# ANOTAÇÃO SEMÂNTICA DE TRANSCRITOS PARA INDEXAÇÃO E BUSCA DE VÍDEOS

Sandro Athaide Coelho
*Instituto de Ciências Exatas – Universidade Federal de Juiz de Fora*
*Campus Universitário, Juiz de Fora, Minas Gerais, Brasil*

Jairo Francisco de Souza
*Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora*
*Campus Universitário, Juiz de Fora, Minas Gerais, Brasil*

**ABSTRACT**

Ao longo dos anos de existência da Internet, o montante de dados disponibilizados em muitos formatos, como texto, áudio e vídeo cresceu significativamente. O fenômeno da explosão de dados gerou grandes desafios, principalmente em arquiteturas que promovam o reuso de dados e na indexação de arquivos multimídia. Ainda hoje, usuários não são capazes de usufruir completamente das vantagens das buscas em arquivos de vídeos, por conta de lacunas na integração e classificação dos dados. Neste artigo apresentamos um *framework* para indexação de grandes volumes de vídeos através da anotação semântica de transcrições automáticas, a qual pode ser configurada para diferentes aplicações. O *framework* emprega técnicas de extração de informação objetivando maximizar a acurácia dos métodos existentes para o idioma Português, habilitando a classificação em tempo real dos arquivos multimídia e permitindo a interoperabilidade destes dados com *datasets* públicos.

**KEYWORDS**

Recuperação de informação multimídia; dados ligados; indexação semântica; objetos de aprendizagem.

## 1. INTRODUÇÃO

Ao longo dos anos de existência da Internet, a facilidade de difusão da informação em seus mais variados formatos (texto, áudio, vídeo) fez surgir o desafio de indexar, formatar e disponibilizar tais conteúdos de forma que possam ser reutilizados e ainda resultem em um aprimoramento na experiência do usuário [2].

Considerando que a maior parte dos buscadores ainda se utilizam basicamente de indexação de informação textual [2], há uma necessidade de indexar de forma eficaz arquivos de áudio e vídeo [26,27]. Para muitos buscadores, a recuperação de vídeo é realizada utilizando metadados, texto encontrado no contexto do vídeo (ao redor do arquivo), entre outras informações textuais [26,27]. Alguns sistemas buscam categorizar vídeos através de técnicas de reconhecimento de objetos e ações. Somar estas técnicas com técnicas baseadas em texto podem trazer resultados relevantes nesta área de pesquisa.

Para facilitar a indexação de vídeos, é comum a seleção e marcação de palavras-chave, de forma manual que identificam e descrevem o conteúdo [16]. Esta prática facilita o processo de busca por deixar explícito palavras de contexto para o vídeo. Entretanto, é possível inferir que o problema não é completamente solucionado, pois além do tempo despendido pela pessoa que está realizando a classificação, tal técnica condiciona a catalogação e formatação da informação às experiências pessoais [9,17], reduzindo a eficácia dos métodos de busca existentes. De forma menos precisa, sistemas web podem fazer uso da indexação de *tags* encontradas em wikis e blogs [15] para tratar esse problema sem muito esforço manual.

Outra técnica usada para melhorar a busca de vídeos é a utilização de textos transcritos do vídeo, permitindo que o vídeo (ou áudio) possa ser recuperado através de palavras faladas ao longo da sua reprodução [10,17]. Ainda que a utilização da transcrição, uma vez que é realizada de forma manual, é muito custosa e raramente está disponível para ser utilizada por sistemas de busca [10,17].

Considerando a problemática acima, é necessário o uso de técnicas próprias para indexação de grandes volumes de vídeos e áudios. Emissoras de rádio como a BBC, Rádio Globo, entre outras, contém um grande acervo de arquivos de áudio criados há décadas e que não possuem catalogação. Para tratar esse problema, o objetivo deste trabalho é propor um *framework* que possa ser utilizado para busca semântica em grandes volumes de arquivos de áudio e vídeo. *Esse framework* emprega técnicas de extração de informação visando (1) maximizar a precisão dos métodos existentes para o idioma português, (2) classificar este tipo de mídia em tempo real, (3) permitir a interoperabilidade dos dados através da disponibilização das informações na forma de Dados Ligados, (4) integrar aos *Datasets* públicos, enriquecendo através de *facets* a experiência de navegação e, finalmente, (4) fornecer interfaces para integração com ferramentas de *Question Answering* [37].

