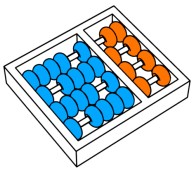
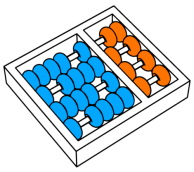


Filipe de Oliveira Costa

"Reconstrução de Florestas para o Problema de Filogenia Multimídia"

CAMPINAS
2013



Universidade Estadual de Campinas
Instituto de Computação

Filipe de Oliveira Costa

"Reconstrução de Florestas para o Problema de Filogenia Multimídia"

Orientador(es)
Prof. Dr. Anderson Rocha (Orientador)
Prof. Dr. Zanoni Dias (Co-orientador)

Texto apresentado ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Computação da Universidade Estadual de Campinas para o Exame de Qualificação Específico de Doutorado em Ciência da Computação.

CAMPINAS
2013

Abstract

In the last years, some researchers have successfully developed approaches to near-duplicate detection aiming at identifying the cohabiting versions of a given document online. However, only recently have there been the first attempts to go beyond the detection task and to find the structure of evolution within a set of related document (e.g., image, video, etc.) overtime. In this work, we propose approaches for identifying the structure of relationships of digital documents in a given set of semantically similar documents, and reconstructing a structure that represents their past history and ancestry information. For this, we aim at designing approaches that calculate the dissimilarities among the documents and group them in distinct trees of history processing automatically. This project is linked to a larger project funded by the European Union denominated "REWIND – Reverse Engineering of Audio Visual Content Data", in partnership with the european universities *Politecnico di Milano* (Italy), *Università di Siena* (Italy), *Università degli Studi di Firenze* (Italy), *Imperial College* (UK), *Universidad de Vigo* (Spain) and *Fraunhofer IDMT* (Germany).

Resumo

Nos últimos anos, muitos pesquisadores têm desenvolvido, com sucesso, abordagens para detecção de duplicatas de documentos com o intuito de identificar cópias semelhantes de um dado documento multimídia (e.g., imagem, vídeo, etc.) publicado na Internet. Entretanto, somente recentemente têm se desenvolvido as primeiras pesquisas para ir além da detecção de duplicatas e encontrar a estrutura de evolução de um conjunto de documentos relacionados e modificados ao longo do tempo. Neste trabalho, nós propomos abordagens para, dado um conjunto de duplicatas próximas de documentos, identificar automaticamente a relação entre elas e reconstruir a estrutura que represente o histórico de geração das mesmas. Para isso, objetivamos o desenvolvimento de abordagens que calculem a dissimilaridade entre duplicatas e as separem corretamente em árvores que representem a estrutura de relação entre elas de forma automática. Este projeto de doutorado estará vinculado a um projeto maior financiado pela União Europeia denominado "*REWIND – Reverse Engineering of Audio Visual Content Data*" (Engenharia Reversa de Dados de Conteúdo Áudio-Visual) em parceria com as universidades européias *Politecnico di Milano* (Itália), *Università di Siena* (Itália), *Università degli Studi di Firenze* (Itália), *Imperial College* (Reuno Unido), *Universidad de Vigo* (Espanha) e *Fraunhofer IDMT* (Alemanha).

Sumário

Capítulo 1

Introdução

Nas últimas décadas, o número de usuários da Internet e de redes sociais sofreu um grande aumento. Com isso, o compartilhamento de conteúdo multimídia se tornou algo simples. Alguns exemplos incluem o Flickr[1] e o Youtube[2] que têm como objetivo o compartilhamento de fotos e vídeos, respectivamente. O Flickr recebe, aproximadamente, 6.000 fotos de alta resolução por minuto, e o Youtube recebe, em média, 100 horas de vídeo por minuto[3]. Infelizmente, a facilidade com que tal conteúdo pode ser compartilhado e distribuído tem algumas consequências nem sempre positivas. Por exemplo, após uma imagem ser postada na Internet, um usuário pode obter uma cópia, modificá-la digitalmente e compartilhar a nova versão da imagem. Em alguns casos, a criação de tais adulterações pode infringir direitos autorais ou até mesmo espalhar pela rede conteúdo digital ilegal ou que afeta a imagem de empresas ou pessoas.

Nesse sentido, algumas pesquisas têm como objetivo detectar duplicatas de documentos digitais. Embora a detecção de duplicatas exatas é direta, esta abordagem não pode ser considerada no caso de imagens alteradas digitalmente [16, 33]. Na última década, muitas abordagens têm sido desenvolvidas com o objetivo de identificar versões modificadas de um determinado documento, dado um conjunto de duplicatas. Na literatura, tais técnicas são denominadas técnicas para detecção e recuperação de duplicatas próximas (*Near-Duplicate Detection and Retrieval* – NDDR) de um documento [16]. Na detecção, o objetivo é verificar, dados dois documentos digitais, se um deles é uma cópia "próxima" do outro. Na recuperação, o objetivo é encontrar todas as cópias "próximas" de um documento em um conjunto.

Uma abordagem mais desafiadora para expandir o problema de NDDR é, dado um conjunto de documentos multimídia os quais um analista sabe que são duplicatas próximas, identificar qual documento é o original deste conjunto e qual a estrutura de geração de tais documentos. Neste cenário, nós estamos interessados em identificar o histórico de geração dos documentos, levando em consideração a relação existente entre eles. Essa identificação se faz útil para rastreamento de conteúdo ilegal sem a necessidade de marcações digitais ou encontrar o documento original dentro de um conjunto de duplicatas, por exemplo.

