**Pontifícia Universidade Católica do Rio Grande do Sul**
**Faculdade de Informática**
**Programa de Pós-Graduação em Ciência da Computação**

# Métodos de Clusterização para Apoio à Classificação Estética de Documentos

Tiago Thompsen Primo

Porto Alegre
2008

**Pontifícia Universidade Católica do Rio Grande do Sul**
**Faculdade de Informática**
**Programa de Pós-Graduação em Ciência da Computação**

# Métodos de Clusterização para Apoio à Classificação Estética de Documentos

Tiago Thompsen Primo

**Dissertação apresentada como requisito parcial à obtenção do grau de mestre em Ciência da Computação**

Orientador: Prof. Dr. João Batista Souza de Oliveira

Porto Alegre
2008

Pontifícia Universidade Católica do Rio Grande do Sul

## Dados Internacionais de Catalogação na Publicação (CIP)

P953m Primo, Tiago Thompsen
Métodos de clusterização para apoio à classificação estética de documentos / Tiago Thompsen Primo. – Porto Alegre, 2008.
112 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS
Orientador: Prof. Dr. João Batista Souza de Oliveira

1. Informática. 2. Algoritmos. 3. Agrupamento de Informações (Informática). 4. Documentos – Estética (Informática). I. Oliveira, João Batista Souza de. II. Título.

CDD 005.1

**Ficha Catalográfica elaborada pelo**
**Setor de Tratamento da Informação da BC-PUCRS**

PUCRS
**Campus Central**
Av. Ipiranga, 6681 - prédio 16 - CEP 90619-900
Porto Alegre - RS - Brasil
Fone: +55 (51) 3320-3544 - Fax: +55 (51) 3320-3548
Email: bceadm@pucrs.br
www.pucrs.br/biblioteca

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada **"Métodos de Clusterização para Apoio à Classificação Estética de Documentos"**, apresentada por Tiago Thompsen Primo, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Sistemas Interativos de Visualização, aprovada em 24/03/08 pela Comissão Examinadora:

Prof. Dr. João Batista Souza de Oliveira – Orientador — PPGCC/PUCRS

Profa. Dra. Vera Lúcia Strube de Lima — PPGCC/PUCRS

Profa. Dra. Rosa Maria Viccari — UFRGS

Homologada em 16/09/08, conforme Ata No. 19/08 pela Comissão Coordenadora.

Prof. Dr. Fernando Gehm Moraes
Coordenador.

**Campus Central**
Av. Ipiranga, 6681 – P32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@inf.pucrs.br
www.pucrs.br/facin/pos

*Dedico este trabalho a Deus e a minha família.*

# Agradecimentos

Agradeço primeiramente as forças positivas do universo que trouxeram paz e serenidade em diversos momentos difíceis durante o desenvolvimento deste trabalho. Graças a essas vibrações de campos superiores meu espírito manteve-se balanceado e focado.

A minha namorada Kelen Bernardi, que suportou momentos difíceis aos quais passei durante a elaboração deste trabalho, mostrando paciência e compreensão inexistentes em grande parte dos seres humanos. Meus pais, Wilson e Carmen Primo, que sempre se mostraram dispostos a me ajudar, me apoiar e não permitir que eu esqueça de meu potencial. Minhas vitórias são conseqüência de seus esforços em me tornar um ser humano íntegro e lutador.

Meu orientador João Batista Souza de Oliveira, que me deu a oportunidade e o suporte necessários para a realização e conclusão deste trabalho.

A Pontifícia Universidade Católica do Rio Grande do Sul e ao Centro de Pesquisa PUCRS/HP, que possibilitaram este período de estudos.

A todos os meus amigos que comigo participaram desta jornada de estudos árdua e enriquecedora, com especial menção aos caras que mais próximos estive durante este período, grandes amigos, Tiago Silva, Thyago Borges e Luis Souza. Não é todo o dia que se encontra pessoas especiais como estas.

# Resumo

Neste trabalho serão abordados estudos referentes à classificação de grande quantidade de documentos de conteúdo variável. Em tal processo quando um grande número de documentos é gerado, existe a necessidade de um usuário verificá-los um a um com a intenção de separá-los em bons (com pouco ou nenhum problema estrutural) ou ruins (que possuem problemas estruturais), processo este considerado lento e oneroso. Considerando este problema, neste trabalho foi desenvolvida uma ferramenta de classificação estética de documentos que visa reduzir esta intervenção humana.

A ferramenta desenvolvida é baseada em métricas que avaliam o quanto um documento automaticamente gerado difere de seu template, criando para cada um destes documentos uma assinatura baseada nas técnicas de *fingerprint*, objetivando primeiramente distingui-los entre si para então utilizar técnicas de clusterização criando grupos de documentos com características semelhantes.

O algoritmo K-Medóides é usado para fazer tal agrupamento, tal algoritmo funciona criando grupos de objetos considerando um destes como base para a criação de cada cluster. A idéia deste trabalho é reduzir a intervenção humana fazendo com que um usuário classifique em bom ou ruim apenas determinados documentos de cada grupo formado pelo algoritmo de clusterização.

São também apresentados resultados de quatro experimentos realizados com esta ferramenta avaliando as contribuições para diminuir a intervenção humana no processo de classificação de documentos.

Palavras-chave: Clusterização, Medidas Estéticas, Classificação Estética.

# Abstract

This work presents a study about classifying a large amount of variable data printing documents. At such process when a large number of documents is generated there is the need for a user to classify then one by one with the goal to separate then in good (those with none or few structural problem) or bad (those with several structural problems), a process that is considered very slow. Considering this problem, in this work we build a tool for aesthetical document classification, winch has the goal to reduce such human intervention.

The developed tool is based on metrics that determines how different are the documents that was automatic generated to their respective template, creating for each one of those documents a signature based in the fingerprinting techniques. After that, clustering techniques are used to create groups which the documents that have similar characteristics.

The K-Medoids algorithm is used to create those groups. Such algorithm works creating groups of objects considering one of then as a base for each created cluster. The main idea is to reduce the human intervention by asking for a user to classify in good or bad specific documents on each cluster.

It is also presented the results of four experiments that was realized with this tool, evaluating the contributions to reduce the human intervention in the document classification process.

Keywords: Clustering, Aesthetical Measures, Aesthetical Classification.

# Lista de Figuras

# Lista de Tabelas

# Lista de Siglas

| | |
|---|---|
| **CNC** | Cloudmark Network Classifier |
| **SR** | Sistema de Reputação |
| **DE** | Distância Euclidiana |
| **SGBD** | Sistema de Gerênciamento de Banco de Dados |

# Sumário

# 1 Introdução

## 1.1 Motivação

Uma forma de comunicação que tem crescido é aquela através de documentos adaptativos ou documentos de conteúdo variável. Um documento de conteúdo variável pode ser composto por conteúdos originários de uma base de dados, obtendo uma personalização através de informações que se alterarão conforme o perfil de cada cliente, aplicando ao documento diferentes mensagens, produtos, textos e figuras.

Conforme (Arts, 2006) 37% das empresas de artistas gráficos no ano de 2006 produziram algum tipo de documento de conteúdo variável, 9% a mais que no mesmo período do ano de 2005. Os mesmos autores ainda mencionam que entre os artistas gráficos que trabalham com geração de documentos, 16% disseram que passaram a utilizar "um pouco mais" (cerca de 1 a 25%) ou "muito mais" (25% ou mais) dados variáveis em seus trabalhos, dentre os editores a porcentagem que considera a comunicação sobre forma de documentos de conteúdo variável como grande oportunidade de venda aumentou de 2%, no segundo trimestre de 2002, para 8% em 2006.

Estes tipos de documentos são principalmente utilizados na área de *marketing* personalizado, visando atrair novos clientes através de propagandas baseadas em uma análise dos costumes e hábitos de uma pessoa ou grupo de pessoas em questão, usando conteúdos personalizados.

Tradicionalmente, uma pessoa que trabalha na criação e/ou geração de documentos, possui domínio sobre o material gerado, já que cria individualmente cada documento a ser distribuído. Nos documentos de conteúdo variável a geração de um documento também parte inicialmente de uma pessoa, mas esta resume o seu papel à criação de um modelo (template) que envolve atributos como tamanhos de letras, posições das imagens e escolha de cores, entre outros, que influenciam na aparência e organização de um documento, mas a interação humana no restante do processo de criação fica restrita ao resultado final, ou seja, ao conjunto de documentos gerados automaticamente.

O processo de criação automática produz incerteza quanto à qualidade dos documentos gerados, já que, o usuário perde o controle do conteúdo gerado, fazendo que este precise verificar um a um se os documentos foram corretamente gerados, processo considerado lento e oneroso.

Considerando o custo associado a esse processo e percebendo que tais tipos de documento parecem ser uma tendência, neste trabalho está sendo proposta uma ferramenta de classificação estética de documentos, que apoiada no uso de técnicas de clusterização tem o objetivo de di-

minuir a intervenção humana no processo de classificação estética visando um ganho de tempo no processo de geração e distribuição de documentos de conteúdo variável.
Para este trabalho considera-se intervenção humana quando documentos são apresentados a um usuário, sendo solicitado a este, que os classifique em bons (possuem poucos ou nenhum problema em sua geração) ou ruins (com problemas em sua geração). A idéia é buscar reduzir esse número de intervenções feitas a um usuário durante o processo de classificação estética.
O desenvolvimento desta ferramenta não busca uma automatização completa do processo, nem um sistema que não necessite intervenção de um usuário. A idéia é que o sistema funcione solicitando o menor número de requisições possíveis a um usuário (semi-automático).
O desenvolvimento desta ferramenta parte do fato de que existem métricas para aferir a qualidade estética de um documento, mas um processo que os separe em bons (com pouco ou nenhum problema em sua geração) e ruins (com erros significativos em sua criação) não foi encontrado dentre a bibliografia pesquisada sobre o assunto.

## 1.2 Objetivos

O objetivo principal deste trabalho é a redução da intervenção humana no processo de classificação estética de documentos. Esta intervenção habitualmente é realizada quando a um usuário é apresentado a um documento gerado automaticamente e este informa se o considera bom ou ruim.
Com informações obtidas através de um usuário, busca-se uma forma de agrupar documentos semelhantes à aqueles que foram classificados visando eliminar uma intervenção posterior de tais documentos, teoricamente reduzindo o número de intervenções humanas.
Para que isto seja possível, o primeiro passo foi buscar uma maneira de representar cada documento individualmente, criando uma espécie de assinatura. Com esta assinatura a tendência é que aqueles documentos com características semelhantes também devem ter assinaturas semelhantes. Imaginando cada documento representado por um ponto, acredita-se que existirão regiões com maiores concentrações de pontos seguidas de regiões com baixas concentrações de pontos. Em um segundo passo foi buscada uma maneira de se delimitar essas regiões de grandes concentrações de pontos de forma a se identificar grupos de documentos semelhantes.
A junção destes passos culmina no desenvolvimento de uma ferramenta de classificação estética de documentos que possibilitará através de intervenções de um usuário a documentos específicos classificar os grupos formados em bons ou ruins. A ferramenta deve lidar com a possibilidade de existirem duas classificações (bom ou ruim), mas não necessariamente dois grupos de documentos, já que podem existir diversos grupos de documentos com diferentes classificações.
A funcionalidade da ferramenta desenvolvida é testada sobre um conjunto de documentos de

entrada com diferentes características, visando avaliar o comportamento do método de agrupamento de documentos utilizado bem como se a ferramenta cumpre com o seu objetivo principal.

## 1.3 Estado da Arte

Para o desenvolvimento deste trabalho buscou-se encontrar em áreas adjacentes técnicas capazes de resolver o problema de classificar documentos esteticamente.
Este método de pesquisa se deu devido ao fato da área de Engenharia de Documentos ser relativamente recente, tendo seu principal congresso *DocEng* (DocEng, 2001) datado de sua primeira versão no ano de 2001 ocasionando dificuldades na busca de materiais sobre o tema proposto.
Dentre os estudos realizados não foi encontrada alguma ferramenta capas de realizar um processo de classificação estética de documentos. Por outro lado, (Balinsky & Pilu, 2005), (Faria & Oliveira, 2006) e (Harrington et al., 2004) apresentam métricas de classificação estética que avaliam o quanto um documento gerado a partir de um template difere do mesmo. Outros autores como (da Silva et al., 2005) abordam questões relativas ao posicionamento de objetos em documentos de conteúdo variável, possibilitando conteúdos que podem ser inseridos neste tipo de documento apresentem maior flexibilidade de formato. O que tais autores não abordam seria uma maneira de classificar estes documentos quanto a sua qualidade estética, ou até mesmo uma maneira de identificar tais documentos de forma a possibilitar a criação de uma identidade para cada documento que possibilite agrupamentos de documentos com características semelhantes, visando uma redução da intervenção humana no processo atual de classificação estética.
Para buscar uma identificação de cada documento que possibilite um agrupamento daqueles com características semelhantes, e posteriormente solicitar que um usuário os classifique (por exemplo, em um grupo com 500 documentos, se um usuário classificar 50 e os demais assumirem as classificações daqueles semelhantes aos classificados), foram encontradas dentro da área de detecção de *SPAM* as técnicas de *fingerprint*.
O princípio básico de tais técnicas é criar uma identificação para algum objeto. No caso dos autores (Perone, 2004) e (Prakash & O'Donnell, 2005), o *fingerprint* é criado para cada email recebido por um usuário, permitindo que cada qual possua uma identificação própria.
Em uma próxima etapa estes *fingerprints* são comparados com outros presentes em uma base de dados e que já são previamente classificados. De acordo com o resultado é determinado se um email é ou não *SPAM*.
Neste trabalho tais técnicas serão utilizadas para identificar os documentos e distingui-los entre si, criando assinaturas e visando o agrupamento destas assinaturas para um posterior processo de classificação estética.
As técnicas de clusterização estudadas por autores como por exemplo: (Han & Kamber, 2000), (Everitt, 1993) e (Jain et al., 1999), visam criar grupos semelhantes, e seu uso neste trabalho

teria o mesmo propósito.
Dentre os métodos de clusterização existentes na bibliografia, os métodos baseados em particionamento ganham ênfase devido a sua popularidade. Tais métodos caracterizam-se em dividir os objetos a serem agrupados em $K$ grupos distintos, sendo este valor $K$ informado por um usuário. Característica esta, não muito vantajosa para o tema proposto neste trabalho, já que torna-se complexo informar um valor $K$ inicial pela imprevisibilidade que os documentos gerados automaticamente possuem.
Por serem métodos largamente utilizados na bibliografia sobre clusterização, foi desenvolvida neste trabalho uma forma a contornar este problema referente à necessidade de informar um valor $K$ inicial, possibilitando o uso de tais métodos.

## 1.4 Organização do Texto

Esta dissertação está dividida nos seguintes capítulos:
**Capítulo 2:** Neste capítulo será apresentada uma visão geral sobre a classificação estética de documentos, culminando na apresentação de técnicas capazes de medir o quanto um documento gerado de forma automática difere de seu documento modelo/template.
Tais métricas, quando associadas a um documento, podem ser usadas como forma de distinguir os documentos gerados entre si.
**Capítulo 3:** Neste capítulo será apresentada a revisão bibliográfica utilizada para o desenvolvimento da ferramenta de classificação estética de documentos proposta neste trabalho. É apresentado um estudo sobre as técnicas de *fingerprint*, provenientes da área de detecção de *SPAM*. Sua principal característica é possibilitar a criação de uma assinatura de um e-mail. Estas assinaturas são utilizadas para comparações que determinam se um e-mail é ou não considerado *SPAM*. Será apresentado também como tais métricas foram utilizadas de forma a criar uma assinatura para cada documento automaticamente gerado.
No decorrer são apresentados estudos referentes aos métodos de clusterização. Tais métodos em sua essência criam grupos (clusters) de objetos com características semelhantes.
**Capítulo 4:** Neste capítulo será apresentado o processo de classificação estética proposto neste trabalho, assim como o algoritmo de clusterização escolhido para trabalhar com a ferramenta de classificação desenvolvida, que em conjunto visam reduzir a intervenção humana no processo de classificação de documentos.
**Capítulo 5:** Neste capítulo serão apresentados os conjuntos de documentos utilizados nos experimentos com a ferramenta de classificação e a forma como foram conduzidos tais experimentos de forma a verificar se o objetivo de reduzir a intervenção humana foi atingido.
**Capítulo 6:** Neste capítulo serão apresentados quatro experimentos realizados para verificar o funcionamento da ferramenta de classificação proposta.

**Capítulo 7:** Finalizando, neste capítulo serão apresentadas as conclusões obtidas com este trabalho bem como trabalhos futuros.

# 2 Medidas estéticas para documentos

Este capítulo oferece uma visão geral sobre a classificação estética de documentos, culminando na apresentação de técnicas capazes de medir o quanto um documento gerado de forma automática difere de seu documento modelo/template.
Tais métricas, quando associadas a um documento, podem ser usadas como forma de distinguir os documentos gerados entre si, esta distinção será abordada em maiores detalhes nos capítulos posteriores.

## 2.1 Documentos de Conteúdo Variável

Derivados da área de *Variable Data Printing* (VDP) ou impressão de dados variáveis ((BRASIL, 2005a),(BRASIL, 2005b),(BRASIL, 2005c), (Purvis et al., 2003)), os documentos de conteúdo variável tem o objetivo, através de informações personalizadas (nome do cliente, cores preferidas, formas de tratamento, entre outras) de atrair a atenção de um determinado público-alvo.
Documentos de conteúdo variável podem ser vistos no dia-a-dia, através de malas diretas, folders, e-mails, entre outros, que tenham o objetivo de fazer com que uma pessoa que recebe este tipo de material tenha a sensação de que seu remetente o conhece e valoriza. Desta forma, é primordial que um documento de conteúdo variável cause um impacto positivo em seus destinatários.
A criação de tais documentos parte inicialmente de um template, geralmente criado por um *designer* que determina a maneira como as informações serão dispostas. As conseqüências seriam vários documentos automaticamente gerados possuindo específicas variações sobre o template proposto, e é importante que cada documento gerado a partir de um template se mantenha o mais semelhante possível a este.
(Balinsky & Pilu, 2005) mencionam que é vital manter a semelhança entre um documento automaticamente gerado com o modelo proposto. (Faria & Oliveira, 2006) mencionam que evitar que documentos com graus de qualidade estética ruins sejam impressos é de importância em campanhas de marketing direto, pois a imagem da empresa que está veiculando a mensagem de está diretamente relacionada à qualidade do material impresso.
Um exemplo de como pode ser este *template* pode ser visualizado na Figura 1.

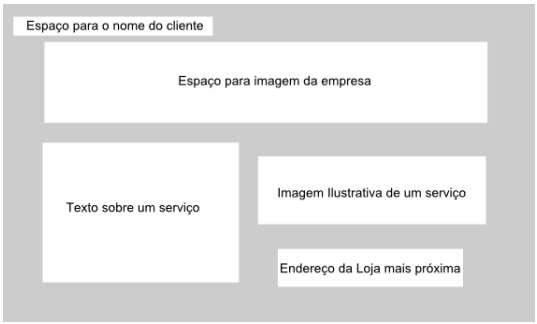


Figura 1 – Exemplo de um *template* de um documento de conteúdo variável

Como forma de buscar garantias no processo de geração automática de documentos, (Faria & Oliveira, 2006) e (Harrington et al., 2004) propõem diversas métricas estéticas que auxiliam a identificar problemas de design que podem ser encontrados em documentos gerados automaticamente. Tais métricas funcionam fazendo uma comparação entre um *template* e variações geradas automaticamente a partir deste. A aplicação destas métricas podem ser usadas como forma de identificação para cada documento, e posteriormente um agrupamento de documentos com características semelhantes visando um processo de classificação.

## 2.2 Criação de um documento de conteúdo variável

Na impressão de dados variáveis as tecnologias que despontam como alternativas para criação de tais tipos de documentos seriam XSL-FO (W3C, 2007) e PPML (*Personalized Print Markup Language*) (DeBronkart & Davis, 2000).
O padrão PPML foi definido pela PODI (*Printing on Demand Initiative*) (PODI, 2007) de forma a prover meios de disponibilizar dados variáveis em documentos que necessitem alta qualidade em sua geração. Baseado em XML, é considerado um padrão de simples uso para a criação de documentos de conteúdo variável.
O padrão XSL-FO, também oriundo do XML, tem a função principal de prover meios onde dados XML possam ser formatados para apresentação em diversas mídias.
A diferença básica entre os dois reside na forma como um documento receberá seus dados variáveis. No padrão PPML são reservados espaços para conteúdos variáveis, onde caso um conteúdo os extrapole, este conteúdo excedente é eliminado do documento gerado. Por exemplo, caso um texto de três parágrafos seja inserido em um espaço onde cabem apenas dois, o último parágrafo é eliminado do resultado final. No padrão XSL-FO, os dados não são cortados. Considerando o exemplo anterior, o parágrafo excedente acabará extrapolando a área que era reservada podendo aparecer sobre os demais elementos pertencentes a um documento.

Neste trabalho, não serão abordados detalhes mais específicos sobre a criação de documentos usando as tecnologias citadas. Autores como (Meneguzzi et al., 2004), (da Silva et al., 2005) e (Giannetti et al., 2006) apresentam trabalhos mais específicos sobre isto.
Um exemplo do processo de criação pode ser visto na Figura 2, onde as regras de negócio (por exemplo, usuários com idade superior a 25 anos, sexo feminino, entre outras), base de informações (locais de onde serão extraídas as informações das regras de negócio) e um *template* (documento modelo, que informa onde os dados variáveis serão inseridos) são base para uma ferramenta de criação de documentos de conteúdo variável, que em seu resultado pode criar diversos tipos de media, como por exemplo, *emails*, *websites*, malas-diretas entre outros.

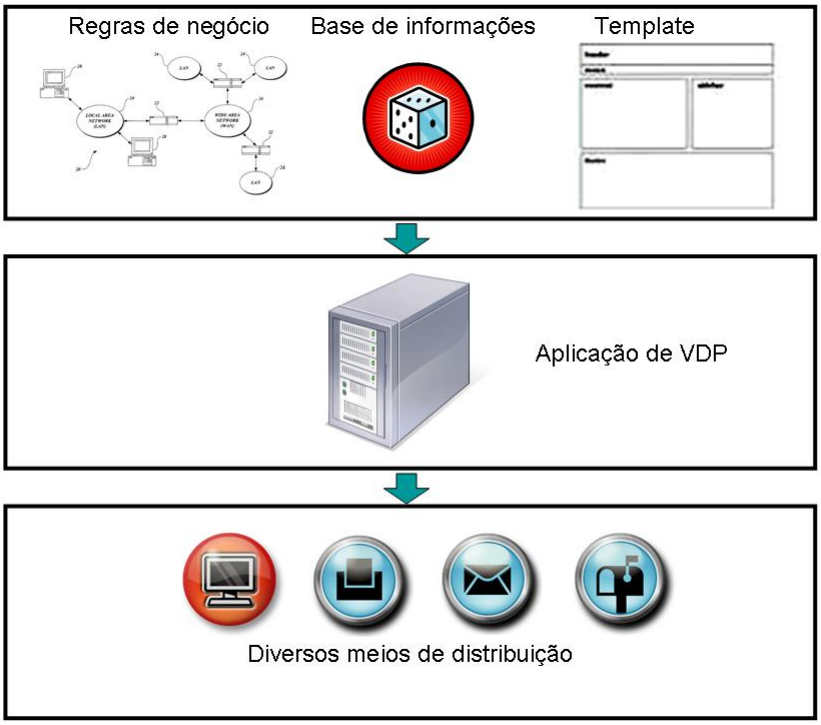


Figura 2 – Exemplo de um *workflow* de criação de documentos de conteúdo variável

## 2.3 Métricas de Qualidade Visual

As métricas de qualidade visual são uma forma de avaliar quantitativamente (através de uma nota) o quanto um documento gerado difere de seu *template*.
Esta medição serve como um primeiro passo para se distinguir um conjunto de documentos gerados automaticamente. Cabe ressaltar que cada documento pode ter uma ou mais notas associadas.
Estas notas podem ser individuais, quando se leva em conta cada métrica aplicada, ou gerais, quando as notas de cada métrica aplicada a um documento em específico são unidas em uma

única nota para representar a avaliação de um documento. (Harrington et al., 2004) menciona cuidados ao se fornecer uma nota geral a um documento, já que, uma única métrica que apresente um resultado ruim pode vir a prejudicar erroneamente a avaliação geral de um documento. Como neste trabalho a idéia é agrupar documentos com características semelhantes, está sendo considerado que cada documento é composto por um conjunto de notas provenientes da aplicação de métricas estéticas. A figura 3, ilustra um exemplo da aplicação de quatro métricas a um documento automaticamente gerado. Ressalta-se, que tais métricas foram aleatoriamente escolhidas e as notas apresentadas são fictícias, bem como, para este trabalho está será a metodologia de aplicação de métricas utilizada.

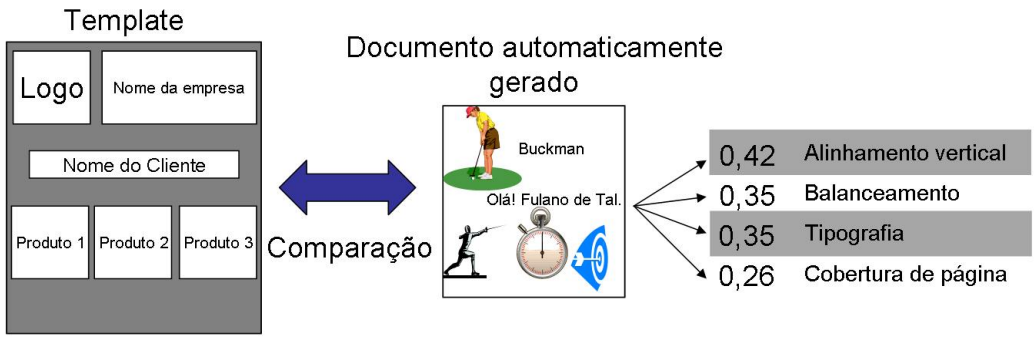


Figura 3 – Aplicando métricas estéticas a um documento automaticamente gerado

Alguns exemplos de métricas utilizadas para medirem falhas estruturais de um documento serão apresentadas nas subseções seguintes. Tais métricas são descritas a partir de (Faria & Oliveira, 2006) e (Harrington et al., 2004)

### 2.3.1 Balanceamento

Esta métrica mensura a distribuição uniforme dos elementos em cada página do documento, podendo ser aplicável a figuras, blocos de texto, tabelas e demais elementos. Podem ser feitos dois tipos de balancementos:

- Balanceamento ao centro
- Balanceamento esquerda-direita

O balanceamento ao centro mede a distância entre os elementos de um documento em relação ao seu centro. O objetivo é de comparar os resultados obtidos em um template com os resultados

obtidos por um documento automaticamente gerado (Figura 4). Em teoria, aqueles documentos automaticamente gerados com valores muito diferentes dos obtidos pelo seu respectivo *template* seriam documentos mal formados. O balanceamento esquerda-direita difere apenas no ponto de comparação, não considerando mais o centro de um documento, mas sim suas bordas esquerda e direita (Figura 5).