Este artigo está organizado como se segue. Na Seção 2 são apresentados os conceitos básicos relativos a extração da informação, a sua representação e as técnicas de marcação semântica. Na seção 3 os principais trabalhos relacionados são listados e discutidos. Na Seção 4 será apresentada a proposta do *framework* Qodra para mecanismos de busca de vídeos e discutida a suas possíveis adaptações. Por fim, na Seção 5 são listadas as considerações finais e trabalhos futuros.

# 2. CONCEITOS E TÉCNICAS

As seções a seguir visam esclarecer os conceitos e técnicas empregadas no *framework*. Os mesmos estão organizados de forma a apresentar o processo de indexação e recuperação da informação, partindo dos arquivos de vídeo até a interface gráfica.

## 2.1 Automatic Speech Recognition (ASR)

O *Automatic Speech Recognition* (ASR) abrange um conjunto de técnicas criadas para permitir o reconhecimento da fala humana. O ASR habilita computadores a processar sinais, que juntamente com modelos estatísticos, realizam o reconhecimento do discurso, resultando nas transcrições textuais [13].

Ao receber um arquivo de áudio como entrada, o decodificador faz uso de algoritmos de processamento de sinais, tais como o Viterbi [4], que perfazem a busca de sequências acústicas utilizando modelos acústicos, de linguagem e o léxico. O modelo acústico é uma representação estatística de cada entrada do vocabulário construído à partir de um ou mais áudios. O modelo de linguagem consiste em um arquivo de regras que são utilizadas para a interpretação bayesiana do sinal acústico. Por fim, o modelo léxico contém o mapeamento das palavras do vocabulário. O resultado do ASR são as transcrições que mais se aproximam do som analisado.

Atualmente, é possível gerar transcrições aceitáveis para áudios em boas condições, como em diálogos bem estruturados, por exemplo. Neste tipo de áudio não ocorre excesso de ruídos e falas simultâneas [10]. Na literatura, são conhecidos problemas que degradam significantemente a acurácia das transcrições. Entre eles estão os áudios provenientes de diálogos livres - fato comum em entrevistas, por exemplo. Isto ocorre devido à utilização de palavras fora do contexto do vocabulário empregado durante a etapa de treinamento supervisionado do modelo. Ainda são apontados outros fatores como a distorção do canal de áudio, sotaques estrangeiros e fenômenos linguísticos como a homofonia [1]. A presença deste último fator resulta na transcrição incorreta de palavras com sons aproximados gerando inconsistências na semântica do texto.

Em ASR, o maior desafio é treinar os modelos e ajustar os dicionários para maximizar a corretude do texto transcrito. Este texto, além de compor o índice, será utilizado na expansão e interligação da informação com fontes de dados ligados disponíveis na web.

Para se conhecer a dimensão do problema, temos a seguir, na tabela 1 um trecho original e o transcrito gerado automaticamente do programa Voz do Brasil realizado no dia 09 de dezembro de 2013. Ao observar o resultado do transcrito, é possível verificar um caso de homofonia como na troca de Atlético por Poético. Há ainda palavras que não constam no dicionário e que por isso não foram transcritas, como Paranaense.

Tabela 1. Exemplo de texto transcrito automaticamente

| | Transcrito |
|---|---|
| Original | **"No fim de semana foi realizada a última rodada do Campeonato Brasileiro da Série A, e na partida entre Atlético Paranaense e Vasco quatro jovens ficaram feridos depois de se envolverem em uma briga."** |
| Transcrito | **"em semana foi realizada o tio aborda do campeonato brasileiro da série a ela parte da liga poético amanhã se faz com o quarto jovens ficaram feridos depois se vou em uma briga"** |

## 2.2 Anotação semântica de áudio e vídeo

Popov *et al* [21] descrevem a anotação semântica como uma técnica de atribuição de metadados para a geração e uso de *schemas* que habilitem novos métodos de acesso a informação, além de aprimorar os já existentes.