[1] `http://www.flickr.com`

[2] `http://www.youtube.com`

[3] Estatísticas coletadas diretamente dos *websites em 21 de Outubro de 2013.*

A identificação da estrutura original de modificações de um dado conjunto de documentos tem uma analogia natural com o processo evolutivo biológico. Na biologia, podemos olhar um processo de evolução como sendo um processo de arborescência, onde espécies evoluem no decorrer do tempo e estas podem ser separadas em diferentes ramificações. Podemos visualizar o processo de arborescência como sendo uma árvore filogenética [20].

Podemos dizer que existem, principalmente, dois fatores importantes e independentes no processo de reconstrução de uma árvore filogenética de um conjunto de documentos semanticamente semelhantes: a função de dissimilaridade relacionada com as diferenças entre cada par de duplicatas no conjunto e o algoritmo de reconstrução da árvore em si.

Algumas pesquisas recentes têm conseguido resultados interessantes no que diz respeito à definição de uma árvore que represente a estrutura de geração de um conjunto de documentos digitais, principalmente imagens e vídeos [7, 9, 10, 29]. Entretanto, é importante notar que existem situações onde se deseja encontrar não uma, mas um conjunto de árvores filogenéticas (floresta) que represente a relação entre os documentos. Este cenário é desejável quando não se tem certeza que todas as duplicatas foram geradas a partir de um único ancestral comum. Uma solução existente propõe que o usuário forneça o número de árvores a serem geradas [11]. Entretanto, a informação do número de árvores a serem geradas pode não estar disponíveis em um ambiente real (e.g., conjunto de documentos multimídia obtido da Internet).

Este projeto de pesquisa visa buscar possíveis soluções para o problema de filogenia de documentos multimídia (imagens e vídeos) onde, dado um conjunto de duplicatas próximas, determinar a árvore (no caso de todos os documentos possuírem um ancestral em comum) ou a floresta (no caso de vários subconjuntos de duplicatas próximas no mesmo conjunto) que represente a relação de geração entre as duplicatas. O projeto estará vinculado a um projeto maior financiado pela União Européia denominado "*REWIND – Reverse Engineering of Audio Visual Content Data*" (Engenharia Reversa de Dados de Conteúdo Áudio-Visual) em parceria com as universidades *Politecnico di Milano* (Itália), *Università di Siena* (Itália), *Università degli Studi di Firenze* (Itália), *Imperial College* (Reuno Unido), *Universidad de Vigo* (Espanha) e *Fraunhofer IDMT* (Alemanha). O projeto tem coordenação no lado brasileiro do Prof. Anderson Rocha e colaboração dos professores Zanoni Dias e Siome Goldenstein, do Instituto de Computação da Universidade Estadual de Campinas (UNICAMP). O lado europeu é coordenado pelo Prof. Stefano Tubaro (*Politecnico di Milano*).

1.1 Objetivo

O objetivo principal deste trabalho é desenvolver abordagens que permitam a identificação da estrutura de geração de um conjunto de duplicatas de documentos multimídia. Os documentos que utilizaremos com objetos de estudo neste trabalho serão imagens e vídeos. Consideramos como meta principal do projeto resolver o problema de encontrar florestas (isto é, quando um conjunto de duplicatas pode ter vários subconjuntos distintos).

1.2 Contribuições científicas e impacto

A realização deste trabalho se faz útil para o auxílio na resolução de problemas que envolvem questões legais. Por meio da filogenia de documentos multimídia é possível verificar, por exemplo, a fonte de imagens ou vídeos ilegais que foram distribuídos indevidamente pela rede (e.g., vídeo de pornografia infantil, imagem com direitos autorais, etc.), dadas algumas duplicatas de tais documento. Nós estamos interessados em encontrar tanto uma árvore quanto uma floresta filogenética (isto é, uma ou mais árvores para cada conjunto de duplicatas).

Por se tratar de um problema relativamente novo, acreditamos que os tipos de abordagens propostos gerarão contribuições científicas importantes para o estudo da filogenia de documentos, tais como:

- Estudo detalhado sobre o problema de filogenia de documentos digitais considerando árvores e florestas, uma vez que abordagens para solucionar este problema são recentes e poucos estudos foram feitos a respeito do tema;
- Definição de algoritmos que, dado um conjunto de duplicatas próximas, encontrem a floresta que represente a relação entre as duplicatas. Os algoritmos gerados auxiliarão diretamente na solução de problemas como distribuição de conteúdo ilegal, infração de direitos autorais e busca de imagens por conteúdo;
- Validação da abordagem gerada neste projeto para solucionar o problema de filogenia de documentos multimídia em um ambiente onde não se sabe a proveniência de tais documentos, uma vez que este tipo de validação foi pouco investigado na literatura;
- Análise de propriedades dos documentos que possam auxiliar no desenvolvimento de novos cálculos de dissimilaridade, uma vez que esse cálculo influencia diretamente nos resultados dos algoritmos para filogenia.

1.3 Organização

Este projeto está organizado da seguinte maneira: o Capítulo 2 apresenta os trabalhos relacionados com filogenia de documentos. A metodologia utilizada para o desenvolvimento do trabalho, bem como os desafios do projeto são apresentados no Capítulo 3. Por fim, o Capítulo 4 apresenta o cronograma do projeto e os indicadores de sucesso do mesmo.

Capítulo 2

Estado da arte

Neste capítulo, apresentamos os trabalhos relacionados ao nosso projeto. Destacamos, brevemente, o problema de detecção de duplicatas e detalhamos um pouco mais o problema de filogenia multimídia.

2.1 Detecção de duplicatas

Uma duplicata próxima é uma versão modificada de um determinado documento que mantém as características semânticas do mesmo. Podemos dar um exemplo considerando imagens, onde uma duplicata é uma versão de uma determinada imagem após sofrer uma ou mais transformações (e.g., recorte, rotação, correção de cor, etc.) [9].