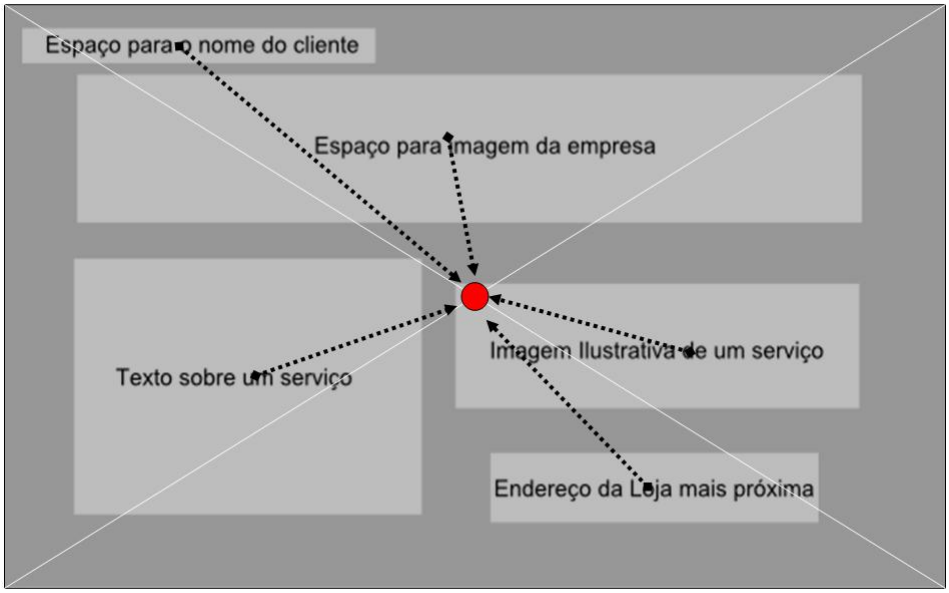


Figura 4 – Balanceamento ao centro

### 2.3.2 Proporção entre largura e altura

Esta medida propõe que os blocos de conteúdo de um documento devem respeitar a regra da proporção entre a altura e a largura na razão aproximada de 1,618. Este número, também conhecido como *golden ratio* (Weisstein, 2007), é adotado por artistas e arquitetos desde o período do **renascimento** como uma tentativa de mensurar a beleza de forma quantitativa, através de relações de ordem e simetria (Figura 6).

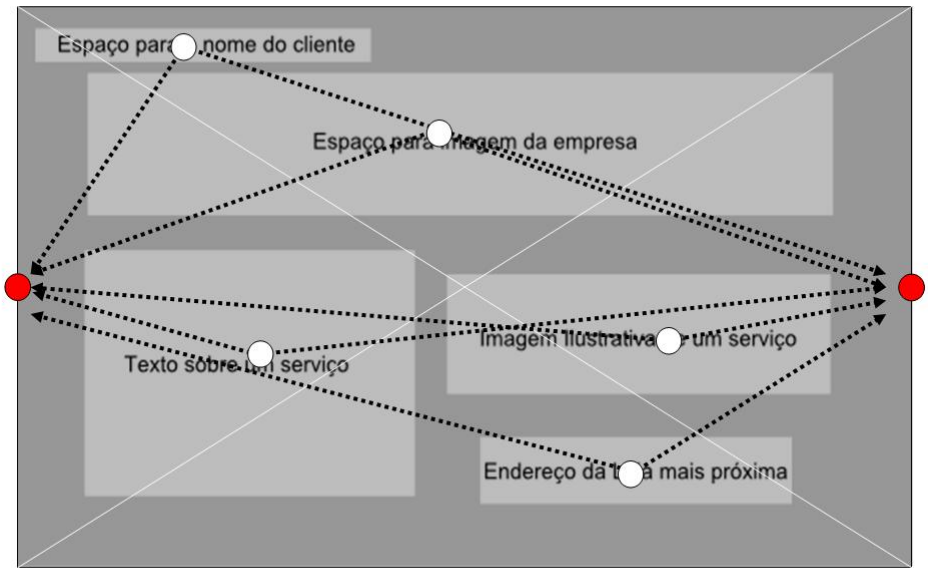


Figura 5 – Balanceamento esquerda-direita

### 2.3.3 Cobertura de Página

Esta métrica mede a quantidade de área em branco de cada página de um documento em relação às demais outras cores. Segundo (Faria & Oliveira, 2006) uma fração ideal de cobertura fica em torno de 50% da página. A expressão para o cálculo desse valor é a seguinte:

$$V = 1 - (|\sum A_i/A_p - 0.5| * 2) \tag{2.1}$$

Nesta expressão a área de uma página é representada por $A_p$ e $A_i$ representa a área de cada elemento básico $i$ de uma página.

### 2.3.4 Tipografia

Esta métrica diz que o tamanho e as famílias das fontes devem ser respeitados, e sua modificação penalizada. Para definir as penalidades uma tabela que contenha as relações de fontes precisa ser criada, definindo penalidades relativas às trocas entre tipos de fonte. A Tabela 1 apresenta um exemplo de como poderiam ser tais penalidades.

Esta tabela serve como referência para indicar o grau de penalidade que um documento sofre quando, por exemplo, seu texto passa da fonte Verdana para a Comic Sans.

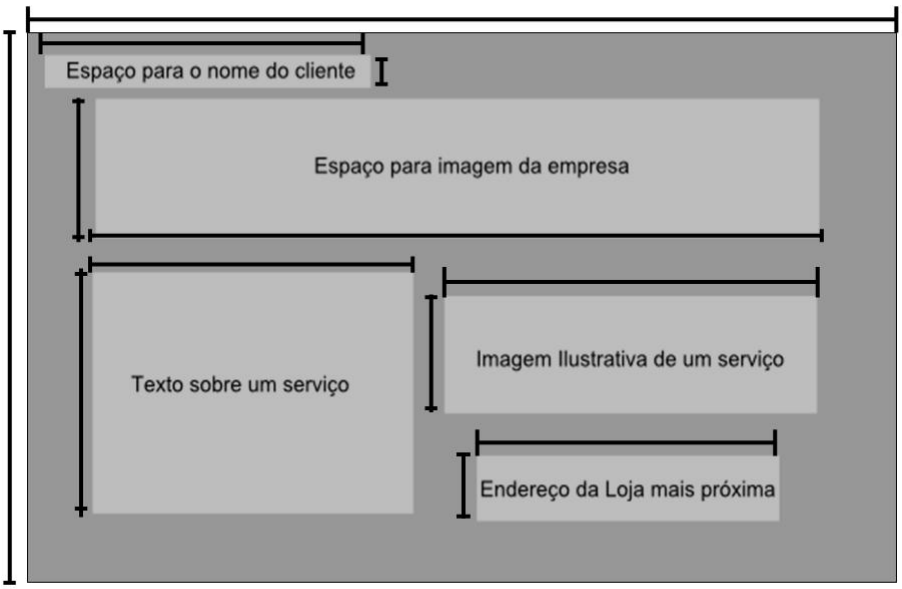


Figura 6 – Proporção entre a largura X altura

Tabela 1 – Tabelas de Penalidades e Fontes

| Fonte | Verdana | Comic Sans | Courier New | Arial |
|---|---|---|---|---|
| Verdana | 1 | 0.6 | 0.45 | 0.8 |
| Comic Sans | 0.6 | 1 | 0.7 | 0.85 |
| Courier New | 0.45 | 0.7 | 1 | 0.75 |
| Arial | 0.8 | 0.85 | 0.75 | 1 |

### 2.3.5 Imagens

Número de imagens por página

O número de imagens de um documento gerado não deve ser diferente do padrão estabelecido em seu *template*, tais mudanças acarretam em uma penalização no documento que apresenta tal problema.

Cores das figuras e do fundo

Esta métrica prevê que as cores de um documento devem ter harmonia com as cores de fundo.
Esse padrão é determinado pelo designer, que cria uma tabela (similar ao método visto na se-

ção 2.3.4), ou forma de representação que determine o grau de penalidade que a mudança de cores origina em um documento automaticamente gerado.

Número de cores utilizadas

O uso de uma paleta de cores pode abrir um leque de opções a serem comparadas e cabe a um usuário delimitar o número de cores que considere útil, ou agradável para o seu documento. O não cumprimento desta métrica acarreta penalizações no documento gerado.

Cor dominante

A cor dominante pode ser descrita como a cor tema de um documento e predomina sobre as demais. Esta cor pode ser definida a cada página. Se considerarmos que cada elemento básico de um documento possue apenas uma cor, podemos definir uma métrica para calcular este escore em função da área dos elementos básicos.

### 2.3.6 Existência

Esta métrica diz que remover um elemento de um documento tem penalidades definidas. Um exemplo de exclusão, que poderia ser aceita, seria a remoção de uma imagem para que um documento automaticamente gerado tenha o seu tamanho compatível com o de um dispositivo móvel.

### 2.3.7 Posição Horizontal e Vertical

Tem o objetivo de medir a diferença entre a posição horizontal dos elementos em cada documento gerado em comparação a seu respectivo *template*. A posição vertical tem o mesmo principio básico, a diferença está no fato que a análise é feita de acordo com a posição vertical ( Figura 7 ).

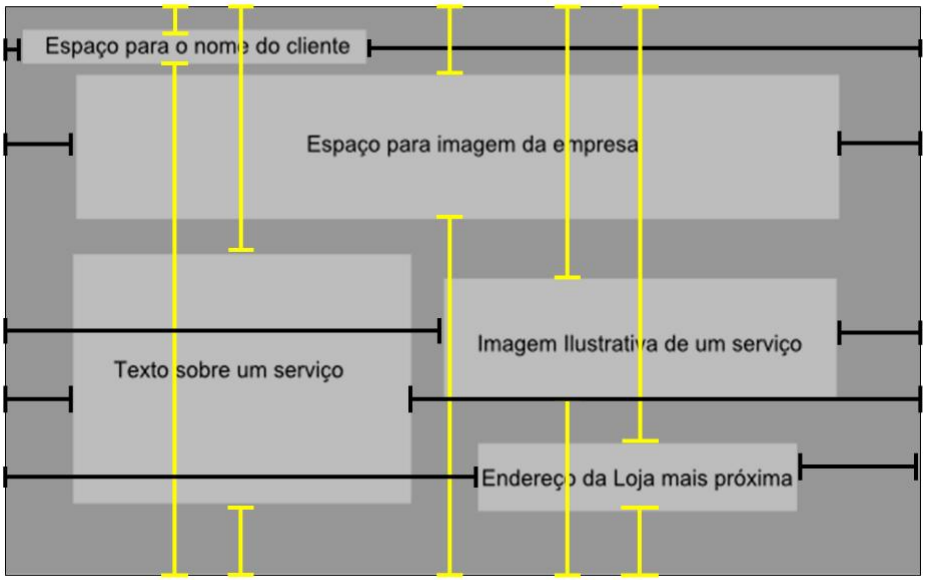


Figura 7 – Posição horizontal e vertical dos objetos de um documento

### 2.3.8 Distância entre vizinhos

Esta métrica prevê uma penalidade relativa ao aumento de distância entre os elementos de um documento gerado quando comparado ao seu referido *template*. (Faria & Oliveira, 2006) propõem uma expressão para o cálculo dessa métrica:

$$V = 1 - (|h_{orig} - h_{inst}|/h_p) \tag{2.2}$$

A altura da página é representada por $h_p$, onde $h_{orig}$ é a altura do envelope ( que neste trabalho refere-se ao retângulo de menor tamanho possível que envolva dois elementos de um documento a serem comparados) original (do *template*) e $h_{inst}$ representa a altura do envelope na instância.

## 2.4 Considerações do capítulo

O objetivo deste capítulo foi trazer uma idéia sobre formas de medir a qualidade de um documento automaticamente gerado.
Tais documentos quando gerados em larga escala escapam ao controle de um usuário, que passa a encontrar dificuldades em verificar um a um os documentos em busca de erros.

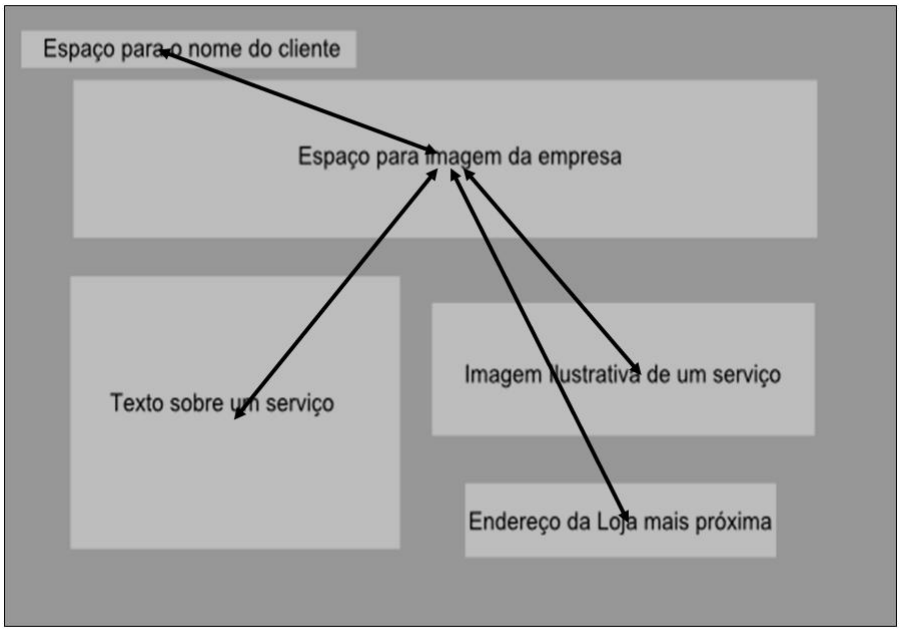


Figura 8 – Distância entre os vizinhos

Como forma de facilitar tal processo, algumas das métricas que auxiliam a medir a qualidade estética foram descritas, assim como a forma que estas podem ser associadas a um documento, seja individualmente (escolhida para este trabalho) ou usando uma única nota para representar o documento em si.
Tais métricas servirão para criar a identidade de um documento, uma forma de distinguir os documentos entre si, visando automatizar o processo de classificação estética que será apresentado no decorrer do texto.

# 3 Clusterização

Até este momento foram apresentadas métricas capazes de avaliar o quanto um documento gerado automaticamente difere do seu respectivo template.
Neste capítulo será apresentada uma revisão bibliográfica utilizada como base para o desenvolvimento de uma ferramenta de classificação estética de documentos ainda a ser proposta.
O capítulo começa expondo um estudo sobre as técnicas de *fingerprint*, provenientes da área de detecção de *SPAM*. Sua principal característica é possibilitar a criação de uma assinatura de um e-mail. Estas assinaturas são utilizadas para comparações que determinam se um e-mail é ou não considerado *SPAM*. Será apresentado como tais métricas foram utilizadas de forma a criar uma assinatura para cada documento automaticamente gerado.
No decorrer são apresentados estudos referentes aos métodos de clusterização. Tais métodos em sua essência criam grupos (*clusters*) de objetos com características semelhantes. Devido a esta característica, este trabalho visa agrupar assinaturas semelhantes e através de intervenções específicas visarem à redução da intervenção humana no processo de classificação estética de documentos.

## 3.1 Fingerprint

### 3.1.1 Técnica de Fingerprint

No presente contexto, as técnicas de *Fingerprint* foram principalmente exploradas na identificação de e-mails como *SPAM* ou não.
O trabalho de (Perone, 2004) menciona o uso das técnicas de *Fingerprint* como forma de criar uma identificação ou assinatura para cada e-mail. De posse desta assinatura é feita uma comparação com as assinaturas de outros e-mails considerados *SPAM*, conforme os resultados, o e-mail recebido passa a ser considerado como *SPAM* ou não.
Para ser criada a assinatura de um e-mail, este passa por um processo de filtragem onde são aplicadas diversas métricas, como por exemplo:

- Aparecimento de determinadas palavras;

- Aparecimento de remetentes específicos;
- Quantidade de imagens e outras;

O resultado da aplicação destas métricas compõe a assinatura de um e-mail.

(Cloudmark, 2007), faz uso de ténicas de *fingerprint* em seus produtos. Um exemplo seria o *Cloudmark Network Classifier*, CNC, uma espécie de servidor de *fingerprints* provenientes de e-mails classificados por usuários de suas ferramentas anti-spam.
(Prakash & O'Donnell, 2005) faz o uso do CNC em seu trabalho, criando uma ferramenta de controle de *SPAM* baseada em Sistemas de Reputação (SR) . SR's funcionam basicamente considerando as avaliações de objetos por usuários de um determinado domínio, as avaliações comuns a grande parte dos usuários podem refletir de forma mais correta a avaliação de algum objeto.
SR's também procuram identificar usuários que de alguma forma queiram prejudicar a qualidade de algum sistema em questão, buscando desconsiderar suas avaliações. Detalhes mais específicos sobre SR's não serão abordados, e podem ser observados no trabalho de (Resnick et al., 2000) ou (Masum & Zhang, 2004).
O SR utilizado no trabalho de (Prakash & O'Donnell, 2005), procura identificar dentre os *fingerprints* classificados por usuários, aqueles que possam ou não *SPAM*.
(Prakash & O'Donnell, 2005) ainda mencionam que o *fingerprint* de e-mails pode ser criado sobre dois pontos de vista:
A abordagem *unique*, considera que cada e-mail tem uma assinatura individual, enquanto a abordagem *oracle* procura juntar e-mails com características semelhantes para então criar uma assinatura para cada grupo de e-mails.
A Figura 9 mostra como seria a criação de *fingerprints* considerando as duas abordagens mencionadas.

A abordagem *oracle* seria útil se um processo de classificação já determinou quais seriam os grupos de documentos bons e ruins, então se poderia comparar cada novo *fingerprint* com o *fingerprint* destes grupos, associando-o a estes.
Como o escopo deste trabalho considera cada documento individualmente, foi escolhida a abordagem *unique*.

## 3.2 *Fingerprint* como alternativa na classificação estética de documentos

As diversas métricas existentes para a classificação estética de documentos tem como resultado um valor, ou seja, um peso referente à penalidade sofrida em conseqüência de uma

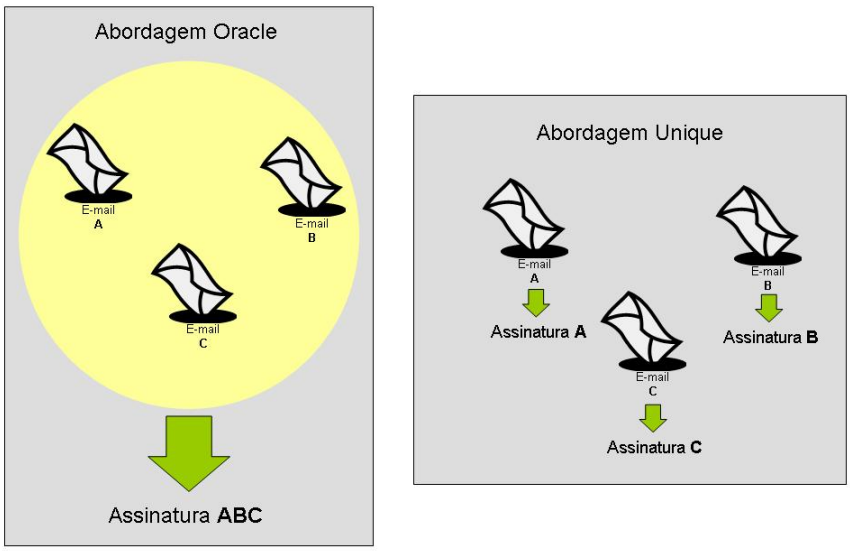


Figura 9 – Exemplo de construção de uma assinatura com a abordagem *unique* e *oracle*

alteração no documento gerado em comparação ao seu template de referência.
Similar ao processo de criação da assinatura de um e-mail, um documento gerado automaticamente também pode ter nele aplicadas diversas métricas, possibilitando o uso de tais técnicas para criar uma assinatura para cada documento automaticamente gerado.
Um exemplo de como é formada esta assinatura pode ser visto na Figura 10, onde o número 1 representa um template exemplo, o número 2 uma comparação de três documentos automaticamente gerados com o template referência bem como a quatro notas hipotéticas resultantes desta comparação e o número 3 ilustra a assinatura de cada documento.

O objetivo de criar a assinatura de um documento é possibilitar um processo de agrupamento de documentos. Este agrupamento procura associar documentos com assinaturas semelhantes e posteriormente, através de intervenções específicas de um usuário, determinar se tais grupos são de documentos bons ou ruins.
Estes agrupamentos serão feitos através do uso de métodos de clusterização, abordados na próxima seção.

## 3.3 Definição e usos dos métodos de clusterização

A clusterização é apresentada como uma técnica corriqueiramente utilizada para a análise estatística de dados, com flexibilidade suficiente para poder ser utilizada em diversas áreas in-

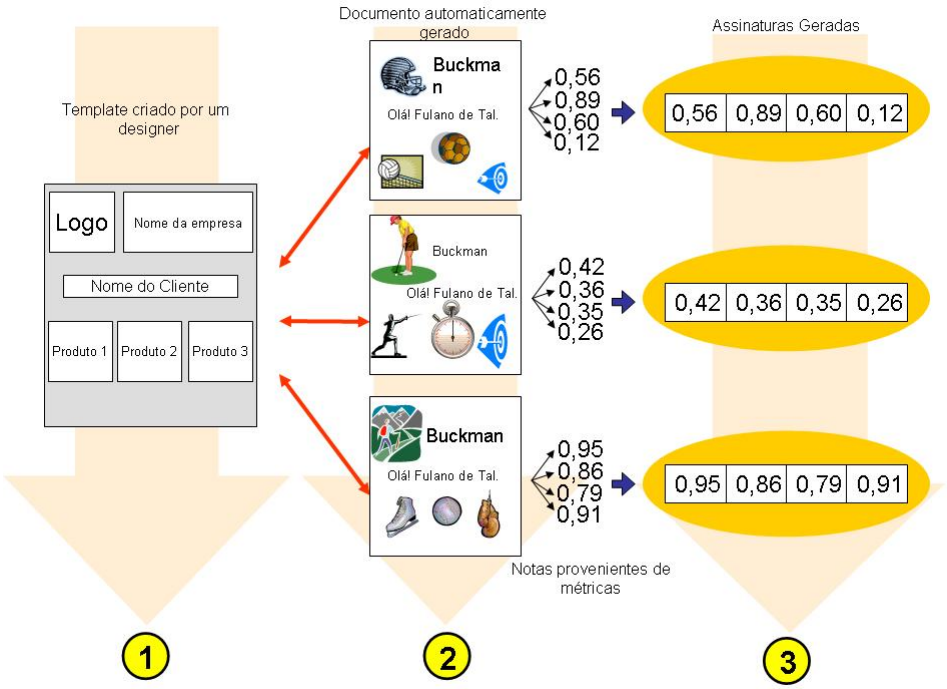


Figura 10 – Exemplo de aplicação de métricas de qualidade em um documento automaticamente gerado.

cluindo aprendizado de máquina, mineração de dados, reconhecimento de padrões, análise de imagens e bioinformática.

(Han & Kamber, 2000) ressaltam a importância das técnicas de clusterização como uma maneira de formação de grupos de objetos semelhantes, característica principal dos métodos de clusterização. Um exemplo, considerando a área da biologia, poderia ser um meio para se distinguir animais através de suas características.

As técnicas de clusterização funcionam de forma a "observar" um determinado conjunto de objetos visando separar *clusters* de objetos semelhantes. Tais características possibilitam um estudo sobre a aplicação dos métodos de clusterização para um processo de classificação estética de documentos.

As técnicas de clusterização serão aplicadas diretamente sobre um conjunto de assinaturas provenientes de documentos avaliados automaticamente, agrupando aquelas consideradas semelhantes. Uma vez realizada esta etapa, intervenções específicas (estas serão abordadas no decorrer to trabalho) são solicitadas, a um usuário, sobre os *clusters* identificados, visando classificá-los em grupos de documentos bons ou ruins. Cabe ressaltar que existem duas classificações, mas nem sempre apenas dois *clusters* serão identificados. A Figura 11 ilustra esta afirmação.

Os *clusters* podem ser formados através de dois processos básicos, um chamado de aprendizado **não-supervisionado**, onde o conjunto de dados de entrada é composto por exemplos não rotulados, ou seja, não existe uma classe (nome ou descrição) associada a cada exemplo

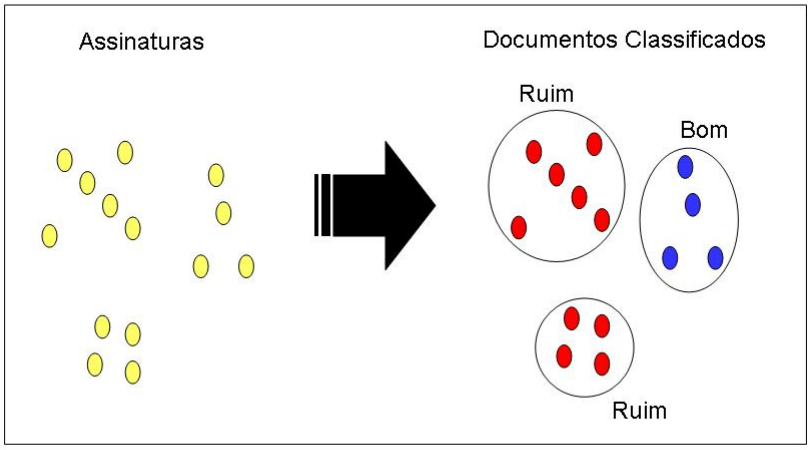


Figura 11 – Exemplo de identificação e classificação de *clusters*.

e aprendizado **supervisionado**, onde um especialista de domínio seria responsável a fornecer uma identificação à cada cluster gerado (Everitt, 1993), (Alpaydin, 2004).
Segundo (Han & Kamber, 2000) a maioria das técnicas de aprendizado de máquina são consideradas como sendo de aprendizado não supervisionado, já que buscam sempre suprir alguma necessidade de um especialista de domínio.
Para a ferramenta desenvolvida neste trabalho considera-se o uso do aprendizado **supervisionado**, já que intervenções específicas serão solicitadas a um usuário.

## 3.4 Medidas de similaridade

Para calcular a similaridade entre objetos, suas características são utilizadas e torna-se possível criar uma associação entre os objetos a serem clusterizados.
(Jain et al., 1999) menciona que pela variedade de características e escalas existentes, as medidas de similaridade precisam ser escolhidas de maneira cautelosa, sendo mais comum calcular a **dissimilaridade** entre os objetos.
Algumas medidas de similaridade se destacam na literatura sobre clusterização, são elas:

- Distância de Mahalanobis ((Mahalanobis, 1936))

- Distância de Manhattan ((Natsoulas, 1989))
- Distância Euclidiana

Considerando que o objetivo deste trabalho não é descobrir a melhor medida de similaridade para o problema da classificação estética de documentos, foi escolhida para este trabalho a Distância Euclidiana (DE) , que dentre as medidas de similaridade existentes, aparece como sendo amplamente utilizada para clusterização de documentos.
Em suma, a DE funciona de forma a representar através de um número a distância entre dois objetos. Na equação (3.1), $D(a, b)$ representa a distância euclidiana entre dois objetos $a$ e $b$, e $i$ representa cada característica dos objetos $a$ e $b$. Sendo $a = (a_1, a_2, ...a_n)$ e $b = (b_1, b_2, ...b_n)$

$$D(a, b) = \sqrt{(a_1 - b_1)^2 + ... + (a_n - b_n)^2} \tag{3.1}$$

## 3.5 Métodos de Clusterização

Esta seção tem como objetivo apresentar uma visão geral dos algoritmos de clusterização mais conhecidos na literatura, assim como o algoritmo implementado para este trabalho. Também será abordado um estudo empírico que foi realizado neste trabalho sobre o possível uso ou não dos algoritmos apresentados para a classificação estética de documentos.