Ampliando um pouco mais o conceito, Körner *et al* [6] afirmam que a anotação semântica possibilita categorizar e descrever de forma precisa os recursos anotados, o que impacta positivamente na precisão dos mecanismos de recuperação da informação. Ao aumentar a relevância com o uso da semântica, é esperado que os mecanismos de busca apresentem informações mais relevantes.

Técnicas de anotação semântica são empregadas para descrever mídias como músicas e vídeos. A escolha da técnica de anotação varia de acordo com o formato alvo, podendo apresentar diferentes abordagens, as quais são discutidas a seguir.

Em músicas, por exemplo, Turnbull *et al* [17] menciona três abordagens. A primeira, considerada como técnica clássica, é a utilização dos metadados armazenados neste tipo de arquivo, como nome do compositor ou artista, nome da música e a data de lançamento do álbum. Estas informações são utilizadas para gerar as *tags* que identificarão os arquivos de áudio. Ainda neste tipo de mídia, uma segunda abordagem apontada pelos autores é a denominada *query-by-similarity*[1]. Nessa técnica as músicas são catalogadas por humanos, tendo pequenos trechos dos arquivos associados a um índice de músicas. A identificação neste caso é realizada computando a similaridade dos trechos de áudio armazenados ao(s) arquivo(s) analisado(s). Por fim, é apresentada uma terceira técnica, considerada pelos autores como a mais genérica e que trata melhor o problema, nomeada como *query-by-text* [1]. Nesta técnica, um modelo é treinado para mapear os eventos acústicos em tags. Em Moyal *et al* [1], este mesmo método é descrito como *Acoustic Keyword Spotting*, onde são empregados modelos acústicos associados a um banco de dados de palavras-chave.

Para os arquivos de vídeo, Waitelonis *et al* [20] afirmam que, geralmente, os metadados não são suficientes para descrever o conteúdo deste tipo de mídia, o que elimina o emprego da técnica clássica dos arquivos de música nesse formato.

O autor aponta como solução relevante a abordagem de marcações e avaliações realizadas por humanos. Tal avaliação é conhecida como Folksonomia [9,22]. Em mecanismos de busca, esta abordagem é empregada pelo Google [36], com o Youtube, e também pelo mecanismo de busca de vídeos acadêmicos YoVisto [12]. Para a indústria, contudo, seus acervos de áudio/vídeos geralmente possuem poucas ou nenhuma *tag* criada manualmente, o que impossibilita o uso desta última abordagem. Neste cenário, tem sido criadas soluções para definição automática de *tags* para catalogação da informação [11].

## 2.3 Busca e apresentação da informação

Em sistemas de recuperação de informação, um usuário precisa traduzir a sua necessidade de informação em uma consulta que, frequentemente, não é clara o suficiente, uma vez que o mesmo não domina completamente o assunto pelo qual está buscando. Neste contexto, determinar um conjunto de documentos relevantes que contém palavras-chave que coincidam com as existentes na consulta geralmente não é suficiente para entregá-lo a informação desejada [16]. Algumas consultas são difíceis de serem tratadas pela maioria dos sistemas de recuperação de informação que calculam relevância por palavra-chave, como "presidente argélia durante 70" ou "quanto a apple gastou em propaganda em 2014?". Na primeira consulta, geralmente os sistemas não interpretam o *token* "70" como data, mas como texto. Na segunda consulta, a relação entre as palavras "quanto", "gastou" e "em 2014" não são determinantes para o resultado gerado e é comum sistemas retornarem sites de vendas para esta consulta.

Com o advento da Web semântica, torna-se viável o aperfeiçoamento dos mecanismos de busca, habilitando humanos e máquinas a descobrir informações utilizando fontes de dados heterogêneas [14]. Neste contexto, abordagens mais eficazes para *Question Answering* (QA) têm sido propostas [8,20]. Por QA, entende-se abordagens que realizam o processamento de consultas em linguagem natural e retornam uma resposta rápida e sucinta, com o contexto suficiente para validar a resposta.

Saedeeh *et al* [14] apresentam uma abordagem de QA que explora fontes de dados heterogêneas descritas em RDF para recuperar informações. Este tratamento transforma frases na forma de perguntas em linguagem natural em consultas que exploram *datasets* semânticos, permitindo ajudar usuários a recuperar a informação desejada de forma intuitiva, sem prévio conhecimento das *tags* de classificação para o item buscado.