Joly et al. [16] definem, formalmente, uma duplicata próxima baseando-se no conceito de transformações toleradas. De acordo com os autores, um documento $\mathcal{D}_1$ é uma duplicata próxima de um documento $\mathcal{D}$ se, dada uma família de transformações $\mathcal{T}$, existe pelo menos uma transformação $T_\beta \in \mathcal{T}$ tal que $\mathcal{D}_1 = T_\beta(D)$. Uma família $\mathcal{T}$ pode conter diversas combinações de transformações gerando, por exemplo, a duplicata $\mathcal{D}_3 = T_3 \circ T_2 \circ T_1(\mathcal{D}), T_{\beta=1,2,3} \in \mathcal{T}$.

Uma duplicata próxima é uma relação de equivalência de pares. Essa relação faz a ligação entre o objeto original (a raiz da árvore) e suas variações geradas por meio de transformações (e.g., compressão, correção de brilho e contraste, recorte, etc.) [24]. Se o documento original $\mathcal{D}$ tem uma duplicata direta $\mathcal{D}_1$ e $\mathcal{D}_1$ tem uma duplicata direta $\mathcal{D}_2$, então $\mathcal{D}_2$ é, por sua vez, uma duplicata de $\mathcal{D}$.

Podemos dizer que existem dois tipos diferentes de filosofias para detecção de duplicatas próximas: a detecção baseada em marcações digitais e a detecção baseada em conteúdo. Detecção por marcações dependem de uma "assinatura" (e.g., *watermarking*, *fingerprint*, etc.) incorporada no documento original antes de sua disseminação [24]. Com esses métodos, podemos detectar o objeto original analisando a presença e as modificações dos padrões de tal assinatura em outros documentos. Em contrapartida, métodos de detecção baseados em conteúdo dependem da análise do conteúdo do documento de forma a extrair características visuais relevantes. Esses métodos identificam quando um conjunto de características estão

próximas das características do objeto original.

Vários métodos de NDDR tem sido desenvolvidos nos últimos anos para uma variedade de aplicações. Técnicas de NDDR tem sido utilizadas por fotógrafos para organização de coleções de fotos [15,31], correspondência entre arquivos multimídia [35], detecção de infração de direitos autorais em imagens e vídeos [4,21] e detecção de falsificações de imagens [14,17]. Tais trabalhos focam na identificação de duplicatas próximas, sem se importar com a estrutura de modificações ou com quais transformações as duplicatas foram geradas.

2.2 Filogenia de documentos

Uma abordagem mais desafiadora para expandir o problema de NDDR é, dado um conjunto de documentos multimídia (e.g., imagens, vídeo, etc.) os quais um analista sabe que são duplicatas próximas, identificar o histórico de geração dos documentos, levando em consideração a relação existente entre eles. Fazer essa identificação tem aplicações diretas nas seguintes áreas [11]:

1. **Proteção de direitos autorais**: rastreamento de documentos digitais sem a necessidade de marcações digitais, como marcas d'água;
2. **Segurança**: a estrutura de modificações de um conjunto de duplicatas de imagens pode fornecer informações importantes de comportamento suspeito e apontar a direção da distribuição do conteúdo;
3. **Novos serviços de rastreamento**: a relação entre as duplicatas próximas pode fornecer novos serviços para o rastreamento da imagem original com elementos chave a fim de determinar o processo de formação de opinião no decorrer do tempo/espaço [18, 29];
4. **Análise forense**: É possível obter resultados melhores se a análise ocorrer no documento original ao invés de em uma duplicata próxima [28];

Durante sua vida útil, um objeto multimídia pode sofrer diferentes operações que podem alterar as características de seu conteúdo de uma maneira bem definida e detectável. Portanto, estabelecer o relacionamento existente entre pares de objetos digitais analisando o conteúdo desses objetos é uma tarefa desafiadora uma vez que existem vários operadores que podem ser aplicados nos objetos para alterar seu conteúdo. Todavia tal tarefa ainda é factível devido aos traços, artefatos ou pistas deixados pelas operações de modificações aplicadas nos documentos.

Encontrar a relação de geração dos objetos dentro de um conjunto de duplicatas é o objetivo da área de pesquisa denominada Filogenia de Documentos. Por se tratar de uma área de pesquisa nova, existem poucos trabalhos sobre o tema. A seguir, apresentamos alguns trabalhos correlatos.

2.2.1 Trabalhos correlatos

Kennedy et al. [18] definiram o problema da relação "pai-filho" entre pares de imagens (Arqueologia de Imagens) e propuseram uma abordagem para detectar, dado um par de imagens,

qual das duas imagens sofreu alguma transformação para que a outra imagem fosse gerada. Para isso, os autores propuseram um mapa de migração visual (*Visual Migration Map* – VMM), representando uma aproximação do histórico da imagem. Contudo, os autores não apresentaram discussões de como encontrar possíveis parâmetros para a família de transformações que uma imagem poderia sofrer para gerar as duplicatas resultantes nem discutiram possíveis algoritmos para a construção da árvore de descendência ligada ao processo evolutivo das imagens.

A verificação temporal das transformações em imagens também foi explorada por Fan and Queiroz [13], onde os autores identificam o histórico de compressão associados a uma imagem, e por Mao et al. [23], no qual os autores discutem sobre informações relevantes fornecidas pelo dispositivo gerador da imagem. Entretanto, nesses trabalhos também não há discussões sobre abordagens para encontrar a estrutura de relacionamento temporal que represente o conjunto de duplicatas próximas.

