatica") <> "" Then
   If CInt(rsTemp("UDidatica")) < 11 Then 'Por definição, as unidades didáticas não passam de 10
      VetTemp(IndiceVetTemp, 1) = rsTemp("UDidatica")
      IndiceVetTemp = IndiceVetTemp + 1
   End If
   rsTemp.MoveNext
   End If
Loop
rsTemp.Close

'Exporta as UDs encontradas para um arquivo texto
Call HomologacaoUDs(VetTemp, IndiceVetTemp - 1)

'Cria o vetor VET com tamanho máximo de alunos iniciais lidos e com dimensões referentes a
'QuantidadesUD + 2 pois além das UDs o vetor terá o e-mail do participante e o total de horas
'que ele totalizou de uso na plataforma
ReDim Vet(AlunosIniciais, (QuantidadesUD + 2))

Do While Not RS.EOF 'Esse Do percorre todos os participantes
   'De acordo com o cenário escolhido, utiliza todos ou parte dos participantes e exclui ou não UD0
   Select Case Cenario
   Case "Cenário A2" 'Faz a AA com todas as UDs, inclusive UD0, excluindo somente os professores
      For contAlunos = 0 To (AlunosIniciais - 1)
         If VetAlunosParticipantes(contAlunos, 0) = RS!aluno Then
            'Se o participante não for aluno, não participará da AA
            If CInt(VetAlunosParticipantes(contAlunos, 1)) <> 4 Then GoTo ApontaParaOutro
         End If
      Next
   End Select

   SQL = "Select aluno, udidatica, data, hora, log From log_pii "
   SQL = SQL & "where aluno = '" & RS!aluno & "' And Curso = " & sCurso & " And udidatica < 11 order by
data, hora"

   Set RS2 = Conn.Execute(SQL)
   RS2.MoveFirst
   For i = 0 To (QuantidadesUD - 1)
      VetTemp(i, 0) = 0 'Limpando vetor temporário
   Next

   Do While Not RS2.EOF 'Esse Do percorre todos registros de um participante específico
      DataAtual = Year(RS2!Data) & Format(Month(RS2!Data), "00") & Format(Day(RS2!Data), "00")
      If DataAtual = DataAnterior Or DataAnterior = "" Then
         'TempoAtual e TempoPraMeiaNoite destinam-se a capturar o tempo daquele participante
         'que tendo entrado em um dia, à noite, saiu no outro dia, logo após a meia-noite
         TempoAtual = RS2!hora
         If HoraAnterior Then
            TempoDecorrido = TempoDecorrido + (TempoAtual - HoraAnterior)
            Else
               TempoDecorrido = 0
         End If
         TempoPraMeiaNoite = #11:59:59 PM# - RS2!hora
         If RS2!Log = 0 And DataAnterior <> "" Then
         'Se no 1o. registro o participante estiver saindo, não entra na AA
            Vet(IndiceAluno, 0) = RS2!aluno
            If UDAnterior = "" And CInt(RS2!udidatica) < 11 Then
               'Quando o participante ficou apenas em uma UD
               For i = 0 To (QuantidadesUD - 1)
                  If VetTemp(i, 1) = RS2!udidatica Then Exit For 'Achou qual UD para atualizar
```

```
            Next
            VetTemp(i, 0) = VetTemp(i, 0) + 1
          End If
          For i = 1 To QuantidadesUD
            Vet(IndiceAluno, i) = Vet(IndiceAluno, i) + VetTemp(i - 1, 0) 'Alimentando o vetor fixo a partir do
temporário
          Next