Do ponto de vista da interface, a disponibilização de filtros sofisticados ou ainda formas especializadas de extração que se apoiam na semântica podem ser utilizadas. Uma das técnicas que facilitam a navegabilidade no resultado da busca é o *faceted browsing*. O *faceted browsing* consiste em um agrupamento dinâmico de itens ou resultados de uma pesquisa em categorias. As categorias representam um determinado domínio e observam seus relacionamentos através dos metadados associados, podendo ser exploradas de acordo com as preferências do usuário. Segundo Uddin [18], esta abordagem supera as limitações da classificação hierárquica, onde a análise da informação acontece de forma bidirecional (*bottom-up*), ao permitir uma análise multidimensionada.

# 3. TRABALHOS RELACIONADOS

Esta seção apresenta os trabalhos relacionados, envolvendo técnicas de indexação e recuperação dos dados para artefatos multimídia (áudio/vídeo) aplicáveis a grandes coleções de vídeos.

O principal mecanismo de busca relacionado à este trabalho é o YoVisto [12]. O YoVisto é um buscador especializado em vídeos acadêmicos e conferências. Sua principal contribuição é utilizar a indexação do conteúdo com baixa granularidade, segmentando e definindo *tags* no vídeo por quadro ou trechos do vídeo. Para extrair a informação, são utilizadas tecnologias de processamento de imagens, *Optical Character Recognition* e Folksonomia. Os metadados extraídos compõem o índice utilizado pela ferramenta. Finalmente, estes metadados são então relacionados com entidades da DBpedia para ajudar o motor de busca a sugerir conceitos e apoiar o usuário durante a busca. Combinando estas técnicas, o YoVisto faz uso das vantagens do conceito de buscas exploratórias apoiados por filtros geográficos e *faceted browsing* utilizando propriedades da DBpedia.

A proposta do Qodra se assemelha com a do YoVisto, ao explorar vídeos acadêmicos. A contribuição do *framework*, neste ponto, é sua arquitetura modular, podendo ser utilizado para indexar qualquer tipo de áudio/vídeo com diferentes técnicas. Na indexação do conteúdo, o YoVisto explora processamento de imagens, OCR e Folksonomia para determinar *tags*. No *framework* proposto, é utilizado ASR e um módulo de anotação semântica para a marcação automática das *tags*. Estas *tags* podem ainda ser avaliadas por humanos através do módulo de *feedback*. O principal avanço da arquitetura do *framework* é poder combinar técnicas e não somente explorar um par de estratégias de indexação. O YoVisto ao não publicar o seu índice ou disponibilizar os dados em outros formatos, além do HTML, perde no aspecto de integração e não se beneficia por completo dos recursos da Web Semântica, além de estar altamente acoplado à sua interface de recuperação, não abrindo chances de explorar novas abordagens.

No processo de indexação, o trabalho mais importante para este projeto é o desenvolvido na BBC [11]. Neste trabalho, os autores utilizaram uma ferramenta de ASR chamada de CMU-Sphinx [8], juntamente com os modelos acústicos HUB4 e o modelo de linguagem Gigaword [5] para realizar as transcrições automáticas do acervo da rádio BBC. O objetivo principal do trabalho é desenvolver uma técnica capaz de classificar de forma automatizada o conteúdo abordado nas entrevistas, de forma a apoiar arquivistas no processo de catalogação do arquivo. A catalogação é realizada utilizando *tags* da DBpedia com o objetivo de descrever de forma não ambígua os arquivos.

Na modelagem computacional do problema, Yves *et al* [11] empregaram como modelo um espaço vetorial, conhecido como *Enhanced Topic-based Vector Space Model* (eTVSM). Segundo Polyvyanyy [9], o eTVSM é uma abordagem utilizada pela área de Recuperação de Informação, que, somada a filtros (*stop words removal* e técnicas de morfologia linguística, como *stemming*, fornece um modelo representativo da linguagem. Esse modelo, combinado a ontologias, consegue codificar e representar o fenômeno linguístico.