De Rosa et al. [29] propuseram uma abordagem para detectar a dependência de imagens em um conjunto de imagens $\mathcal{I}$, considerando a hipótese de que qualquer imagem $\mathcal{I}_i \in \mathcal{I}$ pode ser descrita como a composição de dois componentes separadamente: componentes baseados no conteúdo da imagem e componentes independentes do conteúdo da imagem. Para verificar a dependência entre duas imagens $\mathcal{I}_A$ e $\mathcal{I}_B$, os autores consideram que a informação mútua das imagens pode ser expressa como sendo a soma das informações mútuas entre esses componentes. Os autores assumem que, se existe uma dependência entre duas imagens $\mathcal{I}_A$ e $\mathcal{I}_B$, uma dessas imagens pode ser obtida aproximadamente aplicando sobre a outra funções de processamento de imagens. Assim, o conteúdo das imagens é avaliado para estimar tais funções.

Após a transformação de uma imagem na outra utilizando as funções de processamento de imagens estimadas, os autores calculam o coeficiente de correlação entre as duas imagens baseando-se nos componentes independentes do conteúdo. Nesse caso, os autores utilizaram o ruído de foto-responsividade não uniforme (*Photo Responsivity Non-uniform* – PRNU) das imagens, o qual é causado pela interação entre a luz do ambiente e o sensor de captura da câmera no momento da geração de uma imagem. A estimativa do PRNU é amplamente usada no cenário de atribuição de fonte para câmeras, onde o objetivo é identificar se uma determinada imagem foi gerada por uma câmera sob investigação [6, 22].

Depois de analisar todos os possíveis pares de imagens em $\mathcal{I}$, os autores geram um grafo de dependências, no qual os vértices representam as imagens, as arestas representam a relação entre duas imagens e o peso das arestas representa o valor da correlação entre elas. Para isso, os autores avaliam cada aresta do grafo e descartam as que estão abaixo de um limiar. Primeiramente, os autores visam remover *loops* diretos (é impossível uma imagem i gerar uma imagem j e vice-versa ao mesmo tempo). Em seguida, outras arestas são descartadas de forma que um vértice tenha grau de entrada $= 1$ (assumindo-se que uma imagem não pode ter sido gerada por meio de composição entre duas ou mais imagens). O processo se repete até que se encontre uma árvore que represente a relação entre as imagens.

Dias et al. [9] introduziram e definiram formalmente o problema de Filogenia de Imagens encontrando a estrutura de transformações e seus parâmetros que geraram um conjunto de imagens de duplicatas próximas, definida pelos autores como Árvore Filogenética de Imagens

(AFI). Em geral, podemos considerar que um algoritmo de reconstrução de uma AFI procura construir esta árvore com base em um conjunto de duplicatas e uma função de dissimilaridade d que produz valores baixos quando um par de imagem possui a relação "pai-filho" na árvore, e valores altos para pares de imagens muito diferentes entre si (isto é, se uma imagem provavelmente não foi gerada a partir da outra).

Formalmente, seja $\mathcal{T}$ uma família de transformações em imagens, e seja T uma transformação tal que $T \in \mathcal{T}$. Podemos definir a função de dissimilaridade entre duas imagens $\mathcal{I}_A$ e $\mathcal{I}_B$ como o menor valor de $d_{\mathcal{I}_A,\mathcal{I}_B}$.

$$d_{\mathcal{I}_A,\mathcal{I}_B} = |\mathcal{I}_B - T_{\overrightarrow{\beta}}(\mathcal{I}_A)|_{\text{método de comparação ponto a ponto } \mathcal{L}} \tag{2.1}$$

para todos os possíveis valores de β que parametriza $\mathcal{T}$. A Equação 2.1 calcula a quantidade de dados residuais entre a melhor transformação de $\mathcal{I}_A$ para $\mathcal{I}_B$, de acordo com a família de transformações $\mathcal{T}$ e $\mathcal{I}_B$. Por fim, as imagens são comparadas utilizando algum método de comparação ponto-a-ponto $\mathcal{L}$.

Dado um conjunto de duplicadas de imagem, os autores criam a AFI do conjunto da seguinte maneira: primeiramente, é calculada a dissimilaridade entre cada par de imagens do conjunto. Dado um par de imagens $\mathcal{I}_A$ e $\mathcal{I}_B$, a dissimilaridade é calculada seguindo quatro etapas:

1. *Registro de imagens*, onde é estimada a transformação (rotação, escala, recorte, etc.) que deve ser aplicada na imagem $\mathcal{I}_B$ para que esta tenha as mesmas (ou semelhantes) características geométricas da imagem $\mathcal{I}_A$;

2. *Casamento de cor*, onde se faz a transferência das características de cor da imagem $\mathcal{I}_A$ para uma imagem $\mathcal{I}_B$;

3. *Casamento de compressão JPEG*, que visa garantir que a imagem $\mathcal{I}_B$ possua os mesmos fatores de qualidade de compressão JPEG que a imagem $\mathcal{I}_A$;

4. *Cálculo de dissimilaridade*, feita por meio do cálculo da soma dos erros quadrados (como sendo a técnica $\mathcal{L}$) entre a imagem $\mathcal{I}_A$ e a imagem resultante das três etapas anteriores $\mathcal{I}'_B$.

Fazendo isso para todas as imagens do conjunto, ao final, os autores obtêm uma matriz de dissimilaridade $D_{n\times n}$ (onde n é o número de duplicatas de imagens avaliadas e cada região da matriz representa a dissimilaridade entre um par de imagens). Considerando essa matriz como um grafo completo, os autores encontram a AFI propondo uma extensão, para grafos orientados, do algoritmo para cálculo de árvore geradora mínima de Kruskal [19].

Uma extensão do trabalho de Dias et al. [9] é apresentada em [11]. Nesse trabalho, os autores verificam o comportamento do algoritmo de Kruskal orientado em um ambiente controlado, controlado com perda de nós, ambiente com vários conjuntos de duplicatas e ambientes não controlados. Para um ambiente controlado, os autores assumem que todas as imagens de um conjunto são duplicatas próximas, e a abordagem é executada como apresentada em [9]. Em um ambiente controlado com perda de nós, os autores conhecem a estrutura original da árvore

filogenética, mas o algoritmo é executado considerando somente um subconjunto dos nós. Assim, os "filhos" desses nós devem estar conectados aos ancestrais mais próximos na árvore filogenética final. A Figura 2.1 ilustra este caso.