### 3.5.1 Métodos hierárquicos

Os métodos hierárquicos (Xu & Wunsch, 2005) atuam de forma a produzir uma seqüência aninhada de partições, com um único cluster no topo seguido de vários níveis com diferentes especializações do conteúdo do cluster inicial, terminando em pontos únicos e específicos. Cada nível intermediário pode ser visto como uma combinação de dois *clusters* do nível inferior, ou também como a divisão de um cluster do nível superior.
O resultado de um algoritmo de clusterização hierárquico pode ser representado graficamente como uma forma de árvore, chamada de dendograma. Esta estrutura pode expressar graficamente o processo de união ou divisão entre os *clusters* e todos seus níveis intermediários.
Um exemplo pode ser visto na Figura 12 onde é apresentado um dendograma hipotético com cinco objetos unificados em um único cluster.

Basicamente existem duas abordagens para a geração de *clusters* hierárquicos:

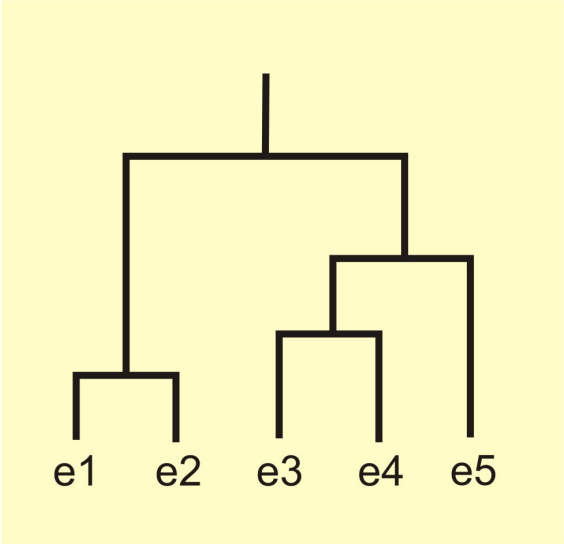


Figura 12 – Exemplo de um dendograma com cinco objetos

**Aglomerativa**: inicia transformando cada objeto em um cluster, para então unir estes *clusters* até que reste apenas um único cluster com todos os objetos (*bottom-up*), ou até que uma determinada condição seja satisfeita. Segundo (Han & Kamber, 2000) grande parte dos algoritmos para *clusters* hierárquicos pertence a esta categoria, diferenciando-se apenas na medida de similaridade utilizada.

**Divisiva**: pode ser considerada como *top-down*. Faz o oposto da técnica aglomerativa, ou seja, inicia com todos os elementos em um único cluster, e este é subdividido até que cada objeto possua seu próprio cluster ou uma determinada condição seja satisfeita. Esta condição pode envolver um número máximo de *clusters* gerados ou a distância entre dois *clusters* mais próximos esteja acima de um determinado limiar.
A Figura 13 apresenta de forma ilustrativa como seria a criação de *clusters* segundo os métodos apresentados.

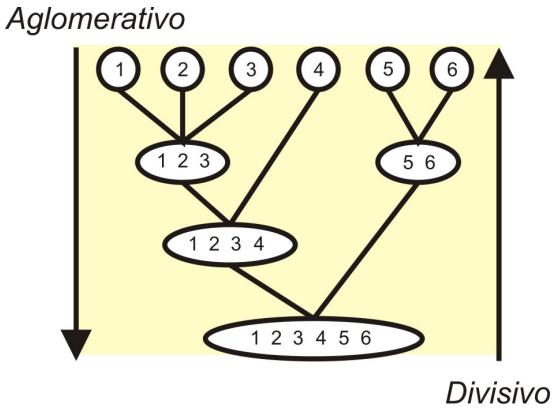


Figura 13 – Ilustração do processo Divisivo e Aglomerativo

Além destas classificações, os algoritmos podem ser também classificados de acordo com a forma que os pares de objetos são escolhidos para o cálculo da função de similaridade, podendo ser *single-link* ou *complete-link*.

### 3.5.2 Diferença entre *single-link* e *complete-link*

A diferença básica entre estas duas técnicas está na maneira com que são escolhidos os pares de objetos de distintos *clusters* para a formação ou não de novos *clusters*.
A técnica de *single-link* funciona buscando os dois objetos mais próximos em *clusters* distintos (vizinhos mais próximos) como forma de determinar a similaridade entre estes *clusters*, determinado se podem ou não pertencer ao mesmo cluster. Os *clusters* unidos são teoricamente aqueles mais próximos.
A técnica de *complete-link* funciona buscando os dois objetos mais distantes de *clusters* distintos (vizinhos mais distantes) como forma de determinar a similaridade entre estes *clusters*, determinando se estes podem ou não pertencer ao mesmo cluster. Os *clusters* a serem unidos são aqueles que teoricamente possuem o menor diâmetro(distância entre dois objetos mais distantes dentro de um cluster).
A Figura 14 ilustra a escolha dos objetos seguindo as técnicas de *single-link* e *complete-link*.

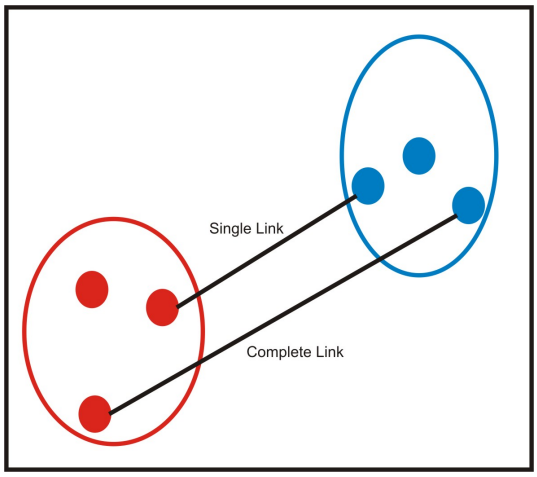


Figura 14 – Single Link vs Complete Link

### 3.5.3 Métodos de Particionamento

Métodos de particionamento (Berkhin, 2006) consistem em dividir os objetos em $K$ *clusters* distintos seguindo uma definição de similaridade. A idéia é que os elementos dentro de

um cluster possuam altos valores de **similaridade** uns com os outros, enquanto a relação com os demais *clusters* apresente grande **dissimilaridade**. Dentre suas técnicas, merecem destaque devido a sua popularidade as técnicas de ***K-Means*** e ***K-Medóides***.
Os métodos de particionamento caracterizam-se pela necessidade de informar previamente um número $K$ de *clusters* a serem criados.

*K-Means*

O algoritmo *K-Means* ((MacQueen, 1967), (Hartigan & Wong, 1979)) é baseado na definição de um ponto central para cada cluster, e a partir deste buscar os objetos mais próximos a este ponto criando os *clusters*.
Este ponto central funciona como uma espécie de centro de gravidade (centróide) de cada cluster, seu objetivo é fazer com que os objetos pertencentes a um cluster possuam alta similaridade entre si e a relação deste cluster com outros *clusters* identificados apresente baixa similaridade. Uma forma de calcular o centróide de um cluster é fazendo a média entre os objetos existentes dentro deste cluster.
Um algoritmo de *K-Means* 1 é apresentado a seguir onde $K$ representa o número de *clusters* a serem criados.

**Algoritimo 1** Algoritmo *K-Means*

Escolha de maneira arbitrária $K$ valores como centróides iniciais;
**repeat**
  associar cada objeto restante com o centróide mais próximo;
  redefine novos centróides para cada cluster calculando uma média entre os objetos de cada cluster;
**until** (não ocorrerem mais mudanças de centróides)

A Figura 15 ilustra em quatro passos a formação de 4 *clusters* com o métodos de *K-Means*. No passo 1, aleatoriamente foram escolhidos 4 pontos que representam 4 centróides (pequenos quadrados), os passos 2 e 3 apresentam as alterações dos centróides e dos objetos que fazem parte de cada cluster e no passo 4 o resultado do processo.

*K-Medóides*

O algoritmo K-Medoids ((Jain & Dubes, 1988), (Bocca et al., 1994)) foi desenvolvido como uma variação do *K-Means*. A diferença básica entre os dois métodos está no fato que a técnica de *K-Means* faz o cálculo de um centróide como ponto de referência para a formação de *clusters*, enquanto a técnica de K-Medoids busca o objeto mais central como ponto de referência.
Este objeto mais central, chamado de medóide, é aleatóriamente escolhido a cada interação

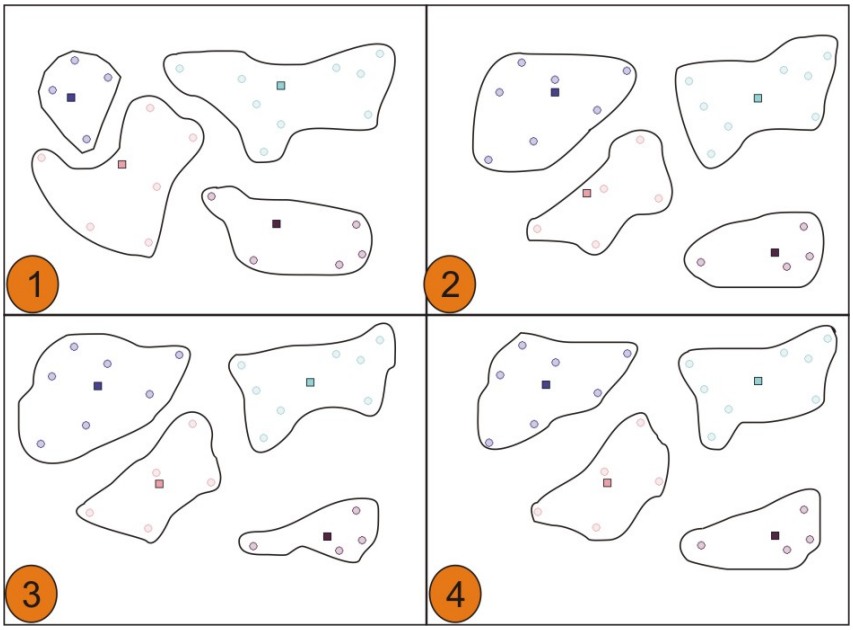


Figura 15 – Formando *clusters* com *K-Means*

deste algoritmo visando encontrar um melhor agrupamento dos objetos.
O número de medóides é diretamente proporcional ao número $K$ de *clusters* a serem criados.
Como forma de determinar os medóides de um conjunto de objetos a ser clusterizado, quatro casos podem ser analisados para que um objeto não medóide ($O_{rand}$, encontrado aleatoriamente dentro de cada cluster já formado) possa vir a ser um bom substituto para o medóide atual ($o_j$), os quatro casos a seguir são examinados para cada objeto ($p_i$) considerando este, um objeto de um cluster em específico e não medóide.

- **Caso 1:** $p_i$ atualmente pertence ao cluster do medóide $o_j$. Se $o_j$ for substituído por $O_{rand}$ como medóide e $p_i$ está mais perto do medóide $o_i$, sendo $i \neq j$, então $p_i$ passa a fazer parte do cluster representado pelo medóide $o_i$.

- **Caso 2:** $p_i$ atualmente pertence ao cluster do medóide $o_j$. Se $o_j$ for substituído por $O_{rand}$ como medóide e $p_i$ está mais perto de $O_{rand}$ então $p_i$ passa a fazer parte do cluster representado por $O_{rand}$.

- **Caso 3:** $p_i$ atualmente pertence a outro cluster que não o representado pelo medóide $o_j$. Se $o_j$ é substituído por $O_{rand}$ como um medóide e $p_i$ continua sendo mais próximo de $o_i$, então $p_i$ permanece em seu cluster atual.

- **Caso 4:** $p_i$ atualmente pertence ao cluster representado pelo medóide $o_i$, sendo $i \neq j$. Se $o_j$ é substituído por $O_{rand}$ como um medóide e $p_i$ é o mais próximo do medóide $O_{rand}$, então $p_i$ é re-posicionado para o cluster representado pelo medóide $O_{rand}$.

Um algoritmo para o método *K-Medóides* (Algoritmo 2), funciona tendo como entrada um número $K$, referente ao número de *clusters* que se pretente encontrar e uma base de dados contendo $n$ objetos e o resultado esperado é um conjunto de $K$ *clusters* que minimizam a soma das dissimilaridades de todos os objetos ao seu medóide mais próximo.

**Algoritimo 2** Algoritmo *K-Medóides*

Escolha de maneira arbitrária $K$ objetos como medóides iniciais;
**repeat**
  associar cada objeto restante com o medóide mais próximo;
  randomicamente escolha um objeto não medóide $O_{rand}$;
  compute o custo total, $S$, de trocar $o_j$ por $O_{rand}$;
  **if** $S < 0$ **then**
    troque o medóide atual com $O_{rand}$ para formar o novo conjunto de medóides;
  **end if**
**until** (não ocorrerem mais mudanças de medóides)

Este algoritmo foi escolhido para ser implementado neste trabalho. Esta escolha foi feita considerando que um dos documentos a serem intervidos por um usuário, seria o mais central de cada cluster. Como o algoritmo de *K-Means* trabalha com centróides, seria necessário ainda, descobrir qual o documento mais próximo deste centróide. O *K-Medóides* elimina esta etapa de processamento.

### 3.5.4 Métodos baseados em densidade

Segundo (Kurniawan et al., 1999) os algoritmos baseados em densidade buscam identificar regiões dentre os objetos a serem clusterizados que apresentem maior número de pontos com maior vizinhança (maior densidade).
Dentre os métodos baseados em densidade, cabe ressaltar o algoritmo proposto por (Ester et al., 1996)(DBSCAN). Sua idéia principal é que para cada objeto a ser clusterizado é determinado por um usuário um "raio" de alcance que visa criar uma "circunferência" ao redor deste objeto. Para que o *cluster* seja formado, se faz necessário também informar um valor $n$ que se refere ao número mínimo de objetos que devem estar dentro desta circunfência. Contemplar tais premissas culmina na formação dos *clusters*.
Como exemplo de funcionamento deste algoritmo, a Figura 16 servirá de exemplo. O retângulo marcado com o número **1** ilustra um conjunto de objetos a serem clusterizados. O retângulo marcado com o número **2** apresenta um raio de alcance hipotético traçado ao redor de cada objeto. Considerando um critério onde dentro da circunferência de cada objeto devam existir

no mínimo outros dois objetos distintos para a formação de um *cluster* os retângulos marcados pelos números **3** e **4** apresentam três *clusters* formados com dois ou mais objetos e outros três objetos que não fazem parte de *cluster* algum.

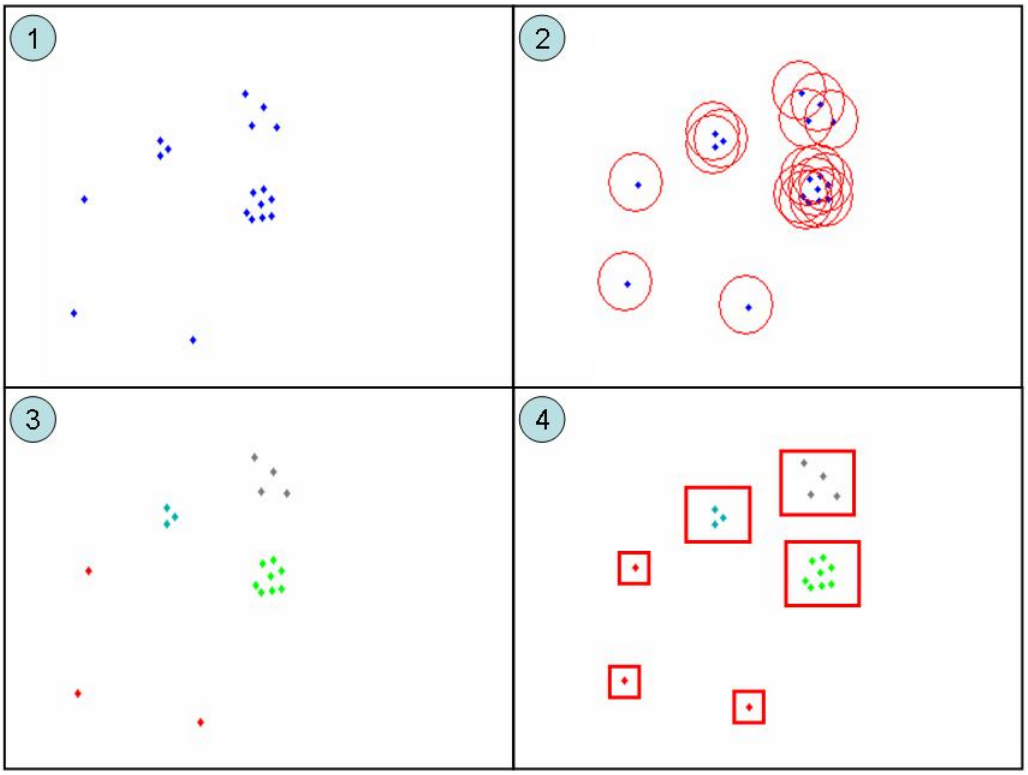


Figura 16 – Passa a passo da formação de *clusters* baseados em densidade.

## 3.6 Considerações do capítulo

Neste capítulo foi apresentada uma revisão bibliográfica acerca do trabalho proposto.
As técnicas de *fingerprint* apresentadas neste capítulo servem como base para a criação da assinatura de um e-mail, ou seja, seu identificador.
Pelo vasto número de medidas de similaridade e algoritmos de clusterização existentes, os algoritmos considerados mais populares na literatura pesquisada foram escolhidos para serem apresentados neste estudo. Destaca-se o uso da distância euclidiana e do algoritmo de *K-Medóides* para serem utilizados com a ferramenta de classificação estética de documentos apresentada no decorrer do trabalho.

# 4 Ferramenta de classificação estética

Neste capítulo será apresentado o processo de classificação estética proposto neste trabalho, assim como o algoritmo de clusterização escolhido para trabalhar com a ferramenta de classificação desenvolvida.
A ferramenta a ser apresentada divide-se em um conjunto de módulos capazes de realizar tarefas específicas ao algoritmo de clusterização utilizado, visando reduzir a intervenção humana no processo de classificação estética de documentos.

## 4.1 Classificando documentos esteticamente

O processo de classificação de documentos abordado neste trabalho tem o objetivo de reduzir a intervenção humana. Atualmente um usuário precisa verificar se cada documento automaticamente gerado está bom (com poucos ou nenhum problema estrutural) ou ruim (com significativos problemas estruturais), processo considerado oneroso e demorado. Como forma de reduzir a intervenção humana, supõe-se que um documento é identificado por uma assinatura composta por notas provenientes de métricas estéticas que mensuram o quanto um documento automaticamente gerado difere de seu template.
Para este trabalho, cada documento terá de três a seis notas em suas assinaturas. Também é considerado que um documento com uma assinatura de três notas é um documento de três dimensões (3D), um documento com quatro notas em sua assinatura, quatro dimensões (4D) e assim por diante. Maiores detalhes sobre os documentos usados para os testes deste trabalho e suas assinaturas serão apresentados no capítulo decorrente.
Em um processo de geração em larga escala de documentos de conteúdo variável, fatores como nomes de destinatários, padrões de consumo, endereços e demais, culminam na criação de documentos de aparência semelhante e assinaturas semelhantes. A Figura 17, mostra um exemplo de dois documentos para duas pessoas com características semelhantes.

Se destes documentos fossem geradas suas assinaturas e estas usadas para representá-los em um plano cartesiano, a tendência é que estas estariam bastante próximas devido a estas pessoas possuírem características semelhantes. Seguindo a mesma linha de raciocínio, em um processo de criação em larga escala deste tipo de documento, infere-se uma grande quantidade de do-


Figura 17 – Semelhanças entre documentos gerados automaticamente

cumentos separados em grupos com características singulares entre si, um exemplo pode ser observado na Figura 18.

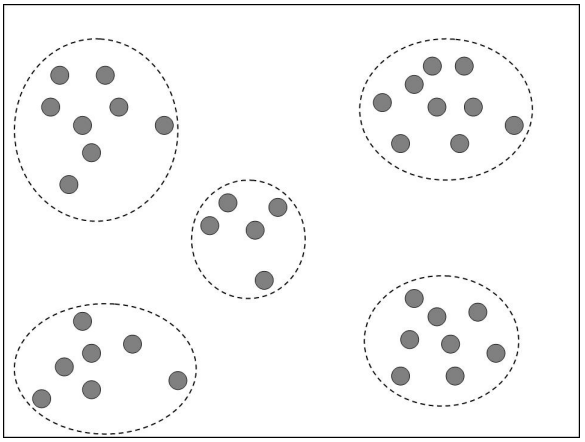

Figura 18 – Documentos representados em um plano através de suas assinaturas

Para delimitar estes agrupamentos de forma que seja possível fazer uso desta informação de forma a classificar documentos estéticamente, está sendo proposto o uso de métodos de clusterização. Acredita-se seu uso ser vantajoso pois anteriormente uma pessoa precisaria verificar um a um os documentos gerados, já com esta delimitação de grupos, a idéia é que através da análise de um pequeno número de documentos de cada grupo delimitado, solicitando que o usuário os classifique em bons ou ruins seja possível inferir que os demais pertencentes a estes mesmos grupos assumam tal classificação, possibilitando uma redução da intervenção humana no processo de classificação estética.

## 4.2 Método de clusterização utilizado

Os métodos de particionamento, conforme previamente abordado no capítulo 3, consistem em dividir um conjunto de objetos a serem clusterizados em $K$ grupos previamente informados por um usuário.

Como o objetivo deste trabalho está em estudar a possibilidade do uso de métodos de clusterização como forma de reduzir a intervenção humana no processo de classificação estética de documentos, e não em fazer uma comparação entre os diferentes métodos de clusterização elegendo o melhor para este problema, a idéia foi fazer o uso dos métodos de particionamento devido a sua grande popularidade na literatura sobre clusterização.

A escolha de um valor $K$ inicial, premissa para os métodos de particionamento, pode ser considerada uma vantagem no sentido de se tornar possível inferir um número de intervenções feitas a um usuário de acordo com o $K$ fornecido, por exemplo, se em cada cluster criado um documento for apresentado a um usuário para que este o classifique em bom ou ruim o número de perguntas será diretamente proporcional ao número $K$, por outro lado, necessitar que seja informado um valor $K$, para um conjunto de objetos que muitas vezes não se tem um conhecimento prévio é um processo complexo. Considerando este problema neste trabalho também será apresentada uma forma de automaticamente encontrar este valor $K$.

Dentre os métodos de particionamento, destacam-se os algoritmos K-Means e K-Medóides. A diferença básica entre os dois métodos reside no fato que a técnica de K-Means faz o cálculo de um centróide representando o ponto mais central de um cluster, enquanto a técnica de K-Medóides busca encontrar o objeto mais central de um cluster, ou seja, um medóide. Neste trabalho, foi feita uma implementação do algoritmo de K-Medóides, este algoritmo, foi proposto por (Kaufmann & Rousseeuw, 1987). Uma representação em pseudocódigo deste algoritmo já foi apresentada no capítulo 3 subseção 3.5.3.

Para funcionar em conjunto com o algoritmo implementado a medida de similaridade utilizada foi a Distância Euclidiana (Capítulo 3 seção 3.4). Outras formas de similaridade poderiam ter sido usadas, mas como o objetivo deste trabalho não é apresentar uma comparação entre as diferentes funções de similaridade que podem ser utilizadas no processo de classificação estética de documentos, optou-se por uma medida de similaridade comumente utilizada.

Sua equação pode ser expressa da seguinte forma: A Distância Euclidiana $D(a, b)$ representa a distância entre dois objetos $a$ e $b$, já $a_n$ e $b_n$ representam $n$ características dos objetos $a$ e $b$.

$$D(a, b) = \sqrt{(a_1 - b_1)^2 + ... + (a_n - b_n)^2} \tag{4.1}$$

Outro cálculo importante para o método de clusterização implementado, seria determinar o custo de escolher um novo medóide, para isto foi utilizada a equação (4.2) onde $C$ representa o custo total de uma troca de medóides, $K$ o número de clusters existentes, $D(a_i, b_j)$ a distância

euclidiana entre as assinaturas de dois documentos $a_i$ e $b_j$ e $n$ o número total de documentos existentes no total disponível a ser classificado, levando em conta que $a_i$ e $b_j$ pertencem a um cluster específico.

$$C = \frac{\sum_K \sum_{i,j} (D(a_i, b_j))}{\frac{n(n-1)}{2}} \tag{4.2}$$

Considerando que um documento previamente considerado como medóide $o_j$ seja substituído por um possível documento considerado um novo medóide $O_{rand}$, se o custo $C$ obtido seja menor para $O_{rand}$ que para $o_j$, $o_j$ é trocado por $O_{rand}$.

## 4.3 Ferramenta de classificação estética de documentos

A ferramenta de classificação estética de documentos que será proposta divide-se em cinco módulos, conforme a Figura 19. O módulo de entrada de dados funciona como ponto de partida para um processo de classificação fornecendo o conjunto de assinaturas a serem classificadas, o módulo de clusterização tem o objetivo de fazer com que o método de clusterização implementado seja executado, o módulo de avaliação armazena as classificações (bom ou ruim) das assinaturas, o módulo de execução determina a quantidade de execuções que serão feitas sobre o conjunto de assinaturas a serem classificadas e o módulo de visualização que apresenta a um usuário os resultados obtidos. Uma visão geral do funcionamento da ferramenta de classificação estética pode ser vista na Figura 20. Nesta figura, o passo **1** representa a criação de um documento modelo (template) por um usuário.
No passo **2**, são criadas diversas variações sobre este template de forma a satisfazer condições que atendam um público alvo. Os padrões de conteúdo destes arquivos serão descritos no decorrer.
No passo **3**, é realizada a criação das assinaturas para cada um dos documentos automaticamente gerados, seguindo os padrões vistos na seção 3.1.
No passo **4** e **5** é iniciado o processo de classificação estética com intervenção de um usuário. Seus padrões de trocas de mensagens e apresentação dos resultados são descritos no decorrer deste capítulo.
No passo **6** a ferramenta finaliza o processo de classificação, separando os documentos bons e ruins.

No decorrer da seção irão ser apresentados em maiores detalhes cada módulo desenvolvido, bem como a linguagem de programação e banco de dados utilizados no desenvolvimento da ferramenta de classificação estética de documentos.