O *framework* proposto tem como objetivo descrever grandes coleções de vídeos utilizando *tags*, assim como a BBC realizou na catalogação do arquivo. No cenário de grandes coleções, não é viável a intervenção humana no processo de transcrição ou geração de *tags.* Apesar do *framework* utilizar uma abordagem similar à da BBC, combinando transcrição e classificação, é necessário ir além do modelo adotado na BBC, uma vez que o custo da busca é muito alto. Para realizar a busca, o *framework* realiza a divisão do arquivo em pequenas partes para classificar trechos do vídeo, o que se torna mais relevante do que classificar o arquivo por inteiro, como na BBC.

# 4. ARQUITETURA

Projetado para ser uma arquitetura modular e aberta, o *framework* proposto tem como o objetivo permitir que desenvolvedores possam experimentar a aplicabilidade de novas abordagens e ainda favorecer a combinação das melhores técnicas de recuperação da informação. Ao construir sua base em conformidade com os padrões RDF e OWL, a arquitetura foi projetada para maximizar a precisão dos métodos de recuperação, fornecendo melhores experiências ao usuário, ampliando de forma significativa a capacidade de explorar formatos semânticos de dados.

Por possuir uma arquitetura modular, composta por 7 módulos: *crawler, automatic search recognition,* anotação semântica, *question aswering,* persistência, interface web e *feedback,* os quais são descritos nas seções seguintes. Assim, diferentes abordagens e ferramentas podem ser plugadas no sistema. A integração externa é realizada inicialmente com as bases de dados públicas DBpedia e Freebase e todo o projeto do *framework* considera os objetivos mencionados na introdução do artigo.

## 4.1 Extração, representação e persistência da informação

O módulo *crawler* é responsável por recuperar vídeos de fontes determinadas, extrair as metainformações dos arquivos e criar os registros baseados no modelo definido pelo programador. São utilizadas ontologias no formato OWL na definição de metadados. Além disso, o módulo é responsável por notificar o módulo de ASR sobre a ocorrência de novos arquivos, os quais serão processados para geração dos seus transcritos.

Ao ser notificado sobre novas entradas, o módulo de ASR realiza as conversões necessárias para processar a transcrição, como a conversão do arquivo no formato de entrada esperado pela ferramenta ASR configurada, segmentação do vídeo e aplicação de um conjunto de filtros que podem ser definidos pelo programador. A segmentação tem como principal objetivo indexar o arquivo de forma granular. É ainda responsabilidade do módulo de ASR direcionar o processamento da transcrição para a instância do idioma a ser transcrito.

No armazenamento dos metadados, o módulo de persistência permite o uso de qualquer gerenciador de banco de dados que possua uma interface de consulta SPARQL e de inserção SPARQL *Update*. Os dados são armazenados em formato RDF, o qual facilita o uso de abordagens atuais de QA, conforme será discutido na seção 4.4. Inicialmente a ferramenta utiliza o banco de dados RDF AllegroGraph.

## 4.2 Transcrição automática

O módulo de ASR executa a transcrição automática, podendo ser definido um pré- e pós-processamento. A fase de pré-processamento pode aplicar filtros e segmentar o vídeo por tempo ou utilizando abordagens de *speaker diarisation* para identificar trechos de áudios de mesmo falante. A fase de pós-processamento permite aplicar tarefas sobre os transcritos. Por exemplo, filtros podem ser aplicados nos trechos transcritos, ou as transcrições de certos segmentos podem ser agrupadas conforme a necessidade do desenvolvedor. Para cada tarefa das do ASR, é possível configurar o uso de ferramentas externas.

Inicialmente, o módulo de ASR realiza transcrições em idioma português utilizando a ferramenta Coruja JLapsAPI. Na fase de pré-processamento, o áudio é extraído do vídeo e segmentado em trechos de menos de dois minutos. É ainda utilizada uma abordagem de *speaker diarisation* que garante o corte no áudio em períodos de pausa da fala.

