Enriquecimento semântico de repositórios de videoaulas: um estudo de caso

João P. R. P. da Silva[2], João Victor de Souza[2], Jairo Francisco de Souza[1], Eduardo Barrére[1]

[1]Departamento de Ciencia da Computação – (UFJF)
36.360-900 – Juiz de Fora – MG – Brasil

[2]Bacharelado em Ciência da Computação – Universidade Federal de Juiz de Fora (UFJF)
36.360-900 – Juiz de Fora – MG – Brasil

{jpradd, joao.souza, jairo.souza, eduardo.barrere}@ice.ufjf.br

Abstract. *An important feature in most educational video repositories is the lack of relevant information in keywords and metadata that enables full access to videos through search engines. One of these repositories is Videoaula@RNP, for which a project was developed with a focus on the semantic enrichment of the educational videos stored there, based on audio transcription, semantic annotation and similarity analysis. This article presents this solution and the impacts generated from its application in the repository.*

Resumo. *Uma característica importante na maioria dos repositórios de vídeos educacionais é a carência de informações relevantes, nas palavras-chave