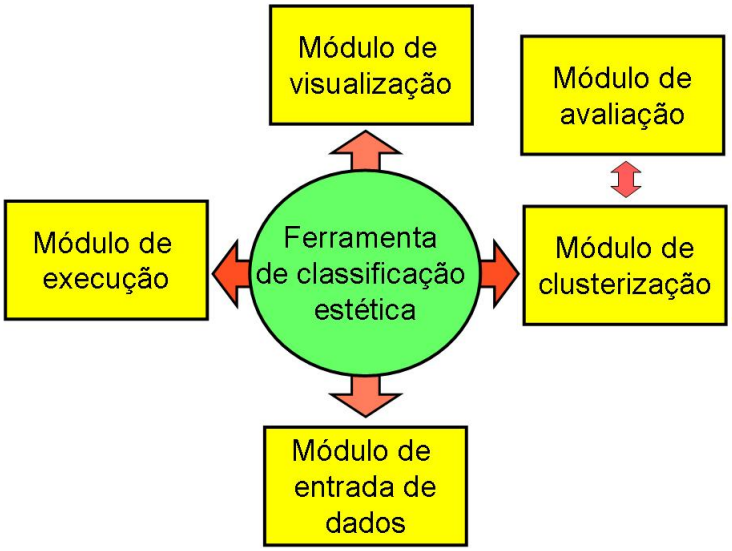


Figura 19 – Visão geral dos módulos da ferramenta de classificação estética de documentos

### 4.3.1 Linguagem de programação e banco de dados utilizados

Para este trabalho, as implementações realizadas foram feitas com a linguagem de programação PHP (Group, 2006) com apoio do Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL (AB, 2006). O uso de tais tecnologias se deu devido à qualidade de documentação, o fato de serem livres e a familiaridade do autor com tais tecnologias.

### 4.3.2 Módulo de execução

O módulo de execução é o ponto de partida do classificador estético, nele é informado o nome do arquivo que possui as assinaturas a serem clusterizadas e classificadas, além de ser informado também o número de execuções que cada arquivo deve fazer.
O módulo também controla se alguma execução atual não pode ser completada, para isto ele verifica se o número de interações atuais é igual ao número total de documentos a serem classificados. Em caso positivo, é inferido que o método de clusterização não foi capas de concluir sua execução, por exemplo, se existirem 100 documentos a serem classificados e 100 intervenções forem feitas a um usuário.

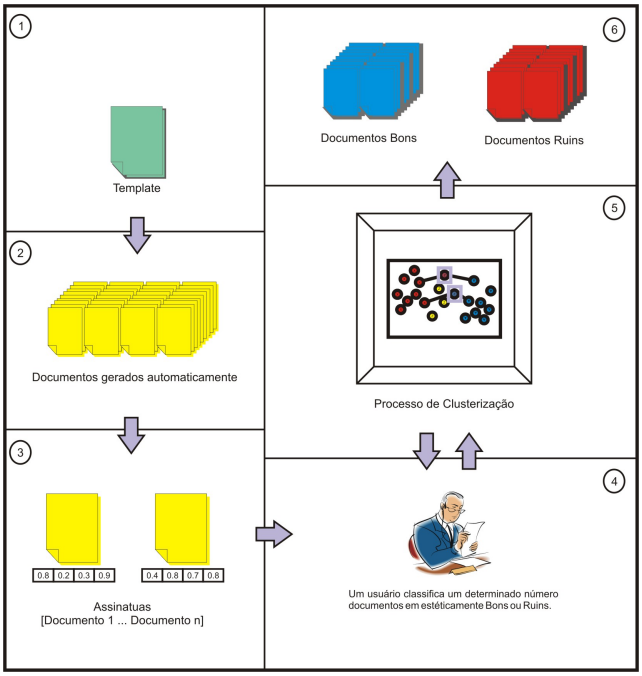


Figura 20 – Visão geral do funcionamento da ferramenta de classificação estética

### 4.3.3 Módulo de entrada de dados

O módulo de entrada de dados desenvolvido tem como base para o seu funcionamento um conjunto de arquivos texto no formato CSV (*Comma-separated values*) o padrão dos arquivos de entrada não fica restrito a forma com os seus conteúdos são dispostos, mas também nos seus nomes, como pode ser observado na Figura 21 marcação **A** representa o número de clusters que existem naquele arquivo, no caso deste exemplo, 2 clusters e marcação **B** que representa a quantidade de assinaturas a serem classificadas.
Considerando a mesma Figura, os valores que variam de 1 a 3, representam respectivamente, o identificador de cada assinatura, a classificação (0 documento ruim, 1 documento bom) e um conjunto de notas, no caso do exemplo, uma assinatura em 3 dimensões.

### 4.3.4 Módulo de avaliação

O módulo de avaliação foi criado de forma a armazenar e gerenciar as classificações dadas a documentos por um usuário a documentos que estão sendo classificados. Tais classificações são armazenadas no SGBD da ferramenta.
A tabela criada no SGBD que armazena tais informações pode ser observada na Figura 22, onde $id$ refere-se apenas a um identificador de execução, $id - documento$ o identificador de cada assinatura de cada documento classificado, $cluster$ seria o cluster ao qual a determinada

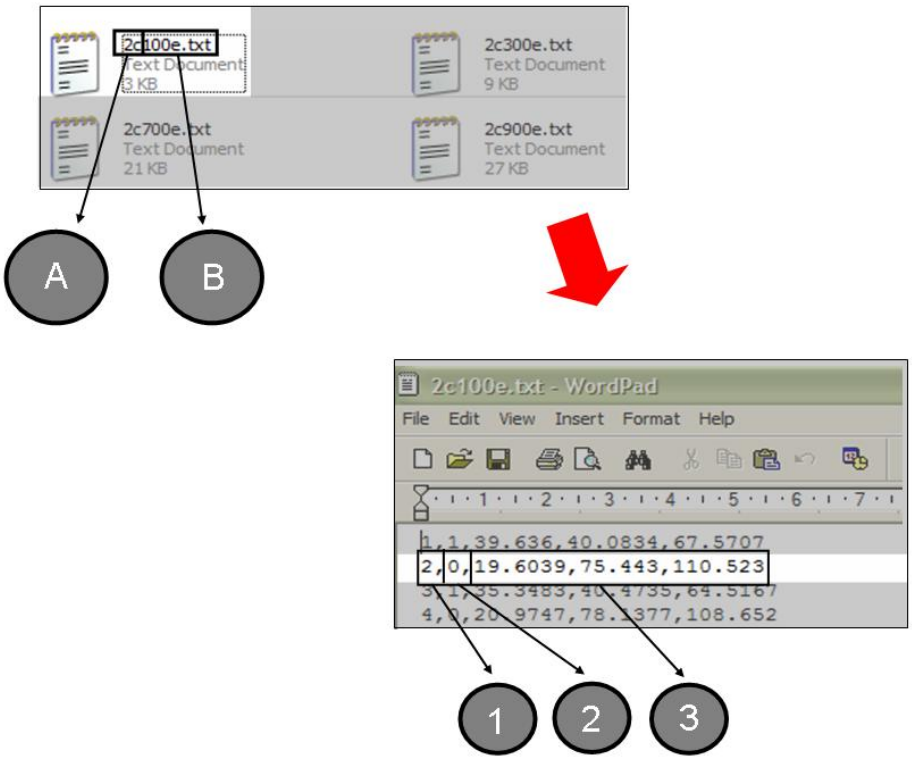


Figura 21 – Arquivo de entrada no formato CSV

assinatura pertence, este campo é utilizado pelo atribuidor de notas que será apresentado no decorrer do texto e $classificacao$ que se refere a nota dada a um documento por um usuário.

| Field | Type | Null | Key | Default | Extra |
|---|---|---|---|---|---|
| id | int(11) | NO | PRI | NULL | auto_increment |
| id_documento | int(11) | YES | | NULL | |
| cluster | int(11) | YES | | NULL | |
| classificacao | int(11) | YES | | NULL | |

Figura 22 – Estrutura da tabela do módulo de avaliação

### 4.3.5 Módulo de clusterização

O objetivo do módulo de clusterização é possibilitar que sejam feitas interações específicas com o algoritmo de clusterização escolhido para este trabalho (K-Medóides) de forma a eliminar a necessidade de um usuário informar um valor $K$. Para que seja atingido este objetivo, este módulo precisa interagir diretamente com o módulo de avaliação, mencionado na subseção 4.3.4.

Antes de serem abordados detalhes mais específicos, ressalta-se o fato de que podem existir $n$ clusters divididos em apenas duas classificações, como exemplo pode-se observar a Figura 23, onde quatro clusters são dispostos em um plano e divididos em duas classificações.
Como forma de iniciar o processo de identificação de um valor $K$ que representa um conjunto de documentos que estão sendo clusterizados partiu-se primeiramente de idéia que no mínimo existirão 2 clusters ($K = 2$), um cluster das assinaturas de documentos considerados bons e um cluster das assinaturas dos documentos considerados ruins, desta forma a ferramenta de classificação estética, através do algoritmo implementado, clusteriza inicialmente com $K = 2$ o conjunto de documentos.

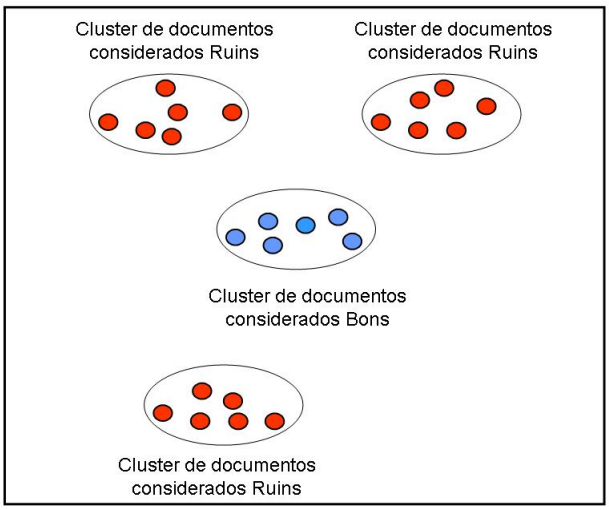


Figura 23 – Ilustração de um espaço com 4 clusters e 2 classificações

Como previamente mencionado, não necessariamente irão existir 2 clusters possibilitando um $K > 2$, a idéia então, é solicitar uma intervenção de um usuário a documentos específicos, fazendo com que este os classifique em bons ou ruins de acordo com o seu gosto, conforme os resultados será determinado se o valor $K$ será ou não incrementado.
A intervenção que irá determinar se o valor $K$ deve ou não ser incrementado é realizada conforme apresentado na Figura 24, onde dentro de cada cluster formado são buscados os dois documentos mais distantes entre si e o medóide, para então apresentá-los a um usuário e solicitar que este o classifique em bom ou ruim.
Um exemplo de como são feitas na prática as intervenções por um usuário em um conjunto de documentos que se dividem em dois clusters é apresentado na Figura 25. Como pode ser observado na etapa 1, no primeiro cluster a ser avaliado, os dois documentos mais distantes e o medóide tiveram a mesma classificação (B), desta forma foram computadas três intervenções a um usuário. Na etapa 2, no segundo cluster a ser avaliado, outras três intervenções foram realizadas com classificações diferentes (R), totalizando seis intervenções.

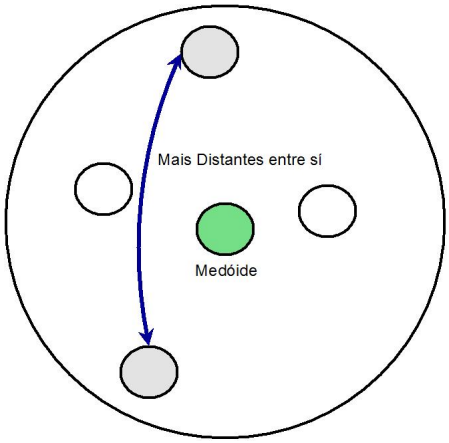


Figura 24 – Documentos selecionados em um cluster para um interação com um usuário

Se em algum dos clusters formados um usuário classifique de forma diferente os documentos, infere-se que o valor $K$ precisa ser incrementado.

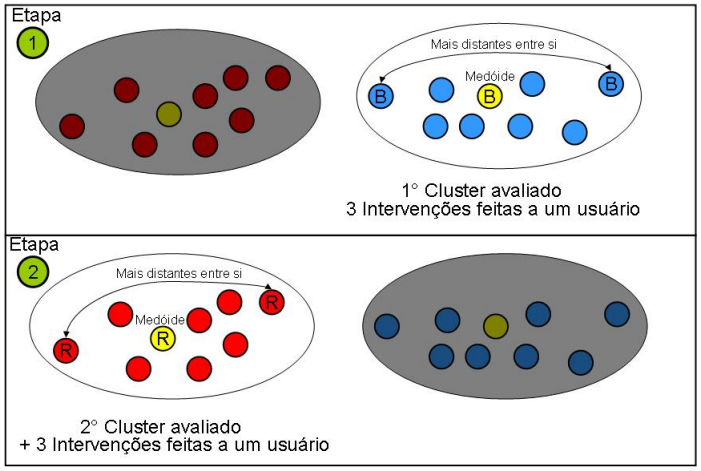


Figura 25 – Exemplo de classificação estética de dois clusters de documentos

Como forma de realizar este incremento, para este trabalho foram implementadas duas técnicas.

- Sem re-alimentação
- Com re-alimentação

A técnica **Sem realimentação**, faz com que o algoritmo de clusterização seja executado novamente passando como novo valor $K$ o somatório em 1 do $K$ anterior, sem informar ao

algoritmo de clusterização onde este deve procurar por um novo cluster.
A técnica **Com realimentação** funciona com o objetivo de ajudar o algoritmo de clusterização, informando possíveis novos medóides para cada cluster novo a ser formado, eliminando o primeiro passo de encontrar o primeiro conjunto de medóides de forma aleatória.
Para elucidar o funcionamento da técnica com realimentação, tem-se como base o exemplo expresso na Figura 26. Na etapa A, é apresentada a clusterização de um conjunto de documentos com um $K = 2$, como pode ser observado, este valor não delimitou corretamente o conjunto de documentos a serem classificados, já que, na primeira intervenção a um cluster ouve uma diferença na classificação dos dois documentos mais distantes, fato este que determina um incremento no valor $K$.
Na etapa B, este método auxilia o algoritmo de clusterização, sendo fornecidos os dois medóides previamente encontrados e mais um novo, que foi escolhido por ter sido dentre os classificados por um usuário o aquele com classificação distinta.
Na etapa C, o algoritmo de clusterização funciona como já descrito delimitando os grupos ao redor de seus medóides, podendo possivelmente ser encontrados novos medóides que melhor representem os clusters formados.
Na etapa D, novamente são feitas intervenções a um usuário de forma a verificar se os clusters foram corretamente formados.

### 4.3.6 Módulo de visualização

O módulo de visualização é responsável por fazer a comunicação entre a ferramenta e um usuário que esteja realizando a classificação de documentos.
Na etapa de preparação dos dados, este módulo coleta informações referentes ao nome dos arquivos que contém os documentos a serem classificados e o número de execuções que serão feitas sobre cada um destes arquivos. Durante o processo de classificação, este módulo é responsável por coletar junto a um usuário sua classificação ( bom ou ruim ) de documentos específicos e no final do processo este módulo apresenta estatísticas referentes ao número de execuções que obtiveram sucesso e a média de perguntas feitas a um usuário em cada arquivo testado.

## 4.4 Considerações do capítulo

Neste capítulo foi apresentado o processo de classificação estética de documentos proposto neste trabalho. Também foi descrito o motivo do uso de um algoritmo de clusterização baseado

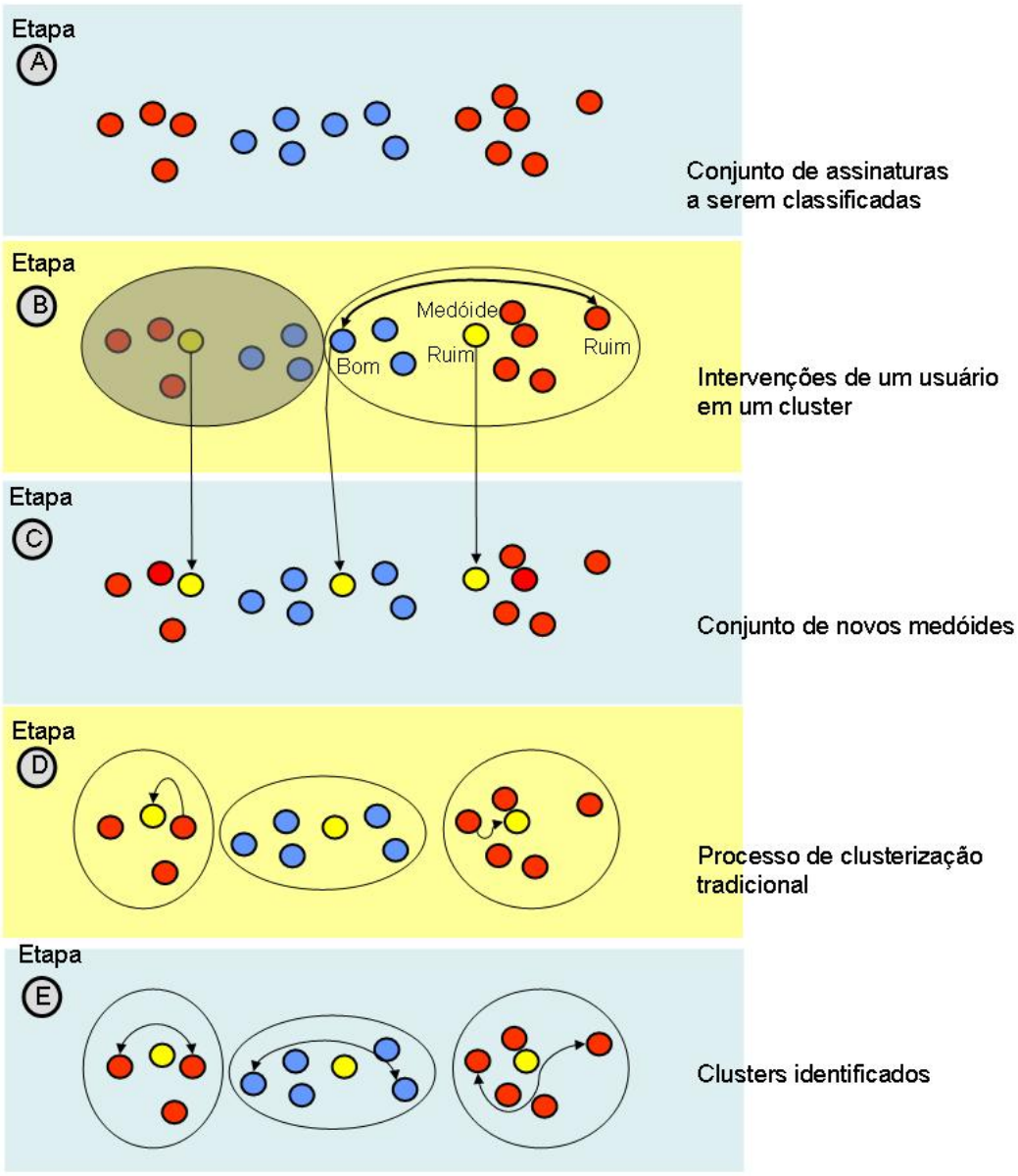


Figura 26 – Exemplo de criação de clusters com realimentação

em particionamento chamado de K-Medóides.
Como este algoritmo necessita de que seja informado um número de clusters inicial ($K$) e este processo é considerado difícil devido à imprevisibilidade das assinaturas dos documentos a serem classificados, foram propostas também duas técnicas que automaticamente identificam este valor $K$, uma técnica sem realimentação de novos possíveis medóides e uma técnica que realimenta a ferramenta de classificação com possíveis novos medóides.

# 5 Metodologia de avaliação

Este capítulo tem como objetivo apresentar os conjuntos de documentos utilizados nos experimentos com a ferramenta de classificação de documentos e a forma como foram conduzidos tais experimentos de forma a verificar se o objetivo de reduzir a intervenção humana no processo de classificação foi atingido.

## 5.1 Preparação dos dados

O objetivo da metodologia de classificação estética proposta neste trabalho é reduzir a interação humana no processo de classificação estética de documentos. Para verificar a funcionalidade da ferramenta a opção ideal seria um experimento, usando documentos e usuários reais, algo difícil de ser feito já que se depende de diversos fatores, como por exemplo, a criação de um template para posterior criação de documentos gerados automaticamente; criar uma assinatura para cada documento gerado através da aplicação de métricas estéticas e finalmente dispor no mínimo de um usuário para utilizar a ferramenta de classificação.
Um processo experimental que faça uso destes passos é lento e oneroso, já que diversas variações de documentos teriam que ser criadas.
Necessitando de uma abordagem que pudesse trazer resultados próximos aos que se alcançaria com este modelo e que tornasse o processo experimental mais rápido foram desenvolvidas duas ferramentas que recriam de forma controlada conjuntos de assinaturas de documentos que simulem variações que podem ser encontradas nos documentos gerados no mundo real.
Como mencionado, conjuntos de documentos gerados podem estar distribuídos em *clusters*, que podem estar mais próximos ou distantes entre si, mas suas classificações dividem-se apenas em bons ou ruins.
De forma a recriar tais situações foi desenvolvido um gerador de assinaturas que tem a capacidade de criar diversas variações de *clusters*, compostos por documentos gerados automaticamente e com características semelhantes.
A primeira etapa do gerador de assinaturas está em criar um conjunto de "sementes" que representem as posições em um espaço de $n$ dimensões que cada *cluster* a ser gerado deve ocupar, permitindo a criação de *clusters* mais próximos ou mais distantes entre si. Estes *clusters* serão populados por assinaturas criadas de forma semelhante ao processo de aplicação de métricas

estéticas. O gerador de assinaturas não irá criar documentos, mas sim, um conjunto de assinaturas baseadas nas posições de cada semente, criando *clusters* que simulem as variações encontradas em documentos de conteúdo variáveis no mundo real. A possibilidade de criar este gerador reside no fato de que a ferramenta de classificação considera apenas a assinatura de um documento. Se esta assinatura foi gerada a partir de um documento real, ou automaticamente gerada, não faz diferença para o seu funcionamento.
Uma segunda ferramenta chamada de atribuidor de notas foi desenvolvida. O objetivo deste atribuidor de notas é eliminar o trabalho de um usuário de classificar em bons ou ruins as requisições feitas pela ferramenta de classificação estética.

### 5.1.1 Gerador de assinaturas

A assinatura de um documento é proveniente da aplicação de métricas estéticas sobre um documento automaticamente gerado em comparação ao seu template de forma a mensurar suas diferenças.
Um documento pode possuir associadas a ele quantas notas um usuário considerar necessárias para criar a sua assinatura. Como cada nota representa uma dimensão, um documento pode ser representado em várias dimensões.
Documentos a serem classificados podem estar distribuídos em *clusters* com as mais diversas variações de coesão e acoplamento. Valores de coesão determinam a distância entre os objetos de cada *cluster*, enquanto o acoplamento determina a distância entre os *clusters* formados. Por exemplo, *clusters* com altos valores de coesão e baixo acoplamento (Figura 27), caracterizam *clusters* bastante definidos e distantes entre si, enquanto *clusters* com baixo valor de coesão e alto acoplamento (Figura 28) caracterizam *clusters* mais difíceis de serem delimitados.

O gerador de assinaturas desenvolvido funciona com cinco variáveis de configuração, que trabalhando em conjunto proporcionam a geração de assinaturas que possam recriar a diversidade de situações que podem ser encontradas. Estas variáveis seriam:

- Quantidade de *clusters* a serem criados ($Q_c$);
- Quantidade de assinaturas a serem geradas ($Q_a$);
- Quantidade de dimensões para cada assinatura gerada ($Q_d$);
- Valor relativo à distância máxima entre as assinaturas de cada *cluster* que controla as variações de coesão ($D_a$);
- Valor relativo à distância máxima entre as "sementes" que controla as variações de acoplamento ($D_s$).

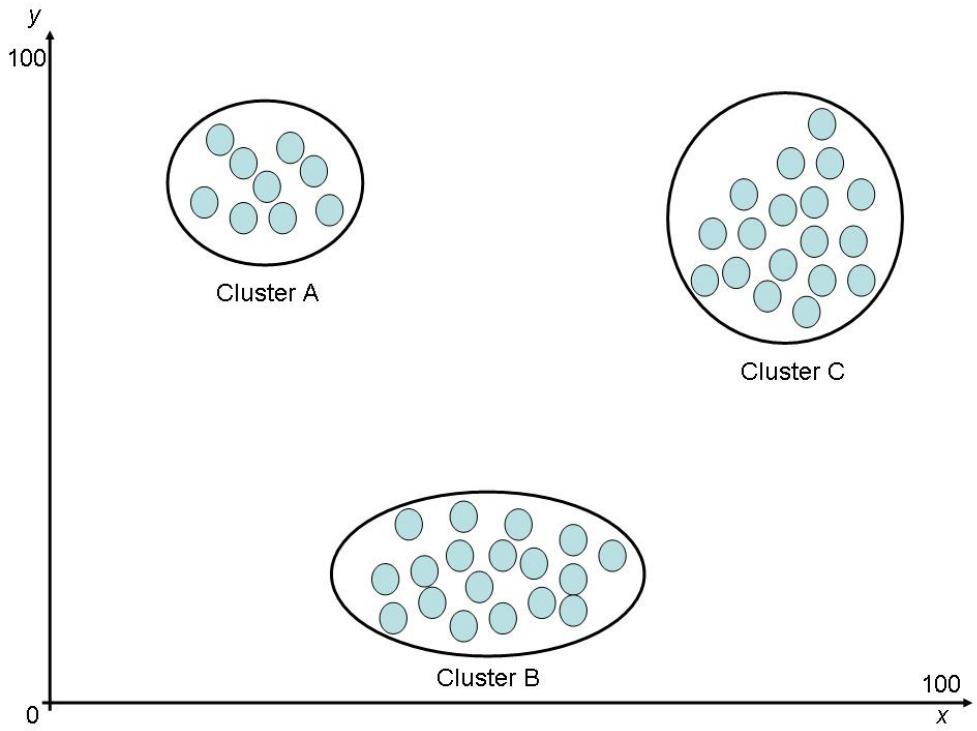


Figura 27 – *Clusters* distribuídos em 2 dimensões com alta coesão e baixo acoplamento.