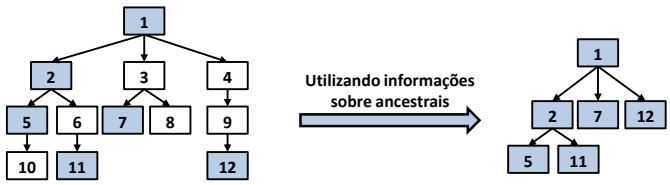


Figura 2.1: A partir de uma árvore filogenética reconstruída por meio de um algoritmo de filogenia de imagens (acima), a estrutura da árvore deve ser reconstruída nos casos de algum nó ser removido (nós brancos) e a árvore resultante é gerada considerando informações dos ancestrais das imagens.

Os autores avaliam a abordagem considerando a porcentagem de acerto no cálculo da raiz (imagem original), das folhas da árvore, da relação "pai-filho" entre cada imagem e a relação de ancestralidade (isto é, se uma imagem é ancestral de outra na árvore filogenética). Para os ambientes controlados, os autores reportam bons resultados, onde a raiz da árvore (i.e., a imagem original) é encontrada em quase 100% dos casos, e as demais relações na árvore são detectadas corretamente em mais de 80% dos casos.

Já em um ambiente não controlado, os autores assumem que não possuem qualquer evidência da relação "pai-filho" entre duas imagens. A robustez do algoritmo de filogenia é avaliada da seguinte maneira: dado um conjunto de duplicatas de imagens obtidas da Internet, é feita a execução do algoritmo de Kruskal orientado sobre elas, gerando uma árvore filogenética a qual os autores consideram, a princípio, que é a original. Em seguida, para cada imagem, é gerada uma duplicata direta. O algoritmo de Kruskal é executado novamente, contemplando as imagens obtidas da Internet e as duplicatas geradas. Por fim, as duas árvores filogenéticas obtidas são comparadas para se analisar se as informações de ancestralidade são mantidas. Para esse caso, os autores reportam uma taxa de reconstrução da árvore com 54% de acerto, considerando o acerto na relação "pai-filho" e de ancestrais da árvore reconstruída em relação à árvore original.

Em [10], os autores aplicam os conceitos apresentados em [9] para encontrar a árvore que representa a relação entre um conjunto de duplicatas de vídeos (Árvore Filogenética de Vídeos – AFV). Basicamente, os autores seguem os mesmos passos para encontrar AFI's, com uma adaptação no cálculo da dissimilaridade. Para esse caso, a dissimilaridade entre dois vídeos é calculada utilizando-se f quadros de cada vídeo, assumindo que estes são temporalmente coerentes. A filogenia do conjunto de vídeos é construída considerando cada quadro separadamente e, para cada vídeo, o seu ancestral direto ("pai") é definido por meio do cálculo de uma matriz de parentesco calculada considerando o resultado do algoritmo de Kruskal orientado para cada quadro separadamente. Os autores reportam que a raiz da árvore é encontrada em 91% dos casos.

Na abordagem apresentada em [7], os autores exploram outros algoritmos para reconstrução

da árvore filogenética de um conjunto de duplicatas de imagens. A primeira técnica é baseada na utilização do clássico algoritmo de Prim [26] adaptado para grafos orientados (*Oriented Prim*). Inicialmente, para cada vértice do grafo, a árvore filogenética é calculada, considerando o nó escolhido como raiz. Em seguida, o custo da árvore (isto é, a soma do peso de todas as arestas) é calculado. Após avaliar todas as possíveis raízes, a árvore filogenética final é a que possuir o menor custo. Essa abordagem apresentou baixa eficácia na reconstrução da árvore filogenética, considerando a relação de ancestralidade entre as duplicatas de um conjunto. A segunda abordagem se baseia na utilização do algoritmo de arborescência ótima proposto por Edmonds [12]. Essa abordagem se mostrou mais eficaz na reconstrução da árvore filogenética, comparada com as abordagens de Kruskal e Prim orientado.

Os autores também realizaram alguns experimentos iniciais com florestas filogenéticas. Em [11], os autores executam o algoritmo de Kruskal orientado para árvores e, caso se deseje encontrar k árvores (onde k é fornecido pelo usuário), remove-se as $k-1$ arestas mais pesadas da árvore gerada inicialmente. A Figura 2.2 ilustra esta situação.

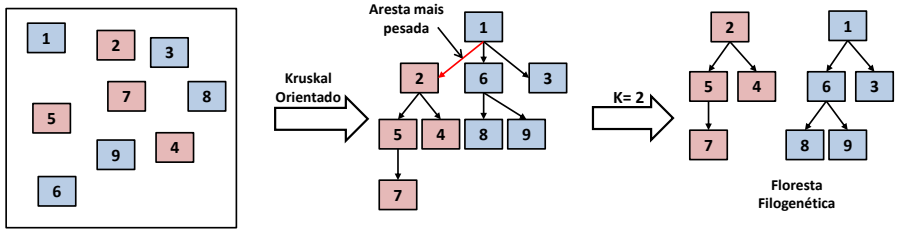


Figura 2.2: Solução inicial para geração de florestas com o algoritmo de Kruskal orientado [11].