O Coruja apresenta resultados satisfatórios para a transcrição. Em experimentos conduzidos para a validação dos dicionários, foram utilizados áudios de diferentes fontes em situações de diálogos livres, ou seja, em condições diversas de ruído. O nível de ruído de cada áudio foi classificado da seguinte forma: (1) pouco ruído são áudios/vídeos gravados em ambientes com sons que variam de 0 a 30 dB, como entrevistas em ambientes fechados que possuem algum isolamento acústico. (2) médio ruído são áudios/vídeos gravados em ambientes com sons que variam de 30 a 60 dB, como entrevistas que ocorrem em uma biblioteca ou sala com a janela aberta. (3) muito ruído são áudios/vídeos gravados em ambientes com sons acima de 60 dB, como os encontrados em ambientes de escritório e ambientes externos. Os resultados foram comparados às transcrições realizadas por um humano e analisadas computando a similaridade entre os textos. A tabela 2 apresenta os resultados gerados. A similaridade corresponde ao valor-F (ou medida-F) calculada pela sobreposição dos *tokens* da transcrição automática e da transcrição manual, excluindo-se *stopwords* e normalizando pelo tamanho da transcrição.

Tabela 2. Comparação de similaridade dos transcritos realizados pelo ASR Coruja JLapsAPI 1.5 em relação à transcrição realizada por um humano

| Nível de ruído | Similaridade (%) |
|---|---|
| Pouco ruído | **65,9869** |
| Médio | **54,6981** |
| Muito ruído | **47,3998** |

## 4.3 Anotação semântica dos vídeos

O módulo de anotação semântica tem como objetivo associar automaticamente *tags* para os vídeos. O módulo recebe como entrada as transcrições geradas pelo ASR. Contudo, transcrições geradas de forma automática possuem muitos *tokens* falso-positivos, conforme pode-se inferir da tabela 1. Assim, o módulo de anotação semântica necessita avaliar a transcrição, identificar o contexto do áudio e escolher qual o conjunto de *tags* que melhor descrevem o vídeo.

O módulo permite que o desenvolvedor acrescente diferentes estratégias na escolha de *tags*. Porém, quatro estratégias estão disponíveis para serem utilizadas: Spotlight Lucene [7], Spotlight JDBM [3], eTVSM [9] e matrizes de coocorrência. É permitido que o desenvolvedor defina o uso de mais de uma estratégia ao mesmo tempo. Neste caso, o módulo é responsável por combinar os resultados das diferentes estratégias.

O Spotlight é uma API *open source* para anotações automáticas que fornece uma interface programática com o objetivo de reconhecer e desambiguar (*entity linking*) textos não estruturados. A API fornece *tags*, *scores* e tipos da ontologia da DBpedia. O módulo de anotação semântica permite a utilização das duas implementações da API: o Spotlight Lucene e o Spotlight JDBM. O primeiro é uma implementação modificada do *TF-IDF weights*, enquanto o segundo é uma implementação do *generative probabilistic model*.

A terceira opção trata-se de um modelo vetorial, conhecido como *Enhanced Topic-based Vector Space Model* (eTVSM). Esta técnica é utilizada em Recuperação da Informação para fornecer um modelo representativo de linguagem, tendo como vantagem sua flexibilidade de configuração. Estruturalmente, o eTVSM é composto por um espaço vetorial com apenas eixos positivos. Cada dimensão é um domínio ou conceito representado por um vetor. Sua utilização se dá pela extração de conceitos dos documentos alvo da análise. Cada termo tem então sua similaridade computada utilizando o eTVSM, onde à partir dos valores, é possível inferir qual é o tópico relacionado ao documento. Esta solução categoriza e realiza a anotação automática.

A última estratégia disponível é a matriz de coocorrência. Este método é usado para identificar a frequência da ocorrência conjunta de dois termos em transcritos. Com esta correlação, é possível reconhecer o domínio do transcrito através da identificação de *clusters* de termos com maior coocorrência.

O módulo de *feedback* fornece uma interface para avaliar as estratégias de anotação semântica. O usuário pode validar as *tags* geradas e adicionar novas *tags* nos vídeos já indexados. Para avaliar as *tags* geradas, o usuário classifica a *tag* em uma das quatro opções: (1) central, (2) relevante, (3) neutro e (4) não relevante.