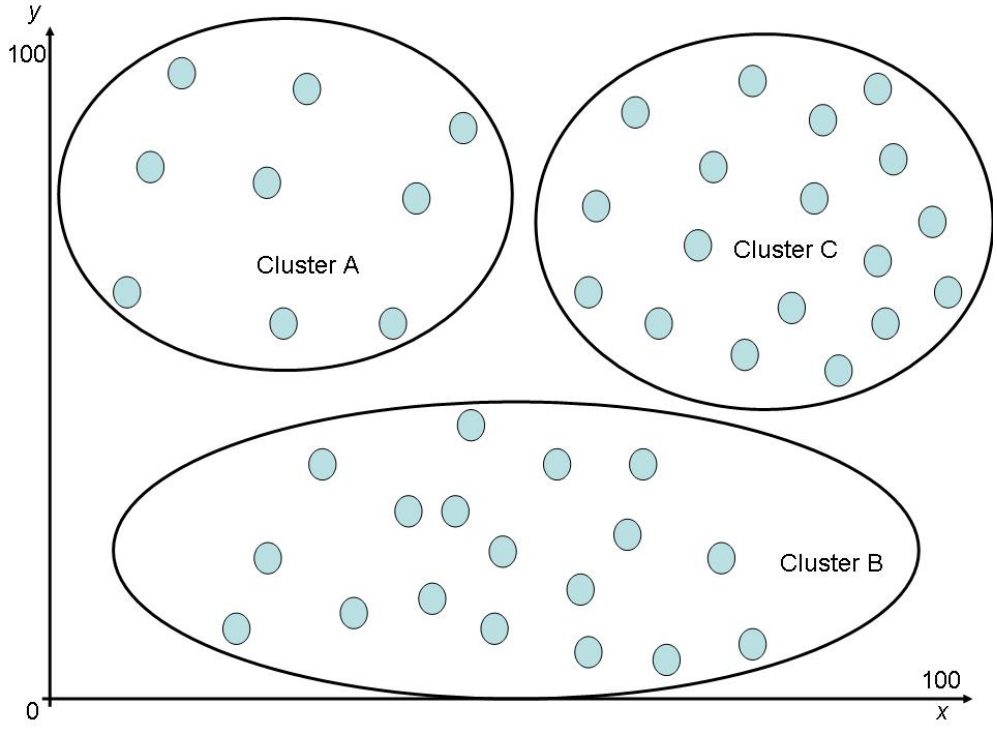


Figura 28 – *Clusters* distribuídos em 2 dimensões com baixa coesão e alto acoplamento.

Uma visão da ferramenta geradora de assinaturas está na Figura 29, onde a interface com o usuário coleta informações relativas à criação de um conjunto de assinaturas. O gerador de "sementes" faz uma distribuição de pontos onde os *clusters* serão gerados, o módulo criador de assinaturas identifica os locais onde cada semente foi criada e gera assinaturas próximas a cada uma destas sementes e o modulo de saída é encarregado de criar um arquivo seguindo o padrão CSV, adotado pela ferramenta de classificação. No decorrer serão apresentados detalhes sobre cada módulo desenvolvido.

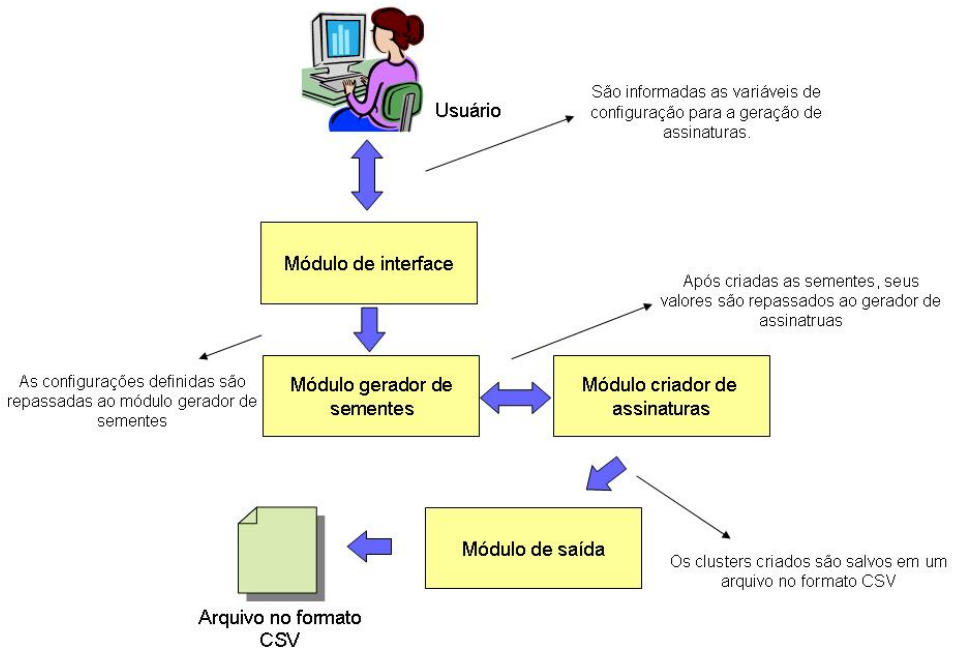


Figura 29 – Visão geral dos módulos do gerador de assinaturas

### Interface com o usuário

Este módulo possibilita que um usuário configure como serão criadas as assinaturas, possibilitando o funcionamento dos demais módulos.

### Módulo gerador de "sementes"

Depois de informados os valores de configuração, o gerador de assinaturas cria um conjunto de "sementes" de acordo com a variável $D_s$: quanto maior seu valor maior a probabilidade das sementes ficarem mais distantes entre si, quanto menores, mais próximas estarão as sementes. Estas sementes são consideradas pontos de referência para a criação dos *clusters* de assinaturas, sendo diretamente proporcionais à quantidade de *clusters* a serem criados. Cabe ressaltar

que cada semente é criada de acordo com a variável $Q_d$, e todas as suas dimensões assumem o mesmo valor. Por exemplo, caso o valor para o posicionamento de uma semente seja $s = 2,345$, e a variável $Q_d = 3$, será criada uma semente de três dimensões onde cada dimensão assume o valor $s$.
Para fazer o posicionamento de sementes a seguinte equação foi utilizada:

$$V_s = D_r \times N_r \tag{5.1}$$

onde, $V_s$ é referente ao valor que será atribuído a cada dimensão de uma semente, $D_r$, informado por um usuário, representa a distância máxima entre as sementes mais distantes (necessária para determinar o nível de acoplamento dos *clusters* gerados) e $N_r$ um número randômico entre zero e um.
Supondo um exemplo onde $D_r = 80$, para a criação de uma semente, tem-se que seu menor valor possível ($V_{smin}$) para a criação de suas dimensões, será quando $N_r = 0$:

$$V_{smin} = D_r \times N_r$$

$$V_{smin} = 80 \times 0$$

$$V_{smin} = 0$$

Considerando o mesmo valor $D_r$, uma semente terá seu maior valor possível ($V_{smax}$) para a criação de suas dimensões quando $N_r = 1$:

$$V_{smax} = D_r \times N_r$$

$$V_{smax} = 80 \times 1$$

$$V_{smax} = 80$$

Interpretando os resultados obtidos no exemplo apresentado pelas variáveis $V_{smin}$ e $V_{smax}$ e considerando o caráter randômico da variável $N_r$, as sementes seriam distribuídas no intervalo entre $0$, como menor que um conjunto de dimensões de uma semente pode ter até $80$ que seria o maior valor que o conjunto de dimensões de uma semente no exemplo apresentado pode ter.
A utilização $N_r$ é feita de forma a garantir que nenhuma semente extrapole o valor máximo de $D_r$. Desta forma, infere-se que altos valores de $D_r$ probabilisticamente permitem que as sementes fiquem mais distantes entre si, enquanto baixos valores possibilitam maior proximidade entre as sementes.

Módulo gerador de assinaturas

Depois de criado o conjunto de sementes, inicia-se um processo de criação de assinaturas. Cada dimensão de uma assinatura é criada em relação a uma semente específica, escolhida aleatoriamente, visando garantir que a assinatura não fique distante da semente usada como base para a sua criação e possibilitando a delimitação de *clusters* distintos.
De forma a garantir a proximidade entre as assinaturas geradas com uma semente em questão, a seguinte equação foi utilizada:

$$D_i = (V_s + (D_a \times N_f)) \tag{5.2}$$

Onde $D_i$ representa o valor que cada dimensão de uma assinatura irá receber, sendo $i$ variando entre 1 e $Q_d$, $V_s$ o valor da dimensão da semente que servirá de base para a dimensão da assinatura a ser gerada, $D_a$ um valor informado por um usuário de forma a delimitar o quanto cada dimensão pode variar sobre $V_s$, possibilitando inferir a coesão dos *clusters* formados e por fim $N_r$ um número randômico no intervalo de zero a um.
Esta equação determina cada dimensão de uma assinatura, assim como garante que cada assinatura gerada forme *clusters* próximos as suas sementes de referência.
Considerando $V_s = 40$ e $D_a = 10$, será apresentado um exemplo onde será criada uma assinaturas composta de duas dimensões.

Considerando a primeira dimensão de uma assinatura a ser criada, sendo o valor $N_f = 0,2$, tem-se:

$$D_{i2} = (40 + (10 \times 0.2))$$

$$D_{i1} = 42$$

Considerando a segunda dimensão de uma assinatura a ser criada, sendo o valor $N_f = 0,8$, tem-se:

$$D_{i2} = (40 + (10 \times 0.8))$$

$$D_{i2} = 48$$

Interpretando-se os valores de $D_{i1}$ e $D_{i2}$, a assinatura deste exemplo teria respectivamente os valores 42 e 48. Considerando que esta assinatura fosse colocada em um plano em duas dimensões, seu valor $x = 42$ e seu valor $y = 48$. A variável $D_a$ torna possível controlar o grau de coesão dos *clusters* formados. Caso seu valor seja elevado, a tendência é que a coesão dos *clusters* formados seja baixa, caso este valor seja baixo, a tendência é que maior será a coesão dos *clusters* formados.

Módulo de saída

O objetivo do módulo de saída é coletar os resultados obtidos pelo gerenciador de assinaturas colocando-os em um arquivo no formato CSV para a serem interpretados pelo atribuidor de notas. Este irá automaticamente classificar em bons ou ruins os grupos de assinaturas criados. Um exemplo da estrutura deste arquivo já foi apresentado na seção 4.3.3.

### 5.1.2 Atribuidor de notas

O objetivo do atribuidor de notas é substituir um usuário no processo de classificar um documento em bom ou ruim, necessário para identificar e delimitar os grupos de *clusters* existentes em um arquivo de assinaturas gerado.
Desta forma o atribuidor de notas classifica (bom ou ruim) aleatoriamente as assinaturas tendo como premissa que cada *cluster* gerado venha a possuir um conjunto de assinaturas de mesma classificação.
As classificações dos *clusters* de assinaturas são feitas de forma aleatória, considerando que no mínimo existam dois *clusters* de classificações distintas, um bom e um ruim.
A vantagem de possuir este conjunto de assinaturas pré-classificadas é eliminar a intervenção de um usuário na classificação de um documento, já que no momento que a ferramenta de classificação estética precisar saber se um documento é bom ou ruim não se torna necessário um usuário para fornecer tal classificação.

## 5.2 Conjuntos assinaturas de teste geradas

O objetivo da ferramenta desenvolvida é verificar a possibilidade de redução da intervenção humana no processo de classificação estética de documentos, hoje realizado por um usuário que verifica um a um a qualidade de cada documento gerado. Desta forma, foi criada uma variedade de casos de teste para verificar o quanto seria possível, reduzir a intervenção humana em tal processo.
Os conjuntos de assinaturas de teste geradas possuem as seguintes variações:

- Quantidade de *clusters* a serem gerados, entre 2 a 10 *clusters*;

Tabela 2 – Documentos de teste gerados, considerando *clusters* e assinaturas

| *Clusters* | Assinaturas | | | | |
|---|---|---|---|---|---|
| | 100A | 300A | 500A | 700A | 900A |
| 1C | X | X | X | X | X |
| 2C | X | X | X | X | X |
| 3C | X | X | X | X | X |
| 4C | X | X | X | X | X |
| 6C | X | X | X | X | X |
| 7C | X | X | X | X | X |
| 8C | X | X | X | X | X |
| 9C | X | X | X | X | X |
| 10C | X | X | X | X | X |

- Quantidade de assinaturas a serem geradas assume valores com 100, 300, 500, 700 ou 900 assinaturas;
- A quantidade de dimensões que cada assinatura irá ter, variando de 2 a 6 dimensões;
- Graus de coesão e acoplamento, que neste trabalho foram divididos em Valores A, Valores B, Valores C e Valores D.

Para a realização dos experimentos, o conjunto de arquivos gerados foi dividido de acordo com os quatro grupos que determinam a variação de coesão e acoplamento das assinaturas geradas. Cada um destes grupos é composto por assinaturas que variam entre um total de sete tipos de possíveis dimensões, dentro de cada uma destas dimensões ainda existem ainda 45 arquivos que variam combinando os valores de 3 a 10 *clusters* e 100, 300, 500, 700 e 900 assinaturas por arquivo gerado, totalizando 1260 arquivos de teste.

Os arquivos gerados para a realização dos testes podem ser descritos apoiando-se na Tabela 2, onde *clusters* representam o número de *clusters* a serem criados (2C dois *clusters*, 3C tres *clusters* e assim por diante), assinaturas representam o número de assinaturas de um arquivo (100A representam 100 assinaturas, 300A trecentas assinaturas e assim por diante) e onde com um **X** são marcados os arquivos que foram criados de acordo com esta relação entre *clusters* e assinaturas.
Por exemplo, a marcação do **X** existente na linha 3C coluna 300A, significa que um arquivo foi criado com 300 assinaturas divididas em 3 *clusters*.

Considerando que todas as possíveis variações em número de assinaturas e número de *clusters* sejam representados na Tabela 2, assumam o valor $CxA$. Cada conjunto $CxA$ pode ser variado de acordo com valores de coesão e acoplamento, bom como do número de dimensões

Tabela 3 – Documentos de teste gerados, considerando Coesão, Acomplamento e Dimensões

| Coesão e Acomplamento | Dimensões | | | |
|---|---|---|---|---|
| | 3D | 4D | 5D | 6D |
| Valores A | $CxA$ | $CxA$ | $CxA$ | $CxA$ |
| Valores B | $CxA$ | $CxA$ | $CxA$ | $CxA$ |
| Valores C | $CxA$ | $CxA$ | $CxA$ | $CxA$ |
| Valores D | $CxA$ | $CxA$ | $CxA$ | $CxA$ |

que cada assinatura de um arquivo venha a possuir.
Nas variações de coesão e acoplamento, considera-se que os Valores A, possuam as assinaturas com maior valor de coesão e menor valor de acoplamento. Os valores de coesão vão gradativamente diminuindo enquanto o acoplamento gradativamente aumenta até o conjunto de assinaturas presentes nos valores D, que possuem menor coesão e maior acoplamento entre os demais. Em relação as variações no número de dimensões, as assinturas podem ter de 3 até 6 dimensões.
A Tabela 3, apresenta as variações de coesão e acoplamento e número de dimensões que um conjunto $CxA$ pode possuir.

Como forma de interpretação da Tabela 3, se buscarmos a linha Valores D coluna 5D, os conjuntos $CxA$ irão possuir os graus de coesão e acoplamento existente nos Valores D, além de suas assinaturas possuírem 5 dimensões.
Para apresentar a distinção entre os valores de coesão e acoplamento que o conjunto de assinaturas geradas possam ter, um arquivo extraido do conjunto $CxA$ possuindo 2 *clusters*, 300 assinaturas e três dimensões foi usado de exemplo. Quando este arquivo aparece nos Valores A, a disposição de suas assinaturas pode ser vista na Figura 30, quando aparece nos Valores B, Figura 31, nos Valores C Figura 32 e finalmente nos valores D Figura 33.

## 5.3 Preparação dos experimentos

O experimento a ser realizado tem o objetivo de medir a quantidade de perguntas (número de interações requisitadas a um usuário, no qual, é informado se o documento é bom ou ruim) realizadas quando os *clusters* de documentos partem de altos valores de coesão e baixos valores de acoplamento para baixos valores de coesão e altos valores de acoplamento. Com isto pretende-se medir o número de intervenções que seriam realizadas a um usuário em cada um destes casos. Para mensurar o número de intervenções realizadas, é computada como realizada uma intervenção a cada vez que é verificada a classificação de um documento.

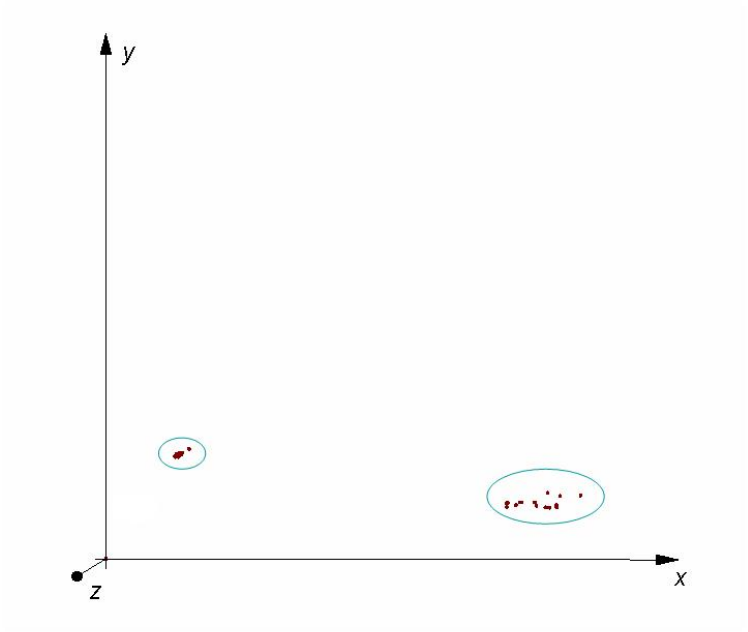


Figura 30 – 2 *clusters*, 300 assinaturas, Valores A, 3 dimensões

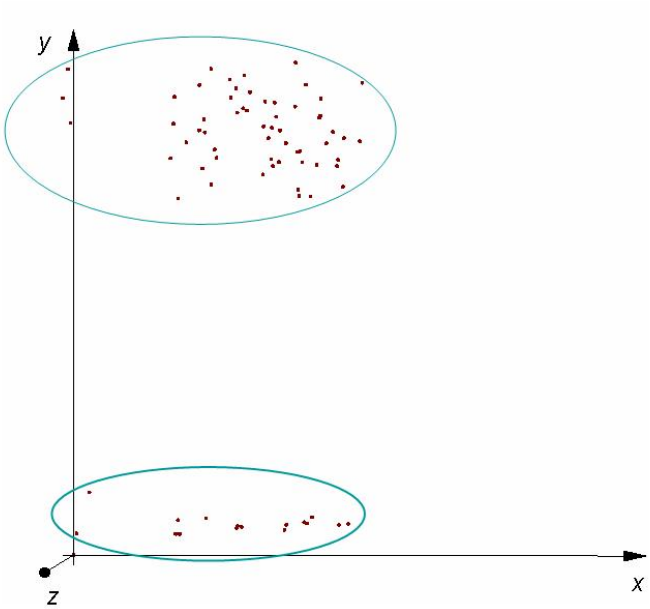


Figura 31 – 2 *clusters*, 300 assinaturas, Valores B, 3 dimensões

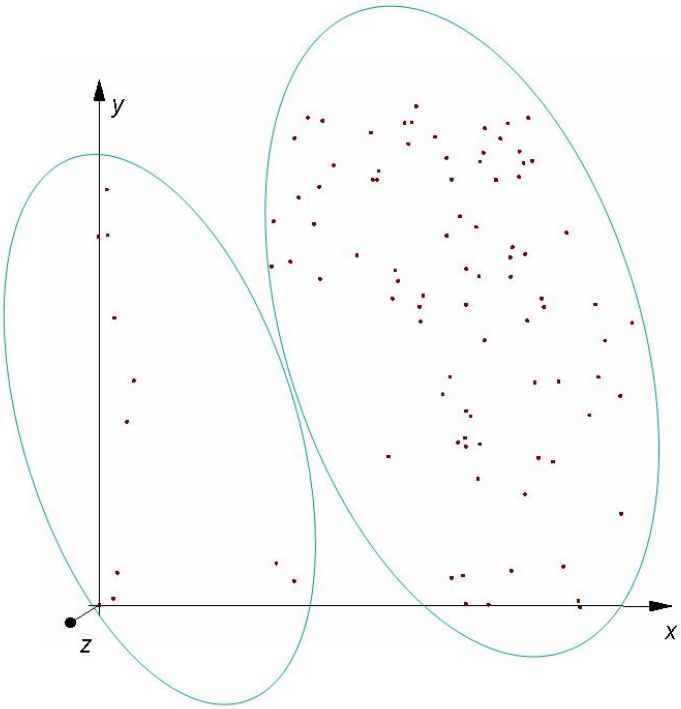


Figura 32 – 2 *clusters*, 300 assinaturas, Valores C, 3 dimensões

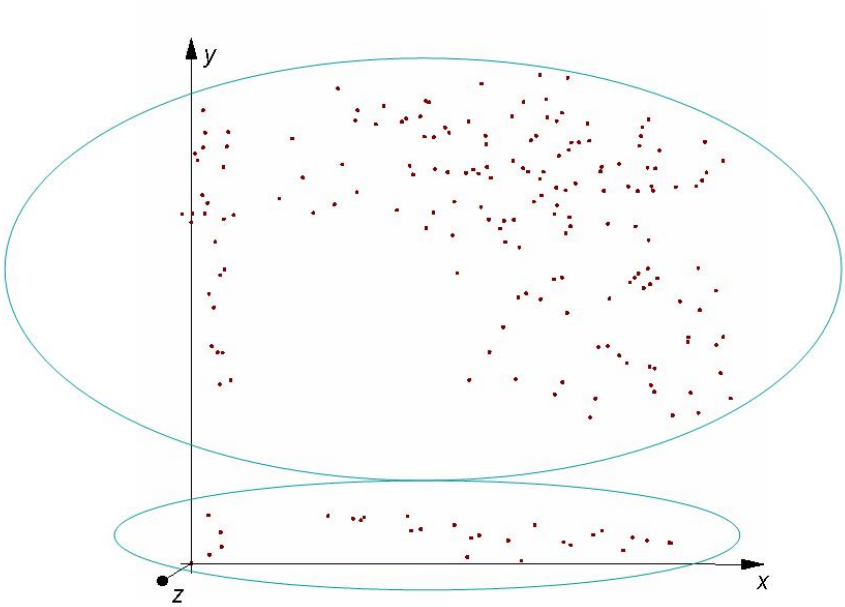


Figura 33 – 2 *clusters*, 300 assinaturas, Valores D, 3 dimensões

Desta forma a idéia é realizar experimentos com casos mais simples, por exemplo, um conjunto de 100 documentos divididos em 2 *clusters* com alto valor de coesão e baixo acoplamento até casos mais complexos que envolvam 700 documentos divididos em 8 *clusters* com baixos valores de coesão e altos valores de acoplamento. Os resultados irão indicar em que casos a ferramenta e os métodos de clusterização utilizados obtiveram melhores resultados.
Também será analisado em quantas oportunidades o algoritmo de clusterização não conseguiu completar uma execução. Uma execução é considerada completa, quando o número de intervenções é inferior ao número de assinaturas de um arquivo sendo testado. Por exemplo, se um arquivo com 100 assinaturas está sendo testado, caso ocorram 100 intervenções, seria o mesmo que um usuário verificar um a um os documentos para classificá-los em bons ou ruins, problema que motivou a elaboração deste trabalho.
Para garantir resultados mais consistentes, cada arquivo de assinaturas foi executado vinte vezes pela ferramenta de classificação e uma média de intervenções foi extraída, bem como, quantas execuções dentre as vinte realizadas foram completadas com sucesso. A opção de vinte execuções para cada arquivo foi feita devido a um problema identificado no algoritmo de clusterização implementado neste trabalho.
Para apresentar este problema, $22$ objetos foram distribuídos em um plano de forma a caracterizarem a formação de dois *clusters* distintos, como pode ser visto na Figura 34, delimitados por retângulos. Como forma de apresentar o problema, o algoritmo de K-Medóides foi executado informando $K = 2$, para o conjunto de objetos. proposto, de forma a verificar como este algoritmo delimitaria este conjunto de objetos refletindo a Figura 34.

Na Figura 35, são apresentados quatro resultados obtidos com este teste. Como pode ser observado, este algoritmo irá produzir resultados que não condizem com o esperado (caso 2 e caso 3), resultados próximos ao esperado (caso 1) e casos onde o resultado esperado foi alcançado (caso 4).

Devido a este fato, o algoritmo de K-Medóides pode levar a resultados muito ruins em relação à formação de *clusters* em um processo de classificação estética. Devido foi optado pela realização de 20 execuções sobre cada arquivo de teste. A escolha deste número se deu considerando o tempo a ser gasto considerando o volume de arquivos a serem testados.
Cabe ressaltar que a ferramenta de classificação, possui como abordado no capítulo 3, duas técnicas para trabalhar com o algoritmo de clusterização implementado. Seriam elas:

- sem realimentação que funciona com os princípios básicos do algoritmo de K-Medóides tradicional, ou seja, a ferramenta de classificação quando necessita quebrar um conjunto de documentos em mais um *cluster*, apenas informa um valor $K$ maior.

- com realimentação que além de informar a necessidade de um $K$ maior, repassar uma

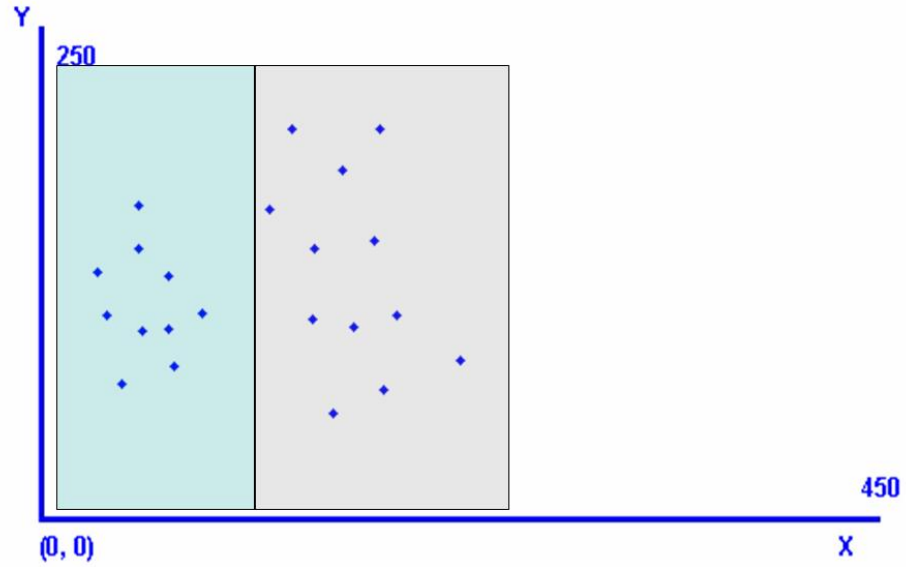


Figura 34 – 22 objetos distribuídos em um plano em duas dimensões distribuídos em dois *clusters*

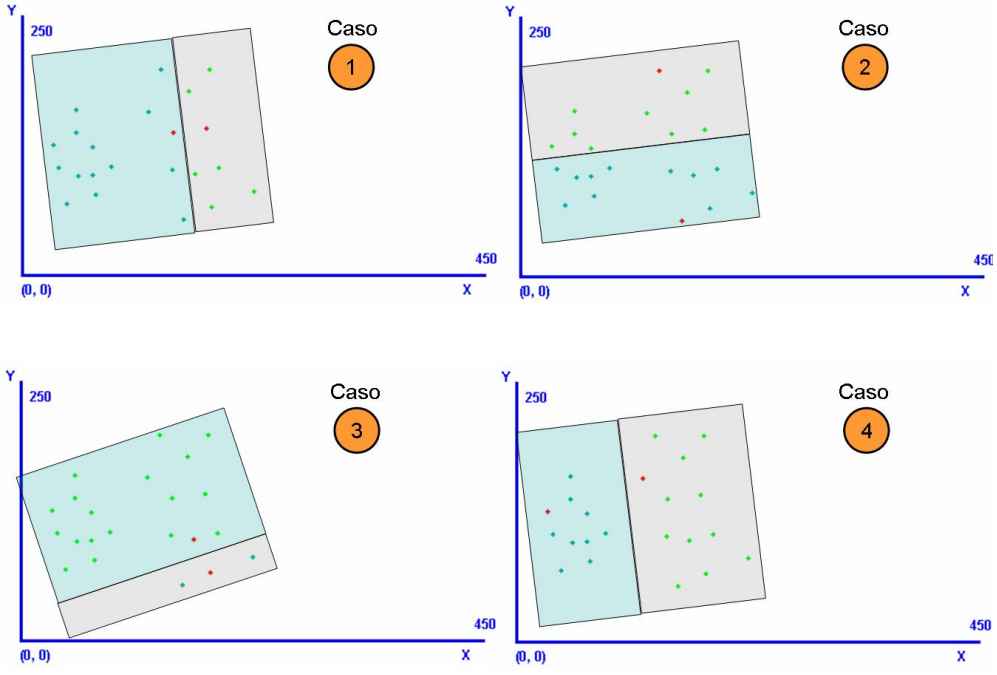


Figura 35 – Quatro execuções do algoritmo de K-Medóides sobre 22 objetos

dica ao algoritmo de K-Medóides relativa a uma região onde este pode vir a procurar por novos *clusters* a serem formados.