Já em [8], os autores assumem que o número de árvores filogenéticas a serem formadas para representar a relação entre as imagens de um conjunto é desconhecido. Assim, o procedimento realizado para adicionar as arestas é realizado da mesma forma que apresentado em [11], com a diferença que as arestas são adicionadas na árvore resultante até o momento em que o peso da próxima aresta for maior que um limiar τ multiplicado pelo desvio padrão do peso das arestas adicionadas no grafo até o momento. Assim, o corte na árvore inicial é feito de forma automática.

Os conceitos sobre filogenia de documentos podem ser aplicados para se encontrar a relação entre duplicatas de outros tipos de documentos, além de imagens e vídeos. Um exemplo é o trabalho de Andrews et al. [1], no qual os autores propõem um modelo generativo de classificação para se identificar a filogenia de nomes. Primeiramente, os autores assumem que um nome não foi simplesmente criado, mas é derivado de outro nome similar. Em seguida, os autores utilizam um algoritmo baseado em Esperança/Maximização para re-estimar a filogenia e os parâmetros do classificador. Com isso, os autores encontram variantes de um nome em um conjunto de nomes (e.g., “Filipe de Oliveira Costa”, “Filipe de Oliveira” , “Filipe Costa”, etc.).

Recentemente, Nucci et al. [25] apresentaram uma abordagem para tratar o problema de filogenia de áudio. Para isso, os autores geraram um conjunto de duplicatas considerando operações de compressão, corte, *fade in* e *fade out* nos arquivos de áudio. Em seguida, a

dissimilaridade entre um par de duplicatas I_A e I_B qualquer é calculada por meio do cálculo de SNR (*Signal to Noise Ratio*).

$$SNR(I_A, I_B) = 20log_{10}\left(\frac{||I_B||_2}{||I_A - I_B||_2}\right) \tag{2.2}$$

Assim, após o cálculo da matriz de dissimilaridade do conjunto de duplicatas, os autores encontram a árvore filogenética do conjunto por meio do algoritmo de Kruskal orientado, proposto por Dias et al. [9].

Capítulo 3

Metodologia e desafios do projeto

3.1 Metodologia

Para a pesquisa do doutorado, iremos abordar dois tipos de análise: a construção da matriz de dissimilaridade entre os documentos de um conjunto de duplicatas e a reconstrução da estrutura filogenética (árvore ou floresta).

3.1.1 Cálculo da matriz de dissimilaridade

O cálculo de dissimilaridade entre um par de duplicatas será realizado, inicialmente, de forma similar à apresentada por Dias et al. [10, 11]. Visando melhorar a eficácia de nossa abordagem, iremos investigar outras formas de se realizar a etapas do cálculo da dissimilaridade (por exemplo, outras formas de fazer o casamento de compressão de vídeos, utilização de outras técnicas para o casamento de cor entre as imagens, etc.).

3.1.2 Algoritmo de reconstrução de Floresta Filogenética

A reconstrução da estrutura (árvore ou floresta) que representa as relações entre as duplicatas de um conjunto será feita utilizando, primeiramente, algoritmos baseados em grafos. Consideramos que a matriz de dissimilaridade de um conjunto é um grafo completo e, a partir desse grafo, selecionamos as arestas que farão parte da floresta filogenética final. A princípio, estudaremos o algoritmo proposto independentemente por Chu & Liu [5], Bock [3] e Edmonds [12], que calcula a arborescência mínima de um grafo, Acreditamos que tal abordagem possa ser promissora para uma solução inicial do problema de filogenia de documentos. Investigaremos, também, abordagens complementares como, por exemplo, o comportamento dos algoritmos propostos em cenários de conexões filogenéticas faltantes (*missing links*), conforme inicialmente estudado em [11]. Finalmente, iremos analisar a proposição de abordagens automáticas para definição de florestas filogenéticas a partir da exploração de propriedades de clusterização [32], simetrização de matrizes, algoritmos de filogenia do campo da Biologia Computacional [2, 30], algoritmos de teoria de grafos, entre outros.

3.1.3 Medidas de avaliação

Assim como é feito por Dias et al. [9], dadas as florestas $\mathcal{F}_{\text{Original}}$ (floresta original, isto é, que representa corretamente as relações das duplicatas próximas de um conjunto) e $\mathcal{F}_{\text{Reconstruída}}$ (a floresta calculada pelo algoritmo de reconstrução), a medida de eficácia de nossa abordagem será baseada no cálculo de quatro métricas de avaliação: a porcentagem de acerto de *RAÍZES* (documento original), de *ARESTAS* (acerto das arestas que compõem a floresta final), das *FOLHAS* (documentos os quais não possuem duplicatas geradas a partir destes) e de *ANCESTRAIS* (relação de ancestralidade entre dois nós da floresta), considerando que a floresta original está disponível. A avaliação das florestas é feita seguindo a equação a seguir.

$$\text{EM}(\text{IPF}_1, \text{IPF}_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2} \tag{3.1}$$

onde EM representa uma das quatro métricas utilizadas para avaliação, IPF_1 é a floresta reconstruída com elementos representados por S_1 e IPF_2 é a floresta original que possúi os elementos S_2.

3.2 Desafios

No desenvolvimento deste trabalho, iremos enfrentar os seguintes desafios:

- Um primeiro desafio do projeto é analisar outros algoritmos para reconstrução de uma árvore filogenética que seja mais eficaz, para então adaptarmos o algoritmo encontrado para tratar florestas filogenéticas. Dias et al. [7] apresentaram resultados que mostram que o algoritmo para encontrar arborescência mínima se mostrou promissor na montagem correta da árvore filogenética. Também iremos analisar outros algoritmos existentes na área de Biologia Computacional. Conforme já estudamos inicialmente, para isso, precisaremos considerar técnicas que simetrizem as informações de dissimilaridade presentes atualmente na matriz de relacionamento entre pares de duplicatas ou adaptar essas abordagens para considerarem um espaço assimétricos, uma vez que os algoritmos existentes em Biologia Computacional funcionam apenas em espaços métricos.