          'O indice é Quantidades + 1 para pegar a última posição do vetor q corresponde ao TTAH do
participante
          'Vet(IndiceAluno, (QuantidadesUD + 1)) = Vet(IndiceAluno, (QuantidadesUD + 1)) +
TempoDecorrido
          'Já coloca o tempo decorrido em horas
          Vet(IndiceAluno, (QuantidadesUD + 1)) = Vet(IndiceAluno, (QuantidadesUD + 1)) +
(Second(TempoDecorrido) + (Minute(TempoDecorrido) * 60) + (Hour(TempoDecorrido) * 3600)) / 3600
          Vet(IndiceAluno, (QuantidadesUD + 1)) = Format(Vet(IndiceAluno, (QuantidadesUD + 1)),
"##0.000")
          For i = 0 To (QuantidadesUD - 1)
            VetTemp(i, 0) = 0 'Limpando vetor temporário
          Next
          TempoDecorrido = 0 'Limpando o contador de tempo
        Else 'Se log for diferente de 0, significa que o partic. nao saiu ainda da Pii, exceto se for o primeiro
registro
          'Percorre o vetor das UDs para descobrir qual UD será atualizada, se a nova UD é válida (< 11)
          If RS2!Log <> 0 And UDAnterior <> RS2!udidatica And CInt(RS2!udidatica) < 11 Then
            For i = 0 To (QuantidadesUD - 1)
              If VetTemp(i, 1) = RS2!udidatica Then
                UDAnterior = RS2!udidatica
                Exit For 'Achou
              End If
            Next
            VetTemp(i, 0) = VetTemp(i, 0) + 1
          End If
        End If
      Else 'Se data for diferente, o aluno pode ter entrado na Pii em um dia e saído no dia posterior
        If RS2!Log = 0 Then 'O participante saiu da Pii no dia seguinte
          Vet(IndiceAluno, 0) = RS2!aluno
          If UDAnterior = "" And CInt(RS2!udidatica) < 11 Then 'Quando o participante ficou apenas em uma
UD
            For i = 0 To (QuantidadesUD - 1)
              If VetTemp(i, 1) = RS2!udidatica Then Exit For 'Achou qual UD para atualizar
            Next
            VetTemp(i, 0) = VetTemp(i, 0) + 1
          End If
          For i = 1 To QuantidadesUD
            Vet(IndiceAluno, i) = VetTemp(i - 1, 0) 'Alimentando o vetor fixo a partir do temporário
          Next
          HorasViradas = (Second(TempoDecorrido) + (Minute(TempoDecorrido) * 60) +
(Hour(TempoDecorrido) * 3600)) / 3600
          HorasViradas = HorasViradas + (Second(TempoPraMeiaNoite) + (Minute(TempoPraMeiaNoite) * 60)
+ (Hour(TempoPraMeiaNoite) * 3600)) / 3600
          HorasViradas = HorasViradas + (Second(RS2!hora) + (Minute(RS2!hora) * 60) + (Hour(RS2!hora) *
3600)) / 3600

          Vet(IndiceAluno, (QuantidadesUD + 1)) = Vet(IndiceAluno, (QuantidadesUD + 1)) + HorasViradas
          TempoDecorrido = 0 'Limpando o contador de tempo
        End If 'Descarto o registro anterior

        For i = 0 To (QuantidadesUD - 1)
          VetTemp(i, 0) = 0 'Limpando vetor temporário
```

```
            Next
            TempoAtual = 0 'Limpando o contador de tempo
        End If
        If RS2!Log <> 0 Then
            DataAnterior = DataAtual
            HoraAnterior = RS2!hora
        Else
            DataAnterior = ""
            HoraAnterior = 0
            UDAnterior = ""
        End If
        RS2.MoveNext 'Próxima data/hora do participante específico
    Loop
    IndiceAluno = IndiceAluno + 1

ApontaParaOutro:
    RS.MoveNext 'Próximo aluno
Loop
'IndiceAluno guarda o total de participantes que tiveram seus registros analisados,
'porém alguns campos referentes ao nome encontram-se sem o e-mail do participante
'por isso se faz necessário retirá-los através do For abaixo.
IndiceAlunoApurado = 0
For i = 0 To IndiceAluno
    If Vet(i, 0) <> "" Then
        IndiceAlunoApurado = IndiceAlunoApurado + 1
    End If
Next
ReDim VetAlunos(IndiceAlunoApurado, QuantidadesUD + 2)
'IndiceAlunoApurado fornece o total correto que participarão da AA
'O vetor VetAlunos guardará os dados que subsidiarão as técnicas de AA
cont = 0
For i = 0 To IndiceAluno 'Alimentando o vetor sem espaços em branco
    If Vet(i, 0) <> "" Then
        For j = 0 To (QuantidadesUD + 1)
            VetAlunos(cont, j) = Vet(i, j)
        Next
        cont = cont + 1
    End If
Next
'Homologacao exporta para um arquivo TXT os vetores para que seja feita a homologação dos cálculos
posteriores no Excel
Call Homologacao(IndiceAlunoApurado - 1, QuantidadesUD + 1)
```

End Function

Private Sub GerandoPerfis(TotGrupos As Integer)