Estas técnicas são testadas com os de arquivos de teste gerados, visando compará-las entre si, obtendo resultados referentes a sua qualidade para o processo de classificação estética.

## 5.4 Resultados esperados

Após executados os experimentos é esperado que a ferramenta de classificação atinja seu objetivo de reduzir a intervenção humana no processo de classificação de documentos em bons ou ruins.
Também é esperado, devido aos diferentes graus de coesão e acoplamento, para o conjunto de assinaturas formados de acordo com os Valores A não sejam encontradas muitas dificuldades em identificar e classificar o conjunto de assinaturas, realizando poucas intervenções e um grande número de execuções completas. Nas assinaturas formadas pelos Valores B, é esperado que a ferramenta de classificação estética encontre poucas dificuldades em identificar e classificar de assinaturas deste grupo, possuindo um pequeno aumento nas intervenções e um número de execuções completas levemente inferiores as assinaturas existentes nos Valores A. Nas assinaturas formadas pelos Valores C é esperado que a ferramenta de classificação estética de documentos encontre dificuldades em identificar e classificar as assinaturas deste grupo, possuindo aumento nas intervenções e um número de execuções completas inferior aos resultados obtidos com as assinaturas existentes no grupo dos Valores B. Nas assinaturas formadas pelos valores D é esperado que a ferramenta de classificação estética de documentos encontre grandes dificuldades em identificar e classificar estas assinaturas, possuindo aumento significativo nas intervenções e um número de execuções completas baixo as assinaturas demais assinaturas.

# 6 Experimentos e Resultados

O objetivo deste capítulo é apresentar os experimentos realizados para verificar o funcionamento da ferramenta de classificação proposta.
Para isto, foram feitos experimentos com os dois métodos implementados neste trabalho, o método com realimentação, que para os experimentos realizados será chamado de **clust** e o método sem realimentação que para os experimentos realizados será chamado de **clust2**. Os arquivos de teste usados nos experimentos realizados, foram descritos previamente na seção 5.2, sendo divididos em Valores A, B, C e D.
Os experimentos visam mensurar qual técnica irá obter o melhor conjunto de resultados considerando as variáveis de teste que serão apresentadas no decorrer do texto.
Cabe lembrar a descrição de alguns termos utilizados nestes experimentos: Uma **intervenção** é considerada quando a ferramenta de classificação estética busca a classificação em bom ou ruim de um **arquivo classificado**. Cada **arquivo** é composto por diversas **assinaturas** sendo que cada uma destas representa um documento.
É dito um **sucesso**, quando um arquivo pode ser classificado pela ferramenta de classificação estética de documentos, ou seja, foi capás eliminar o trabalho de um usuário de verificar um a um a classificação de um conjunto de documentos.

## 6.1 Organização dos experimentos realizados

Os arquivos de teste usados no experimento realizado são aqueles abordados no capítulo 5 seção 5.2, Valores A, B, C e D. Estes valores, são divididos em quatro experimentos.
Nos Experimentos A, seus resultados são medidos de acordo com o conjunto de Valores A, este conjunto é caracterizado por possuir assinaturas consideradas com os maiores valores de coesão e menores valores de acoplamento, sendo assim, é esperado um número menor de intervenções no processo de classificação de documentos, bem como, um elevado número de arquivos classificados.
Nos Experimentos B foi utilizado o conjunto de Valores B, espera-se um número maior de intervenções quando comparados aos resultados dos Experimentos A, bem como um número menor de arquivos classificados.
Nos Experimentos C foi utilizado o conjunto de Valores C, é esperado que os resultados obtidos

sejam de um maior número de intervenções se comparados aos resultados do experimento B, bem como um número menor de arquivos classificados.
E por fim, nos Experimentos D foi utilizado o conjunto de Valores D, estes os quais, considerados o conjunto de assinaturas com os menores valores de coesão e maiores valores de acoplamento, são esperados que sejam obtidos os maiores valores no número de intervenções, bem como, um reduzido número de arquivos classificados.
De maneira a comparar os resultados de cada um dos experimentos realizados, as avaliações que serem realizadas são as seguintes:

- Intervenções/assinaturas
- Execuções/arquivo
- Total de arquivos classificados

Cada uma destas avaliações é realizada para cada experimento, por exemplo, no Experimento X, será feita uma análise de qual método obteve melhores resultados considerando os pontos que serão avaliados, bem como, serão apresentados seus resultados em um experimento de acordo com cada método testado para cada conjuntos de arquivos de dimensões específicas, por exemplo, identificar no Experimento X, qual método obteve melhores resultados para um conjunto de assinaturas em 3, 4, 5 e 6 dimensões.

### 6.1.1 Intervenções/assinaturas

O objetivo desta medida é determinar quantas intervenções em média são necessárias para classificar um conjunto de arquivos.
As explicações desta seção consideram um experimento em específico, onde todas as assinaturas possuem valores similares de coesão e acoplamento. Para determinar esta relação de intervenções/assinaturas, uma série de etapas foi realizada.
A primeira etapa consiste em calcular a relação intervenções/assinaturas para um arquivo classificado pela ferramenta de classificação. Esta etapa é realizada com o apoio da seguinte equação:

$$Mi = \frac{Qint}{Qass} \tag{6.1}$$

Onde $Mi$ refere-se à quantidade de intervenções necessárias para classificar um arquivo específico. A variável $Qint$ representa a quantidade de intervenções de um arquivo em específico, considerando que um arquivo pode ser executado $n$ vezes, sendo este valor informado por

um usuário, é feita uma média de intervenções realizadas para representá-la. A variável $Qass$ refere-se à quantidade de assinaturas de um arquivo em específico.
Supondo que o arquivo 2c300a3D (arquivo com 300 assinaturas divididas em 2 clusters pertencente a um grupo de assinaturas que possuem 3 dimensões), dentre $n$ execuções foram necessárias em média 15 intervenções para ser classificado. Desta forma $Qint = 15$, como este arquivo possui 300 assinaturas, $Qass = 300$. O valor de $Mi$ será:

$$Mi = \frac{15}{300}$$
$$Mi = 0.05$$

Devido a $Mi$ estar entre 0 e 1, $Mi = 0,05$ significa que foram necessárias 5% de intervenções para classificar o arquivo exemplo 2c300e3D.
Em uma segunda etapa é feita a soma de todos os resultados $Mi$ para um experimento em específico e então este valor é dividido pelo total de arquivos com um valor $Mi(Totc)$, de forma a determinar o valor de intervenções por assinaturas de um experimento em específico ($IA$).
Por exemplo, considerando que existem 5 arquivos que tiveram seus valores de $Mi$ respectivamente $0.15, 0.20, 0.25, 0.30, 0.35$, o resultado de $IA$ seria:

$$IA = \frac{(0.15 + 0.20 + 0.25 + 0.30 + 0.35)}{5}$$
$$IA = 0.25$$

Interpreta-se com $IA = 0,25$ , tendo este resultado a mesma lógica do cálculo de $Mi$ em relação a faixa de valores, que em média foram necessárias 25% de intervenções para classificar este conjunto de arquivos de teste.

### 6.1.2 Execuções/arquivo

O objetivo desta medida é determinar em média quantas execuções um determinado conjunto de arquivos realizou com sucesso para um número de execuções $n$ determinadas por um usuário.
Na primeira etapa, é feita uma relação de sucessos que um método teve na classificação de um arquivo em relação a um número máximo de sucessos determinado por um usuário.
Representada pela seguinte equação:

$$Mu = \frac{Qs}{Qe} \tag{6.2}$$

Onde $Mu$ refere-se à quantidade de sucessos de execução de um arquivo específico, sendo que seu valor pode variar entre 0, que indica que nenhuma execução foi completa com sucesso, até 1, que refere-se a todas as execuções completas com sucesso. A variável $Qs$ representa a quantidade de sucessos de um arquivo em específico, considerando que um arquivo pode ser executado $Qe$ vezes, sendo este valor definido por um usuário.
Supondo que o arquivo 2c300a3D (arquivo com 300 assinaturas dividas em 2 clusters pertencente a um grupo de assinaturas que possuem 3 dimensões), dentre 20 execuções máximas ( $Qe$=20 ), tivesse obtido sucesso em 12 dessas ($Qs = 12$) seu $Mu$ seria:

$$Mu = \frac{12}{20}$$
$$Mu = 0.6$$

Devido ao fato de $Mu$ estar em uma faixa de valores entre 0 e 1, considera-se a titulo de interpretação que $Mu = 0.6$ significa que foram feitas 60% das execuções determinadas por um usuário para o arquivo exemplo 2c300e3D.
Em uma segunda etapa é feito o somatório de todos os resultados $Mu$ para um experimento em específico.
Por exemplo, considerando que dentro de um conjunto de experimentos existem 5 arquivos, 2c300a3D, 3c600a3D, 4c300a3D, 5c900a3D, 6c700a3D, sendo seus valores de $Mu$ respectivamente $0.15, 0.10, 0.12, 0.13, 0.19$, a média de $Mu$ ($Mu_{med}$) para estes arquivos seria:

$$Mu_{med} = \frac{(0.15 + 0.10 + 0.12 + 0.13 + 0.5)}{5}$$
$$Mu_{med} = 0.2$$

Interpretando-se $Mu_{med} = 0.2$, diz-se que em média foram feitas 20% de execuções com sucesso.

### 6.1.3 Total de arquivos classificados

O percentual de arquivos classificados ($Pac$) tem o objetivo de medir o percentual de arquivos que no mínimo tiveram uma execução (sucesso), dentre um conjunto de arquivos a serem executados. Para o seu cálculo a seguinte equação é utilizada:

$$Pac = \frac{Na}{Ta} \tag{6.3}$$

Nesta equação, $Na$ representa o número de arquivos que tiveram no mínimo um sucesso,

Tabela 4 – Exemplo de resultados de execuções completas para clust

| Nome do arquivo | Sucessos |
|---|---|
| | clust |
| 2c100e | 14 |
| 2c300e | 1 |
| 2c500e | 0 |
| 2c700e | 15 |
| 2c900e | 16 |

ou seja, uma execução completa de um método testado sobre um conjunto de arquivos e $Ta$ representa o total possível de arquivos que podem ser classificados.

Por exemplo:
Considerando a tabela 4 como base para este exemplo, são apresentados um conjunto de valores na coluna Execuções Completas referentes ao número de execuções feitas para cada arquivo, representados pela coluna Nome do arquivo.

Na tabela 4, $Na = 4$, já que, observando o arquivo 2c500e, nenhum sucesso foi obtido. A variável $Ta = 5$ pois o máximo de arquivos que poderiam ser classificados eram cinco.
Desta forma:

$$Pac = \frac{4}{5}$$

$$Pac = 0,8$$

Interpretando este resultado, observa-se que o valor de $Pac$ para este exemplo representa 80% de arquivos classificados dentre os 5 possíveis.

### 6.1.4 Critérios para escolha do melhor método

Como forma de determinar o melhor método, os experimentos realizados priorizam o total de arquivos classificados como sendo o primeiro critério de escolha entre os métodos, seguido pela relação intervenções/assinaturas e por fim a relação execuções/arquivo.

Tabela 5 – Conjunto de resultados para o conjunto de documentos gerados nos Valores A

| | Dimensões, Técnicas e Valores | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Metodologia | 3D | | 4D | | 5D | | 6D | |
| | clust | clust2 | clust | clust2 | clust | clust2 | clust | clust2 |
| $IA$ | 48,25% | 55,18% | 48,00% | 43,72% | 52,23% | 52,39% | 60,55% | 57,53% |
| $Mu_{med}$ | 41,13% | 47,89% | 41,71% | 36,40% | 42,80% | 42,91% | 48,54% | 49,20% |
| $Pac$ | 86,66% | 100% | 91,11% | 95,55% | 91,11% | 95,55% | 91,11% | 97,77% |

## 6.2 Experimento A

Considerando as variáveis a serem avaliadas, a Tabela 5 apresenta os resultados dos experimentos para o conjunto dos Valores A.

### 6.2.1 Avaliação dos resultados dos arquivos de 3 dimensões

Fazendo uma avaliação sobre este conjunto de testes, percebe-se uma vantagem do método clust na relação intervenções/assinaturas, obtendo um resultado de 48,25% contra 55,18% de clust2. Embora seu percentual de arquivos executados tenha sido alto, considerando um grande número de arquivos classificados, cerca de 86,66% do total, clust2 ainda leva vantagem, já que classificou 100% dos arquivos a serem classificados.

### 6.2.2 Avaliação dos resultados dos arquivos de 4 dimensões

Avaliando o conjunto de arquivos deste experimento, o método clust, teve um aumento em relação ao número de arquivos classificados, passando de 86,66% para 91,11%, embora, não tenha executado o mesmo volume de arquivos que clust2, que teve uma redução para 95,55% em relação ao conjunto de testes em 3 dimensões.
Na relação intervenções/assinaturas, clust2 teve uma melhora significativa, passando de 55,18% obtidos no conjunto de arquivos de testes com três dimensões para 43,72%, obtendo melhores resultados que clust, que ficou em 41,71%.
O metodo clust levou vantagem na relação execuções/arquivo, onde ficou com 41,71% contra 36,40% de clust2.

### 6.2.3 Avaliação dos resultados dos arquivos de 5 dimensões

Avaliando o conjunto de arquivos deste experimento, a diferença entre ambos os métodos foi pequena, a vantagem de clust2, foi considerando o total de arquivos classificados que ficou em 95,55% dos arquivos de teste, contra 91,11% de clust.

### 6.2.4 Avaliação dos resultados dos arquivos de 6 dimensões

Avaliando o conjunto de arquivos deste experimento, percebe-se que os resultados não foram muito distintos, o diferencia a favor de clust2, foi o fato de possuir uma melhor relação intervenções/assinaturas, 57,53% contra 60,55% de clust2.
Outro fator a favor de clust2, seria o fato de obter um total de arquivos classificados bastante alto, 97,77% contra 91,11% de clust.

### 6.2.5 Avaliação geral

Considerando os experimentos B, percebe-se uma vantagem de clust2 considerando a média geral dos testes realizados.
Cabe também ressaltar a qualidade de clust2 neste experimento considerando o número de arquivos classificados, onde seu menor valor foi 95,55% (4 e 5 dimensões), caracterizando uma vantagem quando se considera uma redução na intervenção humana.
O ponto negativo para ambos os métodos foi o baixo desempenho na relação execuções por arquivo, onde ambos os métodos executaram em torno de 50% do total possível de execuções.

## 6.3 Experimento B

Considerando as variáveis a serem avaliadas, a Tabela 6 apresenta os resultados dos experimentos para o conjunto dos Valores B.

Tabela 6 – Conjunto de resultados para o conjunto de documentos gerados nos Valores B

| | Dimensões, Técnicas e Valores | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Metodologia | 3D | | 4D | | 5D | | 6D | |
| | clust | clust2 | clust | clust2 | clust | clust2 | clust | clust2 |
| $IA$ | 43,10% | 36,20% | 63,83% | 48,33% | 54,24% | 46,18% | 52,02% | 50,11% |
| $Mu_{med}$ | 33,71% | 28,26% | 51,33% | 40,57% | 45,31% | 38,69% | 43,08% | 41,25% |
| $Pac$ | 77,77% | 95,55% | 64,44% | 77,77% | 71,11% | 93,33% | 86,66% | 97,77% |

### 6.3.1 Avaliação dos resultados dos arquivos de 3 dimensões

Avaliando o conjunto de arquivos deste experimento, apensar de clust ter obtido uma melhor relação execuções/arquivo 33,71% contra 28,26% de clust2, não garantiu uma melhor relação intervenções/assinatura, ficando com 43,10% contra 36,20% de clust2. Considerando o total de arquivos executados, clust ficou com 77,77%, contra 95,55% de clust2.

### 6.3.2 Avaliação dos resultados dos arquivos de 4 dimensões

Avaliando o conjunto de arquivos deste experimento, ambos os métodos tiveram dificuldades quando se refere ao total de arquivos executados.
Fazendo um comparativo entre ambos os métodos, a vantagem de clust2 ficou na relação intervenções/assinaturas, onde seu resultado foi de 48,33% contra 63,83% de clust, embora clust2 não tenha conseguido classificar grande quantidade de documentos, 77,77%, ainda sim, levou vantagem sobre clust, que teve 64,44%.

### 6.3.3 Avaliação dos resultados dos arquivos de 5 dimensões

Avaliando o conjunto de arquivos deste experimento, o método clust2, levou novamente vantagem sobre clust, um total de arquivos classificados em torno de 93,33% contra 71,11% de clust, e ainda possuindo uma relação intervenções/assinaturas em torno de 46,18% contra 54,24% para clust2, obtendo assim, melhores resultados.

Tabela 7 – Conjunto de resultados para o conjunto de documentos gerados nos Valores C

| | Dimensões, Técnicas e Valores | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Metodologia | 3D | | 4D | | 5D | | 6D | |
| | clust | clust2 | clust | clust2 | clust | clust2 | clust | clust2 |
| $IA$ | 37,51% | 24,85% | 58,21% | 33,03% | 43,58% | 29,05% | 46,88% | 34,64% |
| $Mu_{med}$ | 30,00% | 18,89% | 52,93% | 27,76% | 37,14% | 21,47% | 40,28% | 28,38% |
| $Pac$ | 57,77% | 80% | 64,44% | 84,44% | 60% | 75,55% | 80% | 75,55% |

### 6.3.4 Avaliação dos resultados dos arquivos de 6 dimensões

Avaliando o conjunto de arquivos deste experimento, percebe-se que clust2 leva vantagem novamente no número de arquivos classificados, em torno de 97,77%, embora clust também tenha conseguido um valor alto, 86,66%, os demais resultados foram bastante semelhantes, como o critério de desempate adotado é verificar o total de arquivos classificados, a vantagem ficou com clust2.

### 6.3.5 Avaliação geral

Fazendo uma avaliação geral dos resultados deste experimento, percebe-se que conforme aumentaram o número de dimensões para este conjunto de testes, maiores se tornaram as relações de intervenções por assinatura.
O método clust2 leva vantagem sobre clust, já que obteve os melhores resultados da relação intervenções/assinaturas, bem como resultados bastante superiores a clust, considerando o total de arquivos classificados.
O ponto negativo continua sendo a relação execuções/arquivo, onde o melhor resultado foi obtido por clust, no conjunto de arquivos em 4d.

## 6.4 Experimento C

Considerando as variáveis a serem avaliadas, a Tabela 7 apresenta os resultados dos experimentos para o conjunto dos Valores C.

### 6.4.1 Avaliação dos resultados dos arquivos de 3 dimensões

Avaliando o conjunto de arquivos deste experimento, o método clust2, teve quase metade do valor de clust na relação que considera execuções/arquivo, sendo 30% de clust, contra 18,89% de clust2. Por outro lado, a relação de intervenções/assinaturas ficou em torno de 24,85% para clust2 contra 37,51% para clust. No total de arquivos classificados, clust2 teve ampla vantagem, chegando a 80% contra baixos 57,77% de clust.
Embora tenha sido baixa esta relação execuções/arquivo, as demais avaliações apontam vantagem para o método clust2.

### 6.4.2 Avaliação dos resultados dos arquivos de 4 dimensões

Avaliando o conjunto de arquivos deste experimento percebe-se uma diferença significativa na relação execuções/arquivo, considerando que clust teve 52,93% contra 27,77% obtidos por clust2.
Considerando o total de arquivos classificados a vantagem ainda foi de clust2, que teve um total de 84,44% de arquivos classificados, contra 64,44% de clust, além da relação intervenções assinaturas ter sido melhor, 33,03% em relação a 58,21% de clust.

### 6.4.3 Avaliação dos resultados dos arquivos de 5 dimensões

Avaliando o conjunto de arquivos deste experimento, as diferenças entre os comparativos entre clust e clust2 não diferem muito em relação aos seus resultados. Merece destaque que a relação execuções por arquivo, que até então, considerando os resultados de todos os experimentos, clust obtinha ampla vantagem, neste caso o resultado de clust2 se aproximou de clust, obtendo 21,45% contra 37,14%. Também percebe-se uma aproximação de clust considerando o total de arquivos a serem classificados, 60% de clust contra 75,55% de clust2. A grande vantagem de clust2 neste caso, foi na relação intervenções/assinaturas que ficou em 29,05% contra 43,58% de clust.

Tabela 8 – Conjunto de resultados para o conjunto de documentos gerados nos Valores D

| | Dimensões, Técnicas e Valores | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Metodologia | 3D | | 4D | | 5D | | 6D | |
| | clust | clust2 | clust | clust2 | clust | clust2 | clust | clust2 |
| $IA$ | 25,74% | 24,80% | 27,49% | 22,61% | 23,18% | 17,96% | 30,61% | 24,15% |
| $Mu_{med}$ | 19,46% | 20,16% | 23,41% | 18,28% | 17,65% | 13,87% | 23,25% | 18,04% |
| $Pac$ | 60% | 68,88% | 48,88% | 64,44% | 35,55% | 68,88% | 44,44% | 62,22 % |

### 6.4.4 Avaliação dos resultados dos arquivos de 6 dimensões

Avaliando o conjunto de arquivos deste experimento, pela primeira vez, o número de arquivos classificados de clust, foi superior a clust2, obtendo 80% contra 75,55%. O método clust confirmou também uma melhor relação execuções/arquivo, obtendo 40,28%, contra 28,38% de clust2. Apenas considerando a relação de intervenções/assinaturas clust2 obteve melhores resultados, tendo 34,64% contra 46,88%. Fato que não foi suficiente para garantir que clust2 tivesse melhores resultados com este conjunto de arquivos de teste.

### 6.4.5 Avaliação geral

Comparando os resultados deste experimento com os resultados do experimento B, percebe-se uma significativa redução na relação execuções por arquivo em ambos os métodos, em contra partida, percebe-se uma melhora na relação de intervenções/assinaturas em ambos os métodos. Embora o método clust tenha pela primeira obtido melhores resultados que clust2 no número de arquivos classificados considerando assinaturas de 6 dimensões, a avaliação avaliando os demais resultados clust2 ainda obteve melhores resultados.

## 6.5 Experimento D

Considerando as variáveis a serem avaliadas, a Tabela 8 apresenta os resultados dos experimentos para o conjunto dos Valores D.

### 6.5.1 Avaliação dos resultados dos arquivos de 3 dimensões

Avaliando o conjunto de arquivos deste experimento, os métodos clust e clust2, tiveram resultados bastante semelhantes, tendo baixos valores da relação execuções/arquivo, onde clust obteve 19,46% e clust2 20,16%, os resultados obtidos na relação intervenções/assinaturas, ambos tiveram resultados bastante bons, clust com 25,74% e clust2 24,80% sendo que a vantagem de clust2 tendo sido decidida pela total de arquivos classificados, ficando com 68,88%, contra 60% de clust.

### 6.5.2 Avaliação dos resultados dos arquivos de 4 dimensões

Avaliando o conjunto de arquivos deste experimento, considerando a relação execuções/arquivo clust leva vantagem com 23,41% contra 18,28% de clust2. Considerando a relação intervenções/assinaturas, as diferenças também não foram significativas entre ambos os métodos, clust ficou com 27,49%, clust2 com 22,61%. O detalhe que garantiu a vantagem de clust2, foi no total de arquivos classificados onde clust2 obteve 64,44% contra 48,88% de clust.

### 6.5.3 Avaliação dos resultados dos arquivos de 5 dimensões

Avaliando o conjunto de arquivos deste experimento os resultados obtidos nestes arquivos de teste, seguiram padrões similares ao dos experimentos com 4 dimensões. A diferença principal novamente no total de arquivos classificados, que no método clust2 ficou em 68,88% e no método clust 35,55%.

### 6.5.4 Avaliação dos resultados dos arquivos de 6 dimensões

Avaliando o conjunto de arquivos deste experimento Neste ultimo conjunto de arquivos testados, clust2 apresentou resultados levemente superiores a clust, sua vantagem também foi decidida no número de arquivos classificados que ficou em 62,22% contra 44,44% de clust.

Tabela 9 – Resultados gerais da comparação entre clust e clust2

| Experimento | Resultados gerais | | | | | |
|---|---|---|---|---|---|---|
| | clust | | | clust2 | | |
| | $IA$ | $Mu_{med}$ | $Pac$ | $IA$ | $Mu_{med}$ | $Pac$ |
| Experimento A | 2 | 2 | 0 | 2 | 2 | 4 |
| Experimento B | 0 | 4 | 0 | 2 | 0 | 4 |
| Experimento C | 0 | 4 | 1 | 4 | 0 | 3 |
| Experimento D | 0 | 3 | 0 | 4 | 1 | 4 |
| Totais | 2 | 13 | 1 | 12 | 3 | 15 |

### 6.5.5 Avaliação geral

Em resumo, estes experimentos apresentaram resultados interessantes já que quanto piores os valores da relação execuções arquivo, melhores foram os resultados da relação intervenções/assinaturas.
Cabe ressaltar que nos resultados para 5 dimensões a relação entre execuções/arquivo, atingiu seu menor valor de todos os experimentos realizados obtendo através do método clust2 o valor de 17,96%, em contra partida, a relação intervenções/assinaturas também foi a menor, onde o método clust2, também para o conjunto de resultados com 5 dimensões ficou com o valor de 13,87%.

## 6.6 Considerações finais

### 6.6.1 Considerações sobre clust e clust2

De forma a apresentar uma comparação entre clust e clust2, a Tabela 9, apresenta uma compilação dos resultados apresentados nos experimentos realizados.
Nesta Tabela, são computados os melhores resultados de cada método em cada experimento para cada medida avaliada, bem como um resultado geral que aponta o método com maior eficiência para as medidas avaliadas.