- Outro desafio deste projeto é encontrar novas formas de calcular a dissimilaridade, de forma a melhorar a eficácia do algoritmo de reconstrução de florestas. Uma ideia é trabalhar com compressão, uma vez que a compressão em documentos multimídia muitas vezes é uma compressão com perdas, o que torna o processo irreversível. Mudanças bruscas de cor também podem ser exploradas pelo mesmo motivo. Uma investigação inicial foi realizada com algoritmos de transferência de cor. Dias et al. [9] utilizaram o algoritmo proposto por Reinhard et al. [27]. Em um experimento exploratório, fizemos o mesmo processo para o cálculo de dissimilaridade utilizando a proposta de Xiao e Ma [34] para a etapa de casamento de cor. Somente com essa troca, os resultados apresentaram uma leve melhora em termos de reconstrução da árvore filogenética. Como

esse foi um experimento inicial, pretendemos fazer mais experimentos nesse sentido e explorá-lo também de um ponto de vista mais formal visando encontrar uma forma para o cálculo de dissimilaridade mais precisa. Uma outra possível abordagem é a combinação das características complementares. Por exemplo, podemos considerar as informações de dissimilaridade exploradas por De Rosa et al. [29], principalmente, em relação ao uso do ruído de foto-responsividade não-uniforme [6, 22], assunto que, inclusive, foi foco do trabalho de mestrado realizado pelo candidato.

- A elaboração do algoritmo de reconstrução da floresta filogenética também apresenta desafios importantes. Possíveis soluções são o corte das arestas mais pesadas da árvore filogenética gerada inicialmente e re-execução do algoritmo de reconstrução após o corte e clusterização inicial antes da execução do algoritmo de reconstrução. Outro desafio é, em um ambiente onde não sabemos *a priori* quantas árvores devem ser geradas, decidir qual a forma utilizada para separar as imagens em árvores diferentes e qual o critério de parada do algoritmo (isto é, quando decidir se uma aresta deve ser removida para dividir uma árvore em duas).

- Outro grande desafio do trabalho será trabalhar com um ambiente não controlado, isto é, um ambiente onde não conhecemos a relação entre os documentos do conjunto para futura comparação e verificação de performance dos algoritmos propostos. Este tipo de ambiente foi explorado inicialmente por Dias et al. em [11] e pretendemos expandir este tipo de análise. Uma forma é avaliar as abordagens propostas com arquivos multimídia obtidos da Internet. Nesse caso, teremos que formalizar novas abordagens para verificação da eficácia dos algoritmos propostos.

Capítulo 4

Etapas do projeto

4.1 Cronograma

Este projeto tem as seguintes atividades programadas:

1. Obtenção de créditos em disciplinas;
2. Revisão da literatura;
3. Geração de conjuntos de duplicadas de imagens;
4. Investigação e proposição de novos mecanismos para cálculo das matrizes de dissimilaridade e novos algoritmos para reconstrução da floresta filogenética que represente um conjunto, considerando como objeto de estudo imagens digitais;
5. Qualificação de Doutorado;
6. Programa de Estágio Docente (PED)
7. Geração de conjuntos de duplicadas de vídeos;
8. Investigação e proposição de novos mecanismos para cálculo das matrizes de dissimilaridade e novos algoritmos para reconstrução da floresta filogenética que represente um conjunto, considerando como objeto de estudo vídeos digitais;
9. Doutorado sanduíche no *Politecnico di Milano*, em Milão, Itália;
10. Publicação dos resultados obtidos;
11. Escrita da tese;
12. Defesa da tese de doutorado.

A Tabela 4.1 apresenta as etapas do projeto com durações em bimestres.

O cronograma do projeto possui quatro anos de duração, sendo que o segundo semestre de 2014 está reservado para o doutorado sanduíche no *Politecnico di Milano*, na Itália com os parceiros integrantes do projeto de pesquisa maior a qual esse projeto de doutorado será parte integrante (c.f., Seção 1).

Tabela 4.1: Cronograma do projeto.

Ano		2012			2013						2014						2015						2016		
Bimestre		4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3
Etapas	1	•	•	•	•	•	•																		
	2	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•						
	3				•	•	•																		
	4					•	•	•	•	•	•														
	5									•															
	6										•	•	•												
	7										•	•	•												
	8										•	•	•	•	•	•	•								
	9													•	•	•									
	10									•			•				•				•				•
	11																				•	•	•	•	•
	12																								•

4.2 Disseminação e avaliação

Os resultados das pesquisas são de potencial interesse para as comunidades de Computação Forense, Processamento de Imagens e Visão Computacional. Os indicadores de sucesso serão:

- Publicação de **artigos científicos** nas melhores revistas científicas e conferências nacionais e internacionais;
- Novos algoritmos, modelos e métodos para filogenia de documentos digitais.

4.3 Experimentos iniciais

Alguns esforços iniciais foram realizados considerando, como objeto de estudo, duplicatas de imagens digitais, visando melhorar a eficácia da proposta já existente para filogenia de imagens de Dias et al. [8, 9, 11] (considerada o estado da arte atualmente) para construção de florestas filogenéticas.

Na literatura atual de filogenia multimídia, existem basicamente dois métodos para reconstrução de florestas filogenéticas que estão sendo melhor exporados pela comunidade acadêmica: Kruskal orientado [11] e o algoritmo de arborescência ótima de Edmonds [12]. Visando aumentar a eficácia dos métodos na reconstrução de florestas, nós realizamos uma comparação entre os resultados de ambas as técnicas para observar a complementaridade entre eles e verificar como os resultados na reconstrução das florestas podem ser melhorados por meio da fusão das técnicas.