Caso o módulo seja configurado para utilizar mais de uma técnica ao mesmo tempo, os resultados serão combinados para gerar a lista final de *tags*. O módulo recebe então uma lista de entidades e *scores* fornecido por cada estratégia e aplica uma média ponderada para gerar o *score* final de cada *tag*. Os pesos aplicados em cada estratégia podem ser pré-definidos ou calculados de acordo com o histórico de avaliações manuais

realizadas por especialistas no módulo de feedback. Assim, permite-se reconhecer e priorizar as *tags* geradas pelas técnicas de maior acurácia.

## 4.4 Busca e apresentação dos resultados

Ao ser utilizado um banco de dados com interface SPARQL, é facilitado o uso de abordagens de busca semânticas ou tradicionais, cobrindo assim uma variedade de estratégias de busca que podem ser utilizadas. O módulo de busca permite que métodos de busca possam ser inseridos na aplicação. O módulo recebe a consulta do usuário e possui acesso ao módulo de persistência. A lista de resposta gerada é então processada pela interface web.

Dois métodos estão previamente disponíveis para o implementador, sendo uma busca por palavra-chave e uma classe abstrata nas ferramentas de QA. A busca por palavra-chave permite que palavras contidas nos metadados dos vídeos possam ser encontradas. É utilizada uma linguagem de consulta semelhante à utilizada em sistemas como o Apache Lucene, onde consultas como "física" ou "geometria analítica" correspondem a uma busca pela ocorrência destas palavras em todos os metadados dos vídeos, enquanto consultas como "title:árvore b" corresponde a uma busca por vídeos que contenham a propriedade *title* com o valor "árvore b". Por fim, a busca por métodos de QA permite que sejam configuradas máquinas de busca que recebem uma consulta na forma de um questionamento, como "Quais estruturas de dados são importantes para Recuperação de Informação?".

A interface web é responsável por exibir ao usuário a resposta gerada pelo módulo de busca, processar os metadados dos vídeos retornados e as facetas que serão utilizadas na busca. É possível que o desenvolvedor configure a disposição das áreas de informação, altere os rótulos, defina temas e permita a publicação do *endpoint* SPARQL na publicação dos dados armazenados na aplicação formato de dados ligados. Com o foco na integração, é fornecida uma implementação RESTful. Com o fraco acoplamento provido pelo REST, o desenvolvedor pode criar aplicativos no entorno dos serviços expostos pela interface, ganhando assim flexibilidade para gerar aplicações em diversas plataformas, como dispositivos móveis e desktop.

O mecanismo de busca inicialmente configurado é o openQA (aksw.org/Projects/openQA.html). Este mecanismo tem como essência a combinação de diferentes bases de dados em diversos domínios para analisar e recuperar a informação. A arquitetura do openQA também é expansível, pretendendo ser uma plataforma comum visando facilitar a integração de diferentes abordagens. A interface web é configurada para criar facetas relacionadas com os tipos da ontologia da DBpedia, auxiliando o usuário na seleção dos vídeos durante a busca. As informações correlatas aos vídeos são complementadas utilizando as bases de dados externas DBpedia e Freebase.

# 5. CONCLUSÕES

Sistemas tradicionais de busca de vídeos se utilizam de informação textual com o intuito de processar a consulta do usuário, sendo necessário associar texto ao vídeo, como metadados (título e autor), informação de contexto (texto no entorno do vídeo dentro de uma página web). Empresas de comunicação, como a BBC, utilizam a abordagem por *tags* editoriais, ou seja, um editor ou jornalista adiciona *tags* para descrever os vídeos. Por ser um trabalho manual, torna-se praticamente inviável a anotação de uma grande base de programas antigos de emissoras [11].

Neste trabalho, tratamos do problema de indexação de áudio e vídeos através da transcrição automática dos áudios. Os desafios de indexar este tipo de mídia são consideráveis, ao avaliar as restrições relatadas nas tecnologias envolvidas. Porém, diferentes abordagens têm sido propostas na literatura.

A principal contribuição deste trabalho é apresentar uma arquitetura modular, *open source*, que pode ser explorada e configurada para fazer uso de diferentes soluções. Esta arquitetura se encontra em uso e foi configurada para recuperar vídeo-aulas produzidas pelo Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora.

