Indexação de Grandes Volumes de Áudio e Vídeo usando Reconhecimento de Voz

Gustavo Dibbern Piva[1], Eduardo Javier Huerta Yero[2]

[2] Departamento de Ciência da Computação – Faculdade de Campo Limpo Paulista – Campo Limpo Paulista – São Paulo - Brasil.

gpiva@mpc.com.br, huerta@cc.facamp.br

Abstract. *The amount of digital data available worldwide has grown exponentially in recent years. This phenomenon presents serious challenges to traditional data management mechanisms for indexing and information retrieval. Data in audio and video, usually difficult to manage and index, are among the types of data whose quantity and availability has increased more explosively in recent years. We propose to develop scalable parallel computing mechanisms for indexing and searching large volumes of audio and video using existing voice recognition techniques.*

Resumo. *A quantidade de dados disponível digitalmente no mundo tem crescido em ritmo exponencial nos últimos anos. Este fenômeno apresenta sérios desafios aos mecanismos tradicionais de gerenciamento, indexação e pesquisa de informações. Dados em formato de áudio e vídeo, usualmente difíceis de gerenciar e indexar, estão entre os tipos de dados cuja quantidade e disponibilidade tem aumentado de forma mais explosiva nos últimos anos. Nos propomos a criar mecanismos para indexar e pesquisar grandes volumes de dados de áudio e vídeo utilizando recursos de reconhecimento de voz, técnicas escaláveis de computação paralela e algoritmos de ranking e indexação que permitam gerenciar em tempo hábil grandes volumes de dados.*

1. Introdução

A quantidade de informação digital disponível vem crescendo exponencialmente nos últimos anos. Dados de 2012 estimam que aproximadamente 2.5 hexabytes são gerados por dia, enquanto a quantidade total de dados disponíveis foi estimada em 2.7 zetabytes. A explosão na criação de dados é tal que 90% dos dados disponíveis digitalmente até 2012 tinham sido criados nos últimos 2 anos [Mewawalla2012].

O termo *Big Data* tem sido utilizado para descrever conjuntos de dados desta magnitude, que não podem ser gerenciados por ferramentas tradicionais em tempo hábil. As dificuldades incluem a captura, armazenamento, pesquisa, compartilhamento, transferência, análise e visualização destes dados. Como consequencia, nos últimos anos o investimento em pesquisa e desenvolvimento nesta área tem crescido, tanto por instituições acadêmicas como pela indústria.

Uma parte significativa dos dados armazenados hoje está no formato de áudio e vídeo. O aumento recente no uso de smartphones e câmeras digitais capazes de capturar fotos, vídeos e gravar conversas, junto com a adoção maciça de sites que permitem

compartilhar este tipo de informação (e.g. YouTube, Facebook) tem estimulado o aumento da quantidade de dados disponíveis neste formato. Outras fontes importantes complementam este cenário, tais como imagens produzidas por satêlites, conversas telefônicas, reportagens radiofônicas e televisivas, dentre outras.

Uma das formas de indexar arquivos de áudio e vídeo é através das palavras que neles são faladas. Desta forma, seria possível pesquisar arquivos que contenham um determinado conjunto de palavras e organizar os resultados de acordo com a sua relevância, tal como é feito para pesquisar conteúdo textual na Web. Um sistema com estas características, e capaz de gerenciar de forma apropriada grandes quantidades de arquivos de áudio e vídeo, seria de bastante utilidade em diversos cenários.

Nós propomos desenvolver uma solução para indexar e pesquisar grandes volumes de áudio e vídeo usando técnicas de reconhecimento de voz. Para tanto propomos usar técnicas de processamento paralelo, usualmente utilizadas em cenários de *Big Data*, tais como MapReduce [Dean2008] e Bulk-synchronous parallel (BSP) [Valiant1989]. Além disso, propomos estudar algoritmos de ranking e indexação apropriados que nos permitam pesquisar o conteúdo e apresentar os resultados organizados de acordo com a sua relevância.

2. Trabalhos Relacionados

Os arquivos de aúdio e vídeo na atualidade não são comumente indexados usando técnicas de reconhecimento de fala. A Microsoft tem investido em pesquisa nesta área e disponibiliza comercialmente o MAVIS [Microsoft2011], uma plataforma com características similares às propostas neste documento. Em particular, o MAVIS usa o Windows Azure, a plataforma de Cloud Computing da Microsoft, para realizar o reconhecimento de voz nos arquivos e o SQL Server para armazenar o conteúdo indexado para pesquisa posterior. O MAVIS é uma plataforma comercial, disponível apenas no ecosistema da Microsoft e que não suporta português brasileiro.

Lawto apresenta em [Lawto2011] uma plataforma para indexar reportagens noticiosas vindas de diversas fontes e em várias linguagens (dentre as quais não está o português). Além de focar em um tipo específico de arquivo não há indicativos de que haja preocupação no projeto pela quantidade de informação a ser indexada. A solução proposta por eles indexa diariamente reportagens vindas de um conjunto de fontes bem estabelecido e cujo tamanho é tal que pode ser processado usando técnicas tradicionais.

3. Caracterização do Problema

O problema de processar, indexar e pesquisar um conjunto grande de arquivos de áudio e vídeo usando técnicas de reconhecimento de voz apresenta desafios de características diferentes, alguns dos quais listamos a seguir.