Com base na Tabela 9, clust2 apenas teve desvantagem na relação execuções/arquivo, apresentando melhores resultados nos demais itens avaliados. Dentre os métodos de avaliação utilizados, clust2 em conjunto com o algoritmos de K-Medóides levou vantagem em relação a clust. Os resultados dos experimentos realizados sobre cada arquivo de teste gerado para cada um dos

experimentos realizados podem ser vistos no Apêndice A.

### 6.6.2 Considerações sobre o uso do método K-Medóides

Considerando a relação execuções/arquivo, percebe-se que em todos os experimentos seus resultados não foram muitos elevados, desta forma, infere-se dificuldade dos métodos em classificar os conjuntos de arquivos de teste quando trabalham em conjunto ao algoritmo de K-Medóides, sendo esta dificuldade agravada à medida que os valores de coesão diminuem e os valores de acoplamento aumentam.

Por outro lado o uso do algoritmo de K-Medóides possibilitou bons resultados de acordo com o aumento do acoplamento e redução da coesão, percebendo-se que a relação execuções/arquivo teve uma significativa redução, inferindo-se que uma maior dificuldade em classificar um conjunto de arquivos possibilitou que um menor número de intervenções fossem realizadas para classificá-los.

Um resultado que era esperado, foi conforme a diminuição da coesão dos arquivos de teste e aumento de seu acoplamento, menores foram os totais de arquivos classificados.

Em resumo, o algoritmo de K-Medóides trabalhando em conjunto com o método clust2, obteve os melhores resultados para os experimentos realizados, possibilitando diminuir a intervenção humana no processo de classificação estética de documentos.

Pode não ser a melhor solução, já que, como observado, a relação entre o número de execuções/arquivo não foi considerada boa, ficando abaixo de 52,93%.

Esta baixa relação acaba reduzindo a credibilidade de uso de um sistema de classificação estética, já que, imaginando que em um processo de classificação estética tradicional onde um usuário geralmente irá fazer uma execução de um arquivo, ele tem uma certeza de no máximo 52,93% de que não vai precisar classificar um a um os documentos automaticamente gerados.

Cabe ressaltar que analisando uma média entre todas as intervenções/assinaturas de todos os experimentos executados para o método clust2, que foi aquele com melhores resultados, foram necessárias em torno de 37,54% de intervenções/assinaturas para que os conjuntos de testes fossem classificados.

# 7 Conclusões e trabalhos futuros

Neste trabalho foi abordada uma solução para o problema da classificação estética de documentos. Tais documentos quando gerados em larga escala escapam ao controle de um usuário, que passa a encontrar dificuldades em verificar um a um os documentos em busca de erros.
Como forma de facilitar tal processo, algumas das métricas que auxiliam a medir a qualidade estética de um documento foram descritas, assim como a forma que estas podem ser associadas a um documento, seja individualmente (escolhida para este trabalho) ou usando uma única nota para representar o documento em si.
Tais métricas servem para distinguir os documentos entre si, visando automatizar o processo de classificação. Esta identidade foi criada com o apoio de técnicas de *fingerprint*, que em sua origem servem para a criação da assinatura de um e-mail. Neste trabalho foram utilizadas como uma forma de identificar cada documento automaticamente gerado possibilitando um processo de clusterização.
Pelo vasto número de medidas de similaridade e algoritmos de clusterização existentes, foram escolhidos aqueles considerados mais populares na literatura pesquisada. desta forma foi feito o uso da distância euclidiana e do algoritmo de K-Medóides para serem utilizados com a ferramenta de classificação proposta.
O algoritmo de K-Medóides necessita que seja informado um número de clusters inicial ($K$) e isto é considerado difícil devido à imprevisibilidade das assinaturas dos documentos a serem classificados, por isso foram propostas também duas técnicas que automaticamente identificam este valor $K$, uma técnica sem realimentação de novos possíveis medóides e uma técnica que realimenta a ferramenta de classificação com possíveis novos medóides.
A ferramenta desenvolvida é constituída basicamente de cinco módulos: Módulo de execução, Módulo de entrada de dados, Módulo de avaliação, Módulo de clusterização e Módulo de visualização, que em conjunto visam a redução da intervenção humana no processo de classificação.
Para verificar o funcionamento da ferramenta de classificação, foram desenvolvidas duas ferramentas que de maneira controlada possibilitaram a criação de grupos de assinaturas com diferentes características de forma a verificar funcionalidade da ferramenta proposta.
Para o processo de validação da ferramenta, foi avaliada a relação entre intervenções/assinaturas, execuções/arquivo e total de arquivos executados como forma de verificar em média como que cada método, com ou sem realimentação, se comportou perante os experimentos realizados, constatando que o método com realimentação foi superior em grande parte das avaliações realizadas.
Com relação ao algoritmo de clusterização utilizado, apesar do algoritmo de K-Medóides ter ob-

tido bons resultados de acordo com o aumento do acoplamento e redução da coesão, percebeu-se também que a relação execuções/arquivo teve uma significativa redução, considerando um resultado geral dos experimentos realizados.

Em resumo, o algoritmo de K-Medóides trabalhando em conjunto com o método com realimentação obteve os melhores resultados conseguindo diminuir a intervenção humana no processo de classificação estética. Pode não ser a melhor solução, já que, como observado, a relação entre o número de execuções/arquivo não foi considerada boa em nenhum dos testes realizados.

Como trabalhos futuros acredita-se necessários novos experimentos com a ferramenta proposta sobre um conjunto de documentos reais, e com isto, analisar o desempenho quanto à redução da intervenção humana.

Outra possibilidade, considerando o algoritmo de clusterização implementado, seria buscar uma maneira mais eficiente quanto à escolha de um valor $K$ inicial. Uma alternativa para isso seria armazenar classificações de conjuntos de assinaturas já realizadas para posterior comparação com uma classificação futura. Isso seria possível se as assinaturas a serem classificadas tiverem características semelhantes a um conjunto previamente classificado, possibilitando a utilização dos valores $K$ de classificações anteriores.

Outra possibilidade a ser explorada seria testar diferentes formas de escolher as assinaturas que são perguntadas a um usuário quanto sua classificação, ou até mesmo, abrir outras possibilidades de classificações, além de bom ou ruim.

Modificações também podem ser feitas sobre o algoritmo de clusterização implementado, seja na função de similaridade utilizada, bem como testando diferentes algoritmos de clusterização existentes na bibliografia buscando alternativas para eliminar este valor $K$, ou substituí-lo por alternativas que possam ser mais intuitivas a um usuário de tais tipos de sistemas.

Por fim, estudar maneiras de adaptar a metodologia de classificação estética de documentos proposta a uma ferramenta de criação de documentos de conteúdo variável, visando trazer mais agilidade na criação e distribuição deste tipo de documento.

# Bibliografia

Alpaydin, E. (2004). Introduction to machine learning (adaptive computation and machine learning). Cambridge, USA: The MIT Press.

Arts, T. G. (2006). Variable data printing 2006: Growth and changes in the marketplace. New York, USA: The Industry Measure - Reed Business Information.

Balinsky, H., & Pilu, M. (2005). Emphasis for Highly customized documents. *DocEng '05: Proceedings of the 2005 ACM symposium on Document engineering* (pp. 30–30). New York, USA: ACM Press.

Berkhin, P. (2006). A survey of clustering data mining techniques. San Jose, USA, *Grouping Multidimensional Data*, *12*, 25–71.

BRASIL, E. D. (2005a). Abc da impressão digital de dados variáveis - parte 1. São Paulo, BR, *Professional Publish - Tecnologia aplicada as Artes Gráficas, Design e Criação*, *77*, 26–35.

BRASIL, E. D. (2005b). Abc da impressão digital de dados variáveis - parte 2. São Paulo, BR, *Professional Publish - Tecnologia aplicada as Artes Gráficas, Design e Criação*, *78*, 24–30.

BRASIL, E. D. (2005c). Abc da impressão digital de dados variáveis - parte 3. São Paulo, BR, *Professional Publish - Tecnologia aplicada as Artes Gráficas, Design e Criação*, *79*, 25–29.

Cloudmark, i. (2007). Cloudmark: Anti-spam, spam filter, anti-virus, anti-phishing, and spam-blocking for service providers. Disponível em: http://www.cloudmark.com/gateway/network/ Acessado em setembro de 2007.

da Silva, A. C. B., de Oliveira, J. B. S., Mano, F. T. M., Silva, T. B., Meirelles, L. L., Meneguzzi, F. R., & Giannetti, F. (2005). Support for arbitrary regions in xsl-fo. *DocEng '05: Proceedings of the 2005 ACM symposium on Document engineering* (pp. 64–73). New York, USA: ACM.

DeBronkart, D., & Davis, P. (2000). Personalized print markup language. *XML '00: Proceedings of the XML Europe* (pp. 1–14). Paris, FR: International Digital Enterprise Alliance.

DocEng (2001). Acm symposium on document engineering. Disponível em: http://www.documentengineering.org/doceng01/index.html Acessado em outubro de 2007.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD '96: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). Washington, USA: AAAI Press.

Everitt, B. S. (1993). Cluster analysis. Hoboken, USA: Edward Arnold and Halsted Press.

Faria, A. C., & Oliveira, J. B. S. (2006). Measuring aesthetic distance between document templates and instances. *DocEng '06: Proceedings of the 2006 ACM symposium on Document engineering* (pp. 13–21). New York, USA: ACM Press.

Giannetti, F., Fernandes, L. G., Timmers, R., Nunes, T., Raeder, M., & Castro, M. (2006). High performance xsl-fo rendering for variable data printing. *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing* (pp. 811–817). New York, USA: ACM.

Group, T. P. (2006). Php. Disponível em: http://www.php.net Acessado em outubro de 2006.

Han, J., & Kamber, M. (2000). Data mining: Concepts and techniques. San Francisco, USA: Morgan Kaufmann.

Harrington, S. J., Naveda, J. F., Jones, R. P., Roetling, P., & Thakkar, N. (2004). Aesthetic measures for automated document layout. *DocEng '04: Proceedings of the 2004 ACM symposium on Document engineering* (pp. 109–111). New York, USA: ACM Press.

Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. London, EN, *Applied Statistics*, *28*, 100–108.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. New Jersey, USA: Prentice-Hall.

Jain, A. K., Murty, N., & Flynn, P. (1999). Data clustering: a review. New York, USA, *ACM Computer Survey*, *31*, 264–323.

Kaufmann, L., & Rousseeuw, P. (1987). Clustering by means of medoids. Amsterdam, NL, *Statistical Data Analysis based on the L 1 Norm and Related Methods*, *1*, 405–416.

Kurniawan, A., Benech, N., Yufei, T., Feng, T., Jiying, W., & Malamatos, T. (1999). Towards high-dimensional clustering. *COMP 530 '99: Proceedings of the Database Architecture and Implementation* (pp. 1–43). Hong Kong, CH: COMP.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *BSMSP '67: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). Berkeley, USA: University of California Press.

Mahalanobis, P. (1936). On the generalised distance in statistics. *NISI '36: Proceeding of the National Institute of Sciences of India* (pp. 49–55). Nova Deli, IN.

Masum, H., & Zhang, Y.-C. (2004). Manifesto for the reputation society. Illinois, USA, *First Monday*, *9*, 1–15.

Meneguzzi, F. R., Meirelles, L. L., Mano, F. T. M., de Souza Oliveira, J. B., & da Silva, A. C. B. (2004). Strategies for document optimization in digital publishing. *DocEng '04: Proceedings of the 2004 ACM symposium on Document engineering* (pp. 163–170). New York, USA: ACM.

MySQL, A. (2006). Mysql. Disponível em: http://www.mysql.com Acessado em outubro de 2006.

Natsoulas, A. (1989). Taxicab conics: an exploration into the world of taxicab geometry. New York, USA, *Journal of Computers in Mathematics and Science Teaching*, *8*, 39–47.

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. *VLDB '94: Proceeding of the 20th International Conference on Very Large Data Bases* (pp. 144–155). Los Altos, USA: Morgan Kaufmann Publishers.

Perone, M. (2004). *An overview of spam blocking techniques* (Technical Report). Campbell, USA, Barracuda Networks.

PODI (2007). Print markup language functional specification version 2.1. Disponível em: http://www.podi.org/ Acessado em outubro de 2007.

Prakash, V. V., & O'Donnell, A. (2005). Fighting spam with reputation systems. New York, USA, *Queue*, *3*, 36–41.

Purvis, L., Harrington, S., O'Sullivan, B., & Freuder, E. C. (2003). Creating personalized documents: an optimization approach. *DocEng '03: Proceedings of the 2003 ACM symposium on Document engineering* (pp. 68–77). New York, USA: ACM.

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000). Reputation systems: Facilitating trust in internet interactions. Chapel Hill, USA, *Communications of the ACM*, *43*, 45–48.

W3C, W. W. W. C. (2007). Extensible stylesheet language (xsl) version 1.1. Disponível em: http://www.w3.org/TR/2003/WD-xsl11-20031217/ Acessado em outubro de 2007.

Weisstein, E. W. (2007). Golden ratio. Disponível em: http://mathworld.wolfram.com/GoldenRatio.html Acessado em outubro de 2007.

Xu, R., & Wunsch, D., I. (2005). Survey of clustering algorithms. Nicosia, CY, *IEEE Transactions on Neural Networks*, *16*, 645–678.