Como trabalhos futuros, planeja-se a adição de módulos que combinem OCR e *screen scraping* para extrair e processar as informações visuais nos vídeos, como no YoVisto. Ainda, é planejado que a arquitetura utilize diferentes instâncias de ASR visando permitir a transcrição de áudios em idiomas distintos.

# REFERÊNCIAS


[1] E. T. Ami Moyal, Vered Aharonson and M. Gishri.Phonetic search methods for large speech databases. In Springer, editor, Phonetic Search Methods for LargeSpeech Databases. Springer, London, UK, 2013.

[2] B. Croft, D. Metzler, and T. Strohman. Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company, USA, 1st edition, 2009.

[3] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In Proceedings of the 9th International Conference on Semantic Systems (I-Semantics), 2013.

[4] J. Forney, G.D. The viterbi algorithm. Proceedings of the IEEE, 61(3):268–278, 1973.

[5] D. Graff, J. Kong, K. Chen and K. Maeda. English gigaword. Linguistic Data Consortium, Philadelphia, 2003.

[6] C. Körner, D. Benz, A. Hotho, M.trohmaier, and G. Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In Proceedings of the 19th international conference on World wide web, WWW '10, pages 521–530, New York, NY, USA, 2010. ACM.

[7] P. N. Mendes, M. Jakob, A. Garcia-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics), 2011.

[8] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, et al. The 1996 hub-4 sphinx-3 system. In Proc. DARPA Speech recognition workshop, pages 85–89. Citeseer, 1997.

[9] A. Polyvyanyy. Evaluation of a novel information retrieval model: etvsm, 2007.

[10] L. R. Rabiner and B.-H. Juang. Fundamentals of speech recognition, volume 14. PTR Prentice Hall Englewood Cliffs, 1993.

[11] Y. Raimond and C. Lowis. Automated interlinking of speech radio archives. In C. Bizer, T. Heath,
T. Berners-Lee, and M. Hausenblas, editors, LDOW, volume 937 of CEUR Workshop Proceedings, London, UK, 2012. CEUR-WS.org.

[12] H. Sack and J. Waitelonis. Exploratory semantic video search with yovisto. In Semantic Computing (ICSC),
2010 IEEE Fourth International Conference on, pages
446–447, Sept 2010.

[13] K. Seymore, S. Chen, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, et al. The 1997 cmu sphinx-3 english broadcast news transcription system. In Proceedings of the 1998 DARPA Speech recognition Workshop, pages 55–59, 1998.

[14] S. Shekarpour, A.-C. N. Ngomo, and S. Auer.Question answering on interlinked data. In WWW, pages 1145–1156, 2013.

[15] L. Specia and E. Motta. Integrating folksonomies with the semantic web. In E. Franconi, M. Kifer, and W. May, editors, The Semantic Web: Research and Applications, volume 4519 of Lecture Notes in Computer Science, pages 624–639. Springer Berlin Heidelberg, 2007.

[16] G. Stamou and S. Kollias, editors. Multimedia Content and the Semantic Web: Standards, Methods and Tools. Wiley, Chichester, UK, 2005.

[17] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. IEEE Transactions on Audio, Speech, and Language Processing, 2008.

[18] M. N. Uddin and P. Janecek. Faceted classification in web information architecture: A framework for using semantic web tools. The Electronic Library, 25(2):219–233, 2007.

[19] J. Waitelonis, N. Ludwig, and H. Sack. Use what you have: Yovisto video search engine takes a semantic turn. In T. Declerck, M. Granitzer, M. Grzegorzek, M. Romanelli, S. Ru¨ger, and M. Sintek, editors,
Semantic Multimedia, volume 6725 of Lecture Notes in Computer Science, pages 173–185. Springer Berlin Heidelberg, 2011.

[20] J. Waitelonis, H. Sack, J. Hercher, and Z. Kramer. Semantically enabled exploratory video search. In Proceedings of the 3rd International Semantic Search Workshop, SEMSEARCH '10, pages 8:1–8:8, New York, NY, USA, 2010. ACM

[21] B. Popov, A. Kiryakov, D. Manov, A. Kirilov, and O. M. Goranov. Towards semantic web information extraction. In In proceedings of ISWC, 2003