Os resultados obtidos com esses experimentos serão submetidos para publicação. O artigo pode ser encontrado, em sua forma atual de preparação, no documento anexo a este.

Referências Bibliográficas

[1] N. Andrews, J. Eisner, and M. Dredze. Name phylogeny: a generative model of string variation. In *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 344–355, 2012.

[2] C. Bachmaier, U. Brandes, and B. Schlieper. *Drawing phylogenetic trees*. Springer, 2005.

[3] F. Bock. An algorithm to construct a minimun directed spanning tree in a directed network. *Developments in Operations Research*, pages 29–44, 1971.

[4] X. Cheng and L.-T. Chia. Stratification-based keyframe cliques for removal of near-duplicates in video search results. In *ACM Intl. Conference on Multimedia Information Retrieval(MIR)*, pages 313–322, 2010.

[5] Y. J. Chu and T. H. Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400, 1965.

[6] F. O. Costa, M. Eckmann, W. J. Scheirer, and A. Rocha. Open set source camera attribution. In *IEEE Conference on Graphics, Pattern and Images (SIBGRAPI)*, pages 71–78, 2012.

[7] Z. Dias, S. Goldenstein, and A. Rocha. Exploring heuristic and optimum branching algorithms for image phylogeny. *Elsevier Journal of Visual Coimmunication and Image Representation*, 24(7):1124–1134, October 2013.

[8] Z. Dias, S. Goldenstein, and A. Rocha. Toward image phylogeny forests: Automatically recovering sematically similar image relationships. *Forensic Science International*, 231:178–189, 2013.

[9] Z. Dias, A. Rocha, and S. Goldenstein. First steps towards image phylogeny. In *IEEE Intl. Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2010.

[10] Z. Dias, A. Rocha, and S. Goldenstein. Video phylogeny: Recovering near-duplicate video relationships. In *IEEE Intl. Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2011.

[11] Z. Dias, A. Rocha, and S. Goldenstein. Image phylogeny by minimal spanning trees. *IEEE Transactions on Information Forensics and Security (TIFS)*, 7(2):774–788, April 2012.

[12] J. Edmonds. Optimum branchings. *Journal of Research of National Institute of Standards and Technology (NIST)*, 71B:48–50, 1967.

[13] Z. Fan and R. L. Queiroz. Identification of bitmap compression history: Jpeg detection and quantizer estimation. *IEEE Transactions on Image Processing*, 12(2):230–235, 2003.

[14] J. Fridrich, D. Soukal, and J. Lukas. Detection of copy-move forgery in digital images. In *Digital Forensics Research Conference (DFRWS)*, 2003.

[15] A. Jaimes, S. fu Chang, and A. Loui. Duplicate detection in consumer photography and news video. In *ACM Workshop on Multimedia and Security*, pages 423–424, 2002.

[16] A. Joly, O. Buisson, and C. Frélicot. Content-based copy retrieval using distortion-based probabilistic similarity search. *IEEE Trans. Multimedia*, 9(2):293–306, 2007.

[17] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and subimage retrieval. In *ACM Workshop on Multimedia and Security*, pages 869–876, 2004.

[18] L. Kennedy and S.-F. Chiang. Internet image archaeology: Automatically tracing the manipulation history of photographs on the web. In *Proc. ACM Intl. Conference of Multimedia*, pages 349–358, 2008.

[19] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. of the American Mathematical Society*, 7(1):48–50, 1956.

[20] B. Lewin. *Genes VI.* Oxford Univ. Press, 1997.

[21] Z. X. H. Ling, F. Zou, Z. Lu, and P. Li. Robust image copy detection using multi-resolution histogram. In *ACM Intl. Conference on Multimedia Information Retrieval(MIR)*, pages 129–136, 2010.

[22] J. Lukáš, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security (TIFS)*, 1(2):205–214, 2006.

[23] J. Mao, O. Bulan, G. Sharma, and S. Datta. Device temporal forensics: an information theoretic approach. In *IEEE international Conference on Image processing (ICIP)*, pages 1485–1488, 2009.

[24] Y. Maret. *Efficient Duplicate Detection Based on Image Analysis.* Ph.d. thesis, École Polytechinique Fédérale de Lausanne, 2007.

[25] M. Nucci, M. Tagliasacchi, and S. Tubaro. A phylogenetic analysis of near duplicate audio tracks. In *IEEE Intl. Workshop on Multimedia Signal Processing (MMSP)*, 2013.

[26] R. C. Prim. Shortest connection networks and some generalizations. *Bell system technical journal*, 36(6):1389–1401, 1957.

[27] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics Applications*, 21:34–41, 2001.

[28] A. Rocha, W. Scheirer, T. E. Boult, and S. Goldenstein. Vision of the unseen: Current trends and challenges in digital image and video forensic. *ACM Computing Surveys (CSUR)*, 42(26):26:1–26:42, October 2011.

[29] A. De Rosa, F. Uccheddu, A. Costanzo, A. Piva, and M. Barni. Exploring image dependencies: a new challenge in image forensics. *SPIE-IS&T Electronic Imaging*, 7541:774–788, 2010.

[30] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[31] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *European Conference on Computer Vision (ECCV)*, pages 414–431, 2002.

[32] P. H. A. Sneath and R. R. Sokal. Unweighted pair group method with arithmetic mean. *Numerical Taxonomy*, pages 230–234, 1973.

[33] E. Valle. *Local descriptor matching for image identification systems.* Ph.d. thesis, Universite de Cergy-Pontoise, June 2008.

[34] X. Xiao and L. Ma. Color transfer in correlated color space. *ACM Intl. Conference on Virtual Reality continuum and series (VRCIA)*, pages 305–309, 2006.

[35] D.-Q. Zhang and S. F. Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *ACM Workshop on Multimedia and Security*, pages 877–884, 2004.