Volume: o tamanho do conjunto de dados impossibilita o uso de técnicas de processamento sequenciais. É desejável também que o sistema resultante do projeto seja capaz de se adequar ao aumento da quantidade de dados sem perder sua eficácia. O volume de dados também exige o uso de técnicas de indexação, de forma tal que os arquivos de áudio e vídeo não precisem ser processados a cada consulta feita.

Heterogeneidade: Os arquivos encontram-se usualmente disponíveis em vários formatos, cada um com características específicas.

Reconhecimento de voz: No processo de reconhecimento de voz, existem vários problemas a serem levados em conta, tais como, o ruído ambiente onde o som foi gravado, distorção do canal de gravação, o sotaque do interlocutor, a falta e/ou excesso de fluência do interlocutor, o uso de gírias, neologismos e expressões regionais, como mais alguns dos problemas que deverão ser encontrados no reconhecimento e transformação dos sons em texto [Saon2012].

Indexação: o mecanismo de indexação de resultados deve permitir, preferencialmente, acessar o trecho do arquivo de áudio e/ou vídeo onde aparecem as palavras pesquisadas.

4. Infraestrutura da Solução

Nos propomos a desenvolver uma infraestrutura capaz de ser uma solução para o problema. Para isso, a solução deverá contar com os seguntes elementos.

Sistema para o processamento paralelo de conjuntos de dados de grandes proporções: sistemas deste tipo devem ser capazes de acomodar o aumento do tamanho do conjunto de dados a ser processado sem perder desempenho. Usualmente esse objetivo é atingido pela combinação do aumento do poder computacional (e.g. adicionando nós de processamento) com o uso de um modelo de programação escalável que não crie gargalos quando o número de nós de processamento e o tamanho dos dados aumenta. Modelos como MapReduce e BSP tem se mostrado capazes de atender estas restrições e deverão ser considerados como candidatos neste projeto.

Software para reconhecimento de voz: não é objetivo deste projeto fazer contribuições na área de reconhecimento de voz. Neste projeto escolheremos algum software já existente e testado, de preferência que reconheça português brasileiro, tal como o apresentado em [Neto2011].

Algoritmo de indexação: o resultado produzido pelo software de reconhecimento de voz deve ser indexado e armazenado em um Sistema de Gerenciamento de Bancos de Dados (SGBD), de forma tal que possa ser consultado posteriormente. O algoritmo de indexação deve levar em consideração as características particulares do conteúdo sendo indexado e preferencialmente deve permitir que os arquivos de aúdio e vídeo possam ser acessados diretamente nos trechos em que as palavras pesquisadas aparecem.

SGBD: O SGBD agirá como ponte entre a parte batch do sistema (processamento paralelo dos arquivos de aúdio e vídeo usando um software de reconhecimento de voz e posterior indexação dos resultados) com a parte online, que consiste em responder às pesquisas feitas pelos usuários.

Algoritmo de ranking: os resultados das pesquisas devem ser apresentados aos usuários organizados de acordo com a sua relevância, tal como acontece com os algoritmos tradicionalmente usados para pesquisar texto na web.

Interface para consulta: esta interface, que será disponibilizada ao usuário final para pesquisar o conteúdo indexado, pode tomar a forma de uma aplicação standalone ou, preferívelmente, de uma aplicação Web acessível desde um navegador comum..

5. Contribuições Esperadas

Projetamos a construção de uma infraestrutura que permita processar um conjunto grande de arquivos de aúdio e vídeo usando técnicas de reconhecimento de voz e indexar os resultados obtidos para sua posterior pesquisa. A infraestrutura consistirá em uma plataforma de hardware formada por um conjunto de nós de processamento ligados por uma rede dedicada de alta velocidade, um programa paralelo para processar os arquivos, um software de reconhecimento de fala instalado em cada um dos nós de processamento, um algoritmo de indexação, um SGBD instalado em um ou vários dos nós de processamento e uma interface de pesquisa, que pode ser standalone ou ser baseada na Web, em cujo caso o servidor Web também deverá ser hospedado em um ou vários dos nós de processamento.

6. Referencias

Dean, J.; Ghemawat, S. (2008).MapReduce: Simplified Data Processing on Large Clusters. ACM Digital Library, 107-113.

Lawto, J.; Gauvain, J.; Lamel, L.; Grefenstete, G.; Gravier, G.; Despres, J.; Guinaudeau, C.; Sébillot, P. (2011). A scalable video search engine based on audio content indexing and topic segmentation. CoRR (abs/1111.6265)

Mewawalla, C. (2012). Big Data. Global investment themes: telecoms, media and technology, 3-19.

Microsoft Audio Video Indexing Service (2011). Disponível em: <http://research.microsoft.com/en-us/projects/mavis/>. Acesso em: 05/07/2014.

Neto, N.; Patrick, C.; Klautau, A.; Trancoso, I. (2011). Free Tools and resources for Brazilian Portuguese speech recognition. Journal of the Brazilian Computer Society. Volume 17 (1) pp 53-68. ISSN 0104-6500.

Saon,G.;Jen-Tzung C. Large-Vocabulary Continuous Speech Recognition Systems: A look at Some Recent Advances IEEE Signal Processing Magazine November 2012. 18-33

Valiant, L. G. (1990). A bridging model for parallel computation. Communications of the ACM, 103-111.