# A Apêndice 1

## A.1 Experimentos A

### A.1.1 3D

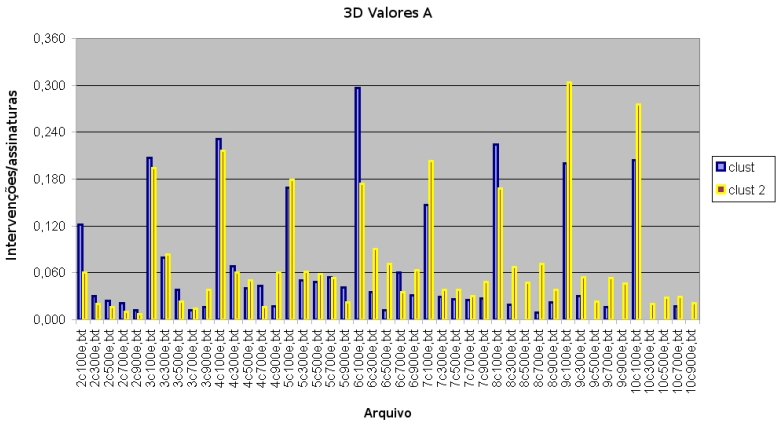


Figura 36 – 3D Experimentos A média de perguntas

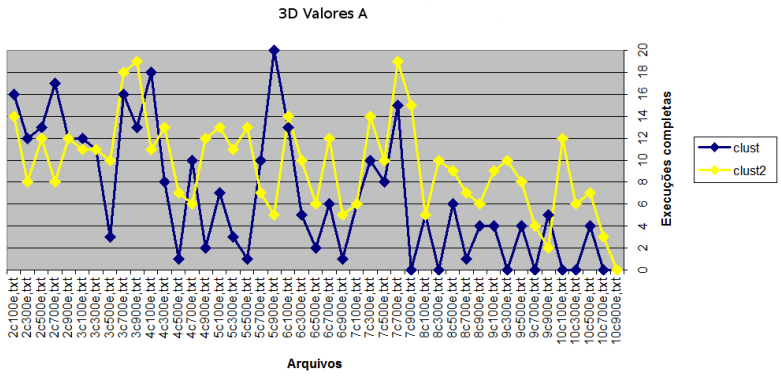


Figura 37 – 3D Experimentos A execuções completas

### A.1.2 4D

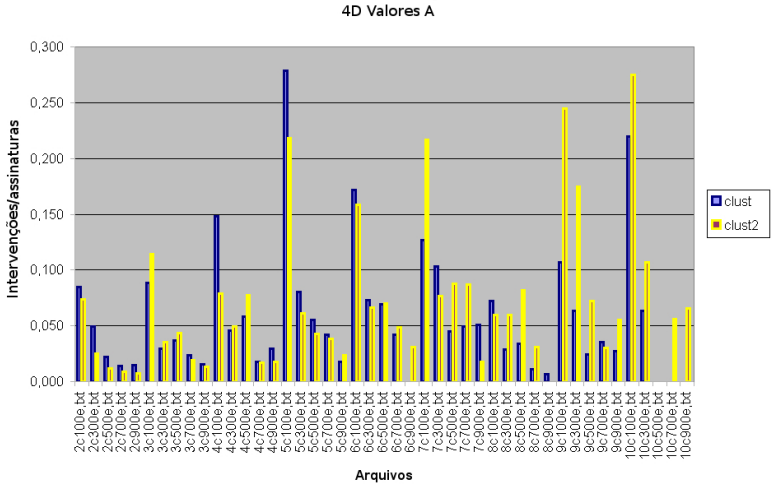


Figura 38 – 4D Experimentos A média de perguntas

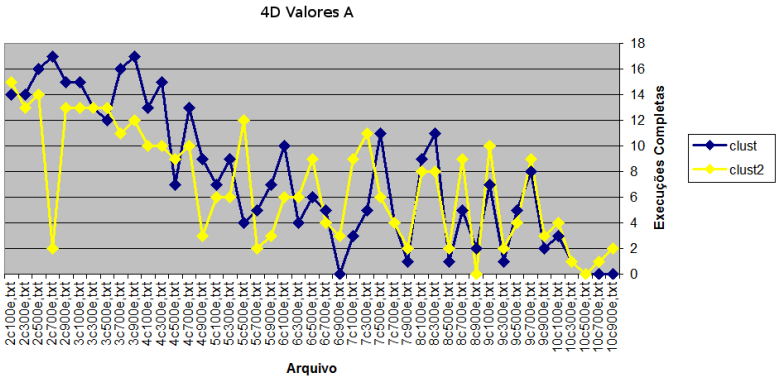


Figura 39 – 4D Experimentos A execuções completas

### A.1.3 5D

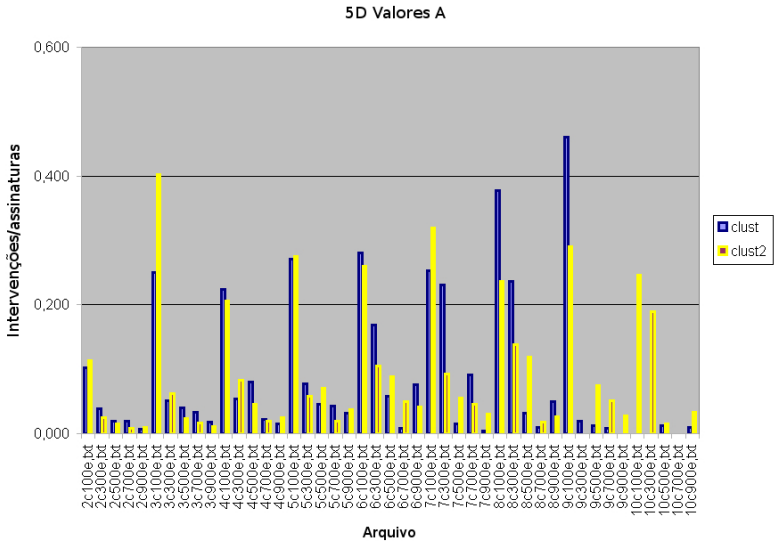


Figura 40 – 5D Experimentos A média de perguntas

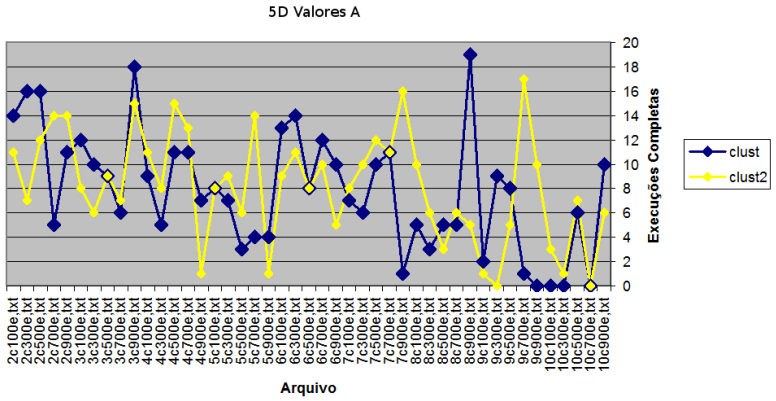


Figura 41 – 5D Experimentos A execuções completas

### A.1.4 6D

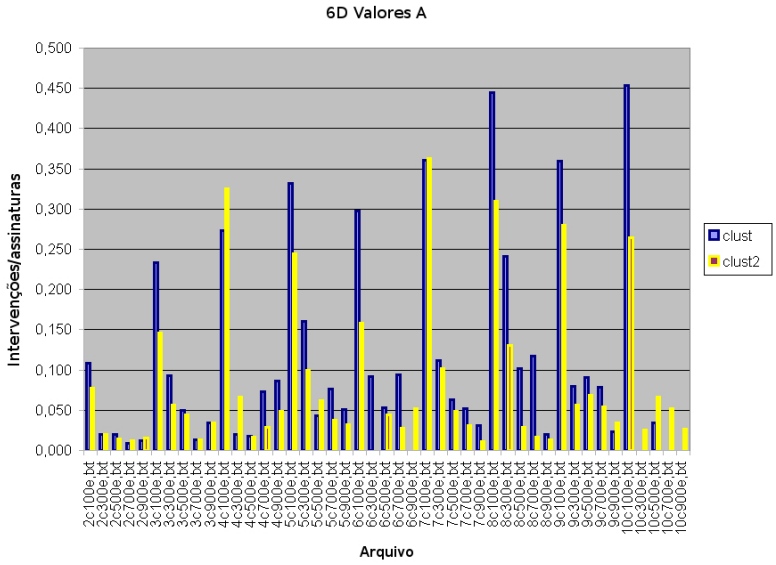


Figura 42 – 6D Experimentos A média de perguntas

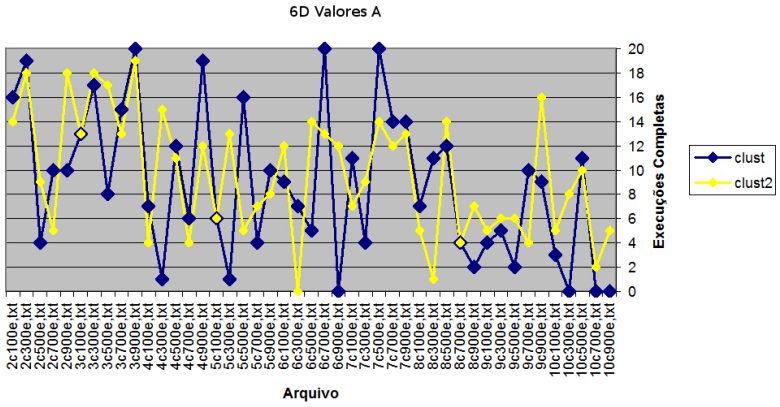


Figura 43 – 6D Experimentos A execuções completas

## A.2 Experimentos B

### A.2.1 3D

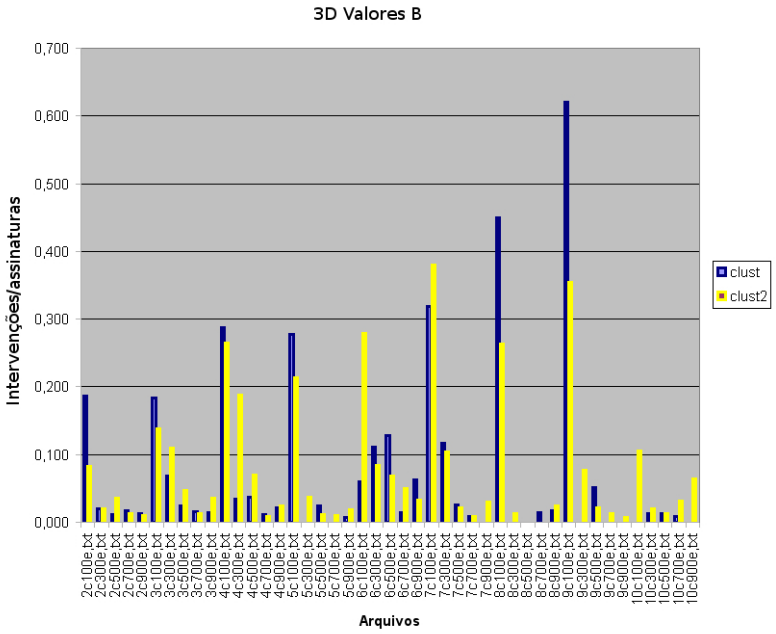


Figura 44 – 3D Experimentos B média de perguntas

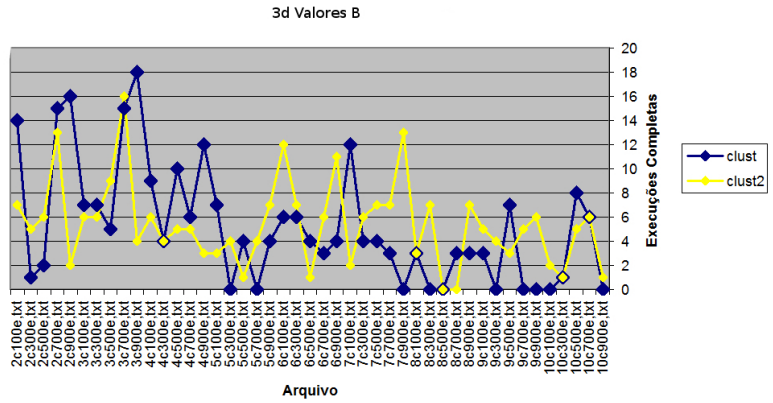


Figura 45 – 3D Experimentos B execuções completas

### A.2.2 4D

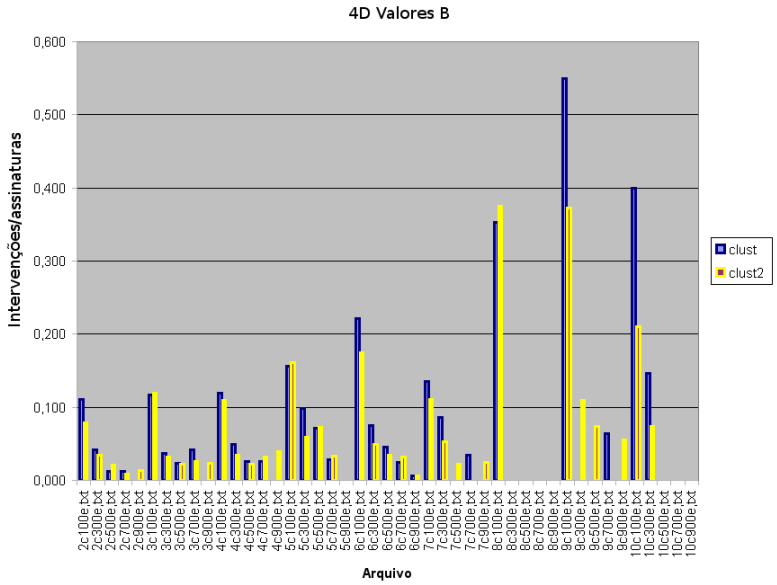


Figura 46 – 4D Experimentos B média de perguntas

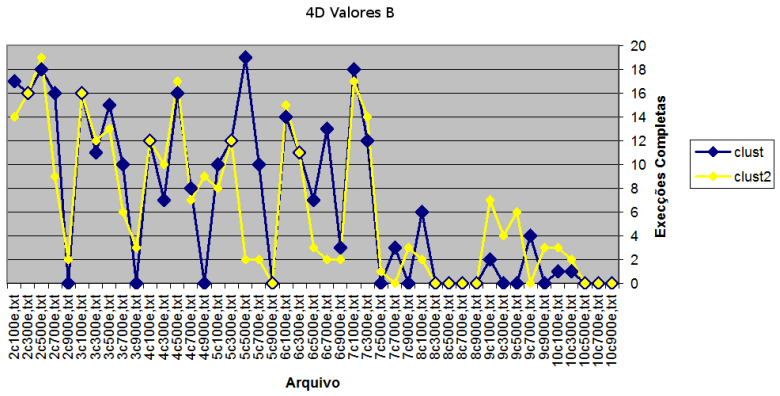


Figura 47 – 4D Experimentos B execuções completas

### A.2.3 5D

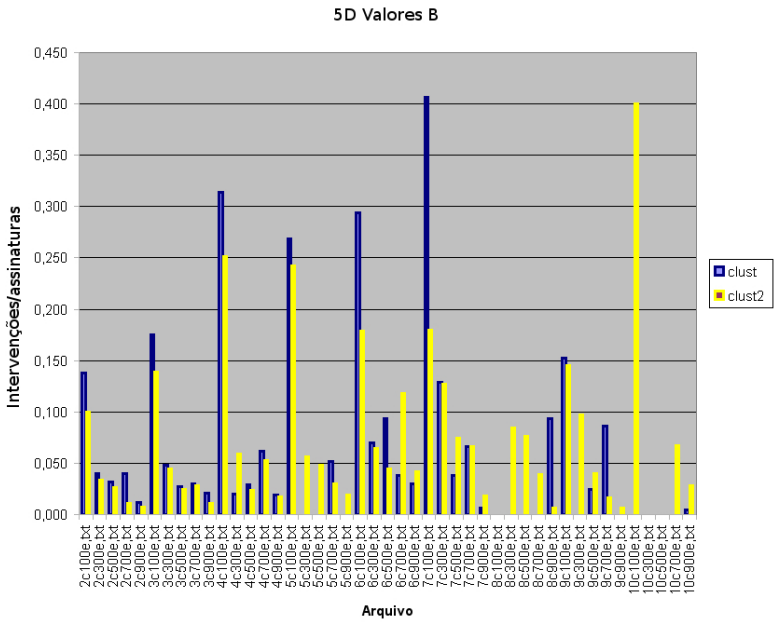


Figura 48 – 5D Experimentos B média de perguntas

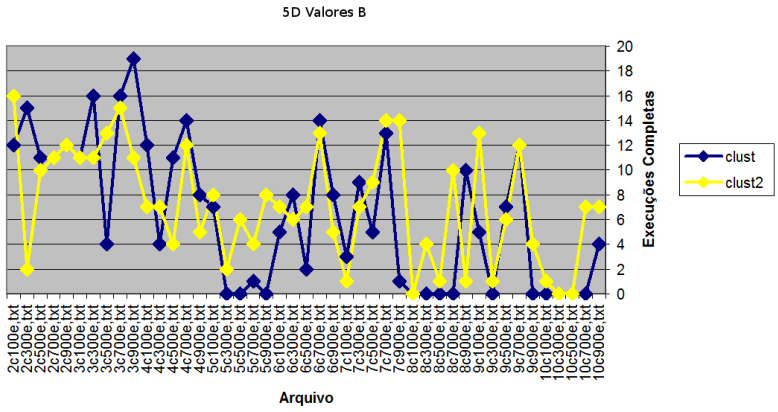


Figura 49 – 5D Experimentos B execuções completas

### A.2.4 6D

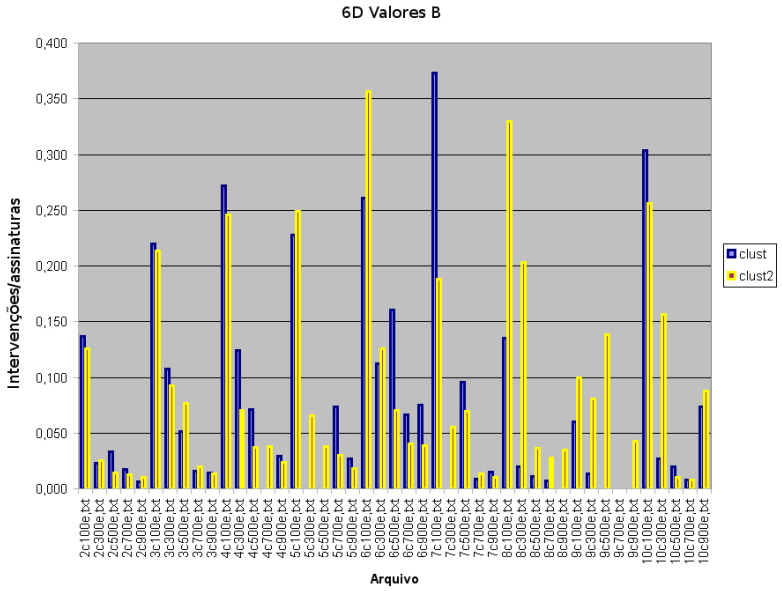


Figura 50 – 6D Experimentos B média de perguntas

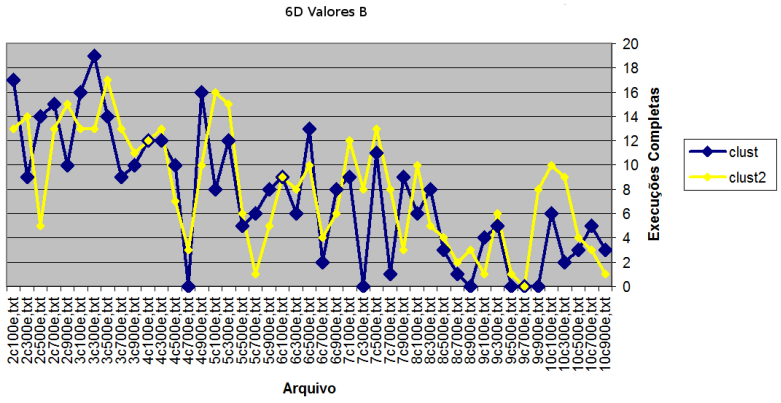


Figura 51 – 6D Experimentos B execuções completas

## A.3 Experimentos C

### A.3.1 3D

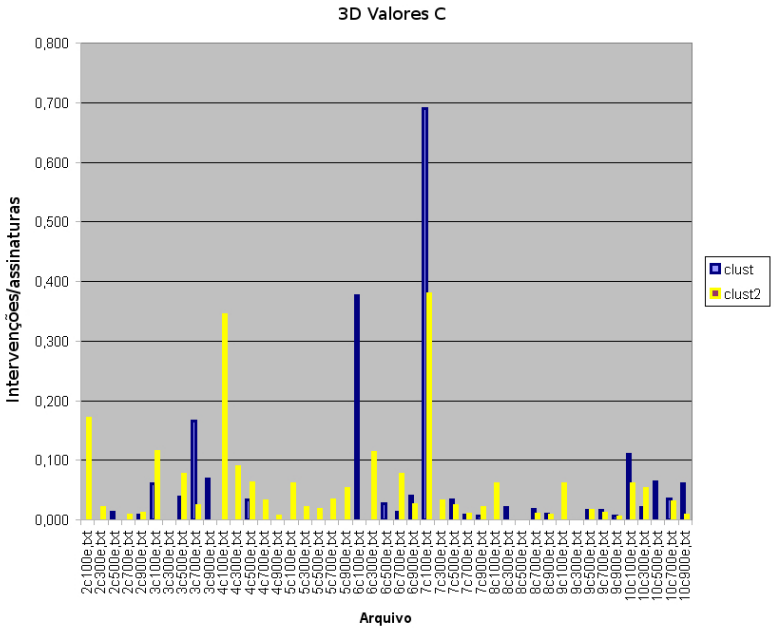


Figura 52 – 3D Experimentos C média de perguntas

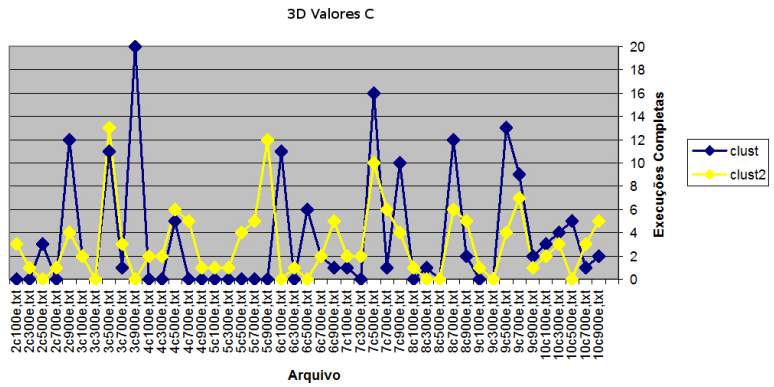


Figura 53 – 3D Experimentos C execuções completas

### A.3.2 4D

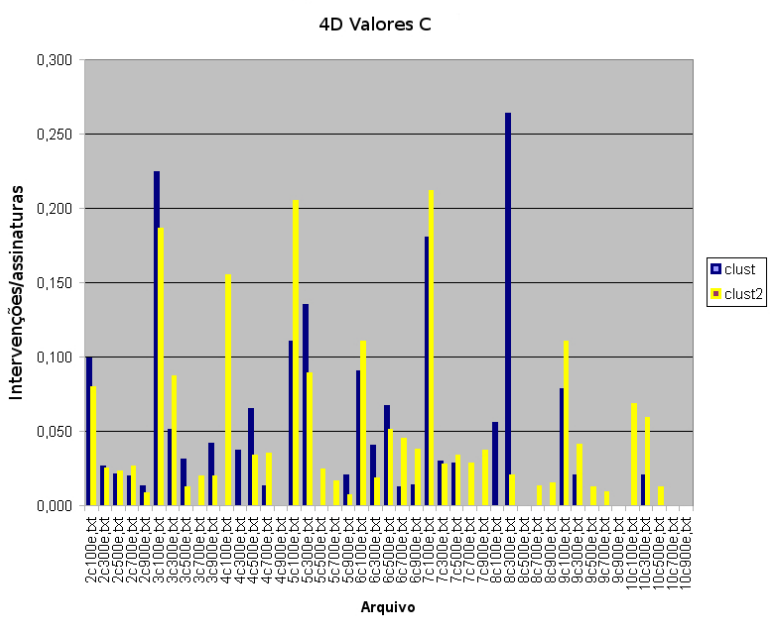


Figura 54 – 4D Experimentos C média de perguntas

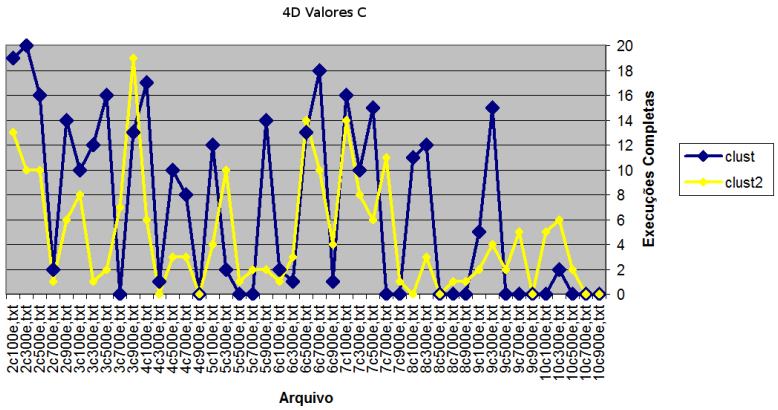


Figura 55 – 4D Experimentos C execuções completas

### A.3.3 5D

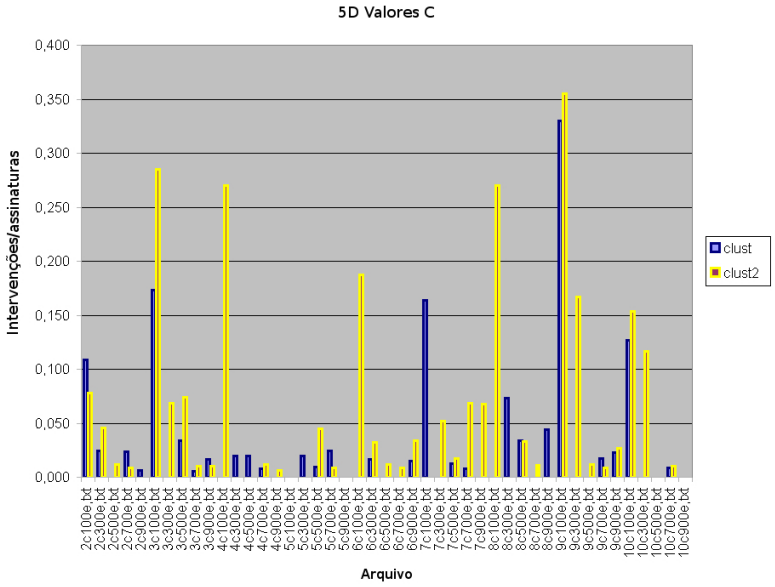


Figura 56 – 5D Experimentos C média de perguntas

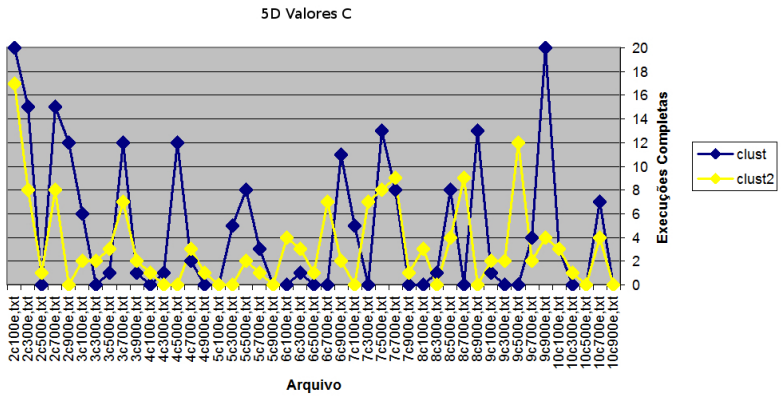


Figura 57 – 5D Experimentos C execuções completas

### A.3.4 6D

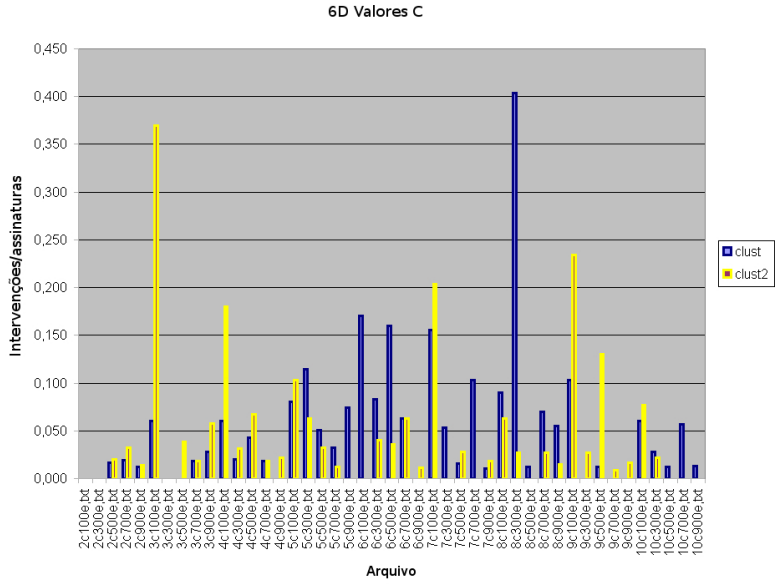


Figura 58 – 6D Experimentos C média de perguntas

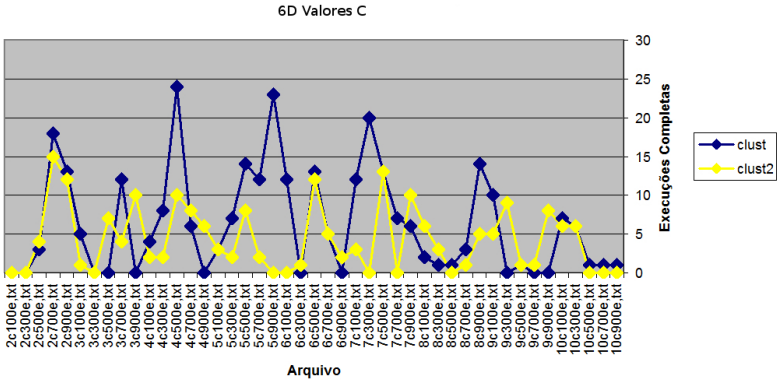


Figura 59 – 6D Experimentos C execuções completas

## A.4 Experimentos D

### A.4.1 3D

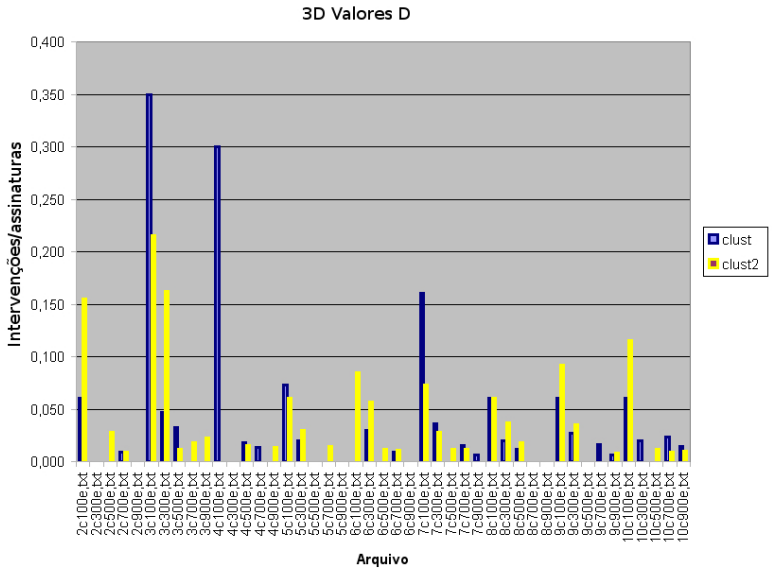


Figura 60 – 3D Experimentos D média de perguntas

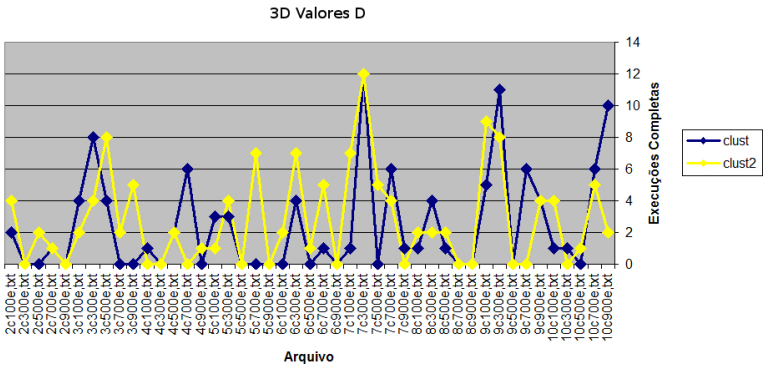


Figura 61 – 3D Experimentos D execuções completas

### A.4.2 4D

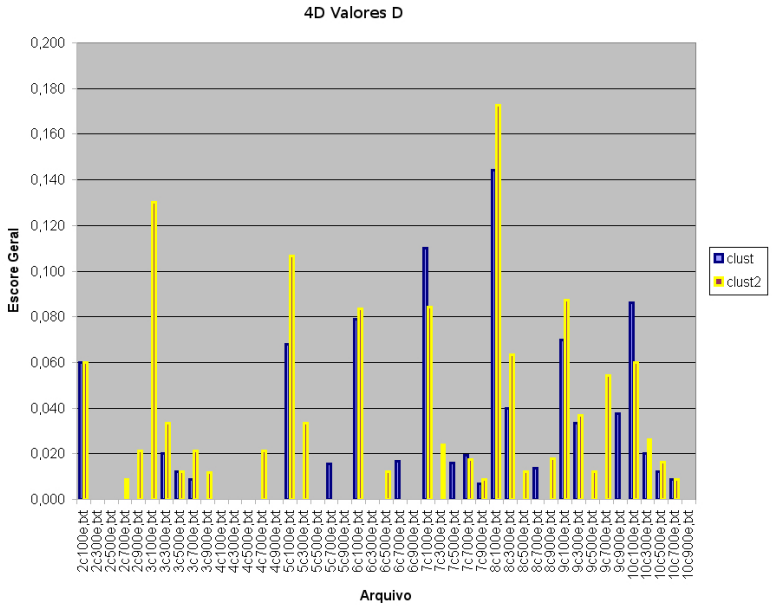


Figura 62 – 4D Experimentos D média de perguntas

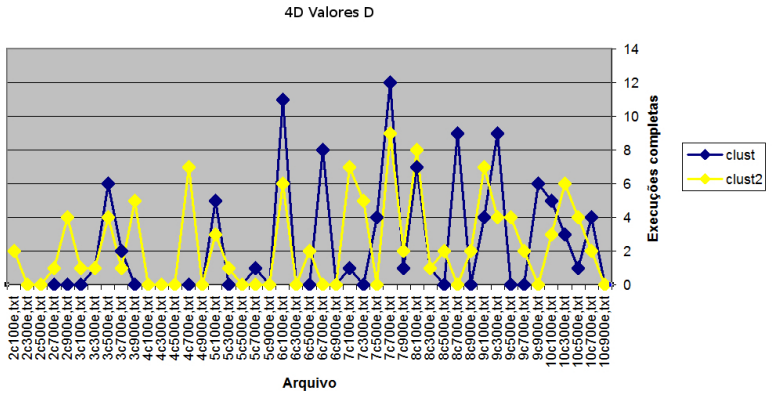


Figura 63 – 4D Experimentos D execuções completas

### A.4.3 5D

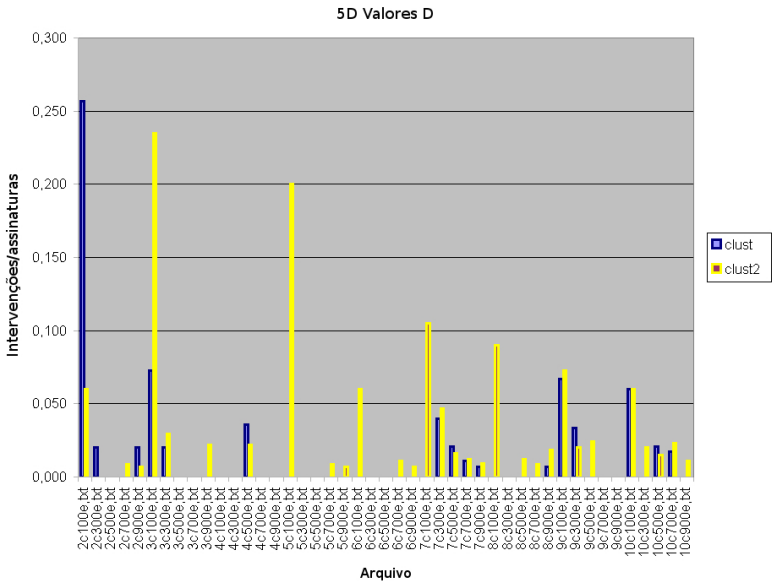


Figura 64 – 5D Experimentos D média de perguntas

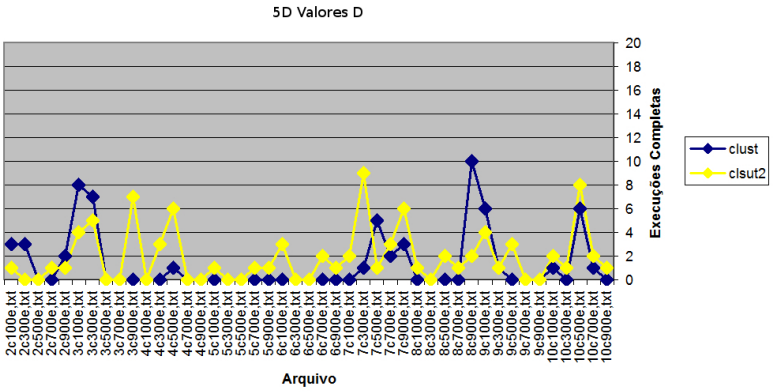


Figura 65 – 5D Experimentos D execuções completas

### A.4.4 6D

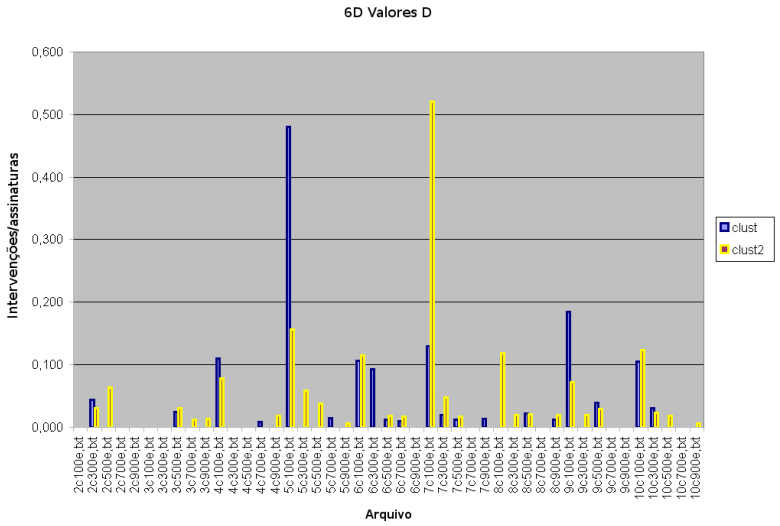


Figura 66 – 6D Experimentos D média de perguntas

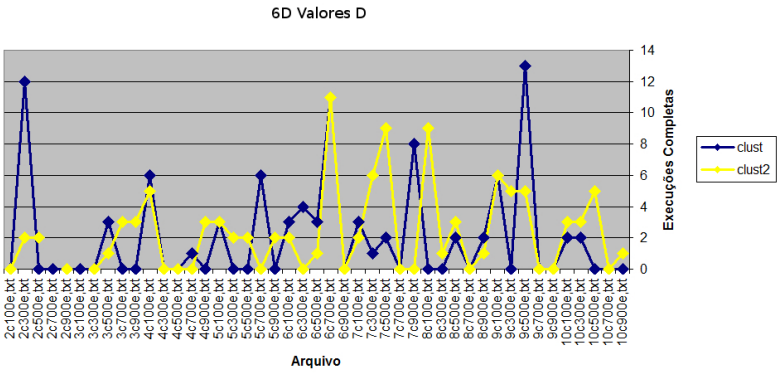


Figura 67 – 6D Experimentos D execuções completas