```
'O objetivo dessa sub é ler os elementos que formaram cada grupo existente em VetGrupos
'com o intuito de, a partir da identificação do nome(e-mail) do elemento, recuperar a partir do vetor
'VetAlunos, as frequências do mesmo, somando esses valores com os valores dos outros integrantes
'desse grupo. Após fechar esse cálculo dentro de um grupo, fornecer ao professor que solicitou
'a Análise de Agrupamento as médias das frequências dos n participantes do grupo em relação às UDs
'e também o tempo médio dos participantes na plataforma Pii.

Dim TodosComponentes As String
Dim ContInterno, ContExterno As Integer
Dim Tamanho, Posicao As Integer
Dim Elemento As String
Dim NumElementos As Integer
Dim PosicaoElementoVetAlunos As Integer
```

```
For ContadorExterno = 0 To TotGrupos
   NumElementos = 0
   TodosComponentes = VetGrupos(ContadorExterno, 3)
   Tamanho = Len(TodosComponentes)
   Posicao = InStr(1, TodosComponentes, ";") 'Acha a posição do separador dos elementos que formam o grupo
   Elemento = Left(TodosComponentes, Posicao - 1)
   PosicaoElementoVetAlunos = RecuperaPosicaoAluno(Elemento)
   'If PosicaoElementoVetAlunos or PosicaoElementoVetAlunos = 0 Then
      For ContadorInterno = 4 To (QuantidadesUD + 4)
         VetGrupos(ContadorExterno, ContadorInterno) = VetGrupos(ContadorExterno, ContadorInterno) +
VetAlunos(PosicaoElementoVetAlunos, ContadorInterno - 3)
      Next
   'End If
   RestanteElementos = Right(TodosComponentes, Tamanho - Posicao)
   NumElementos = NumElementos + 1
   Do While RestanteElementos <> ""
      TodosComponentes = RestanteElementos
      Tamanho = Len(TodosComponentes)
      Posicao = InStr(1, TodosComponentes, ";")
      If Posicao <> 0 Then
         Elemento = Left(TodosComponentes, Posicao - 1)
         RestanteElementos = Right(TodosComponentes, Tamanho - Posicao)
      Else
         Elemento = TodosComponentes
         RestanteElementos = ""
      End If
      PosicaoElementoVetAlunos = RecuperaPosicaoAluno(Elemento)
      If PosicaoElementoVetAlunos Then
         For ContadorInterno = 4 To (QuantidadesUD + 4)
            VetGrupos(ContadorExterno, ContadorInterno) = CDbl(VetGrupos(ContadorExterno,
ContadorInterno)) + CDbl(VetAlunos(PosicaoElementoVetAlunos, ContadorInterno - 3))
         Next
      End If
      NumElementos = NumElementos + 1
   Loop
   'A última dimensão do grupo armazena a quantidade de elementos existentes nele
   VetGrupos(ContadorExterno, ContadorInterno) = NumElementos

Next

Open "C:\DIRETORIO\GruposFinaisComPerfis.txt" For Output As #1
For i = 0 To (IndiceGrupo)
   Print #1, VetGrupos(i, 0) & " ------> " & "( " & VetGrupos(i, 3) & " )" & Chr(13)
   Exporta = "Médias: "
   For ContadorInterno = 4 To (QuantidadesUD + 4)
      Exporta = Exporta & CStr(Format((VetGrupos(i, ContadorInterno) / VetGrupos(i, QuantidadesUD + 5)),
"##0.0")) & Chr(9)
   Next
   Print #1, Exporta & Chr(13)
Next
Close #1
```

End Sub

```
'Essa função retorna o índice de VetAlunos onde se encontra Nome
```

Public Function RecuperaPosicaoAluno(Nome) As Integer

```
Dim Posicao As Integer
```

```
For Posicao = 0 To IndiceAlunoApurado
   If VetAlunos(Posicao, 0) = Nome Then Exit For
Next
RecuperaPosicaoAluno = Posicao

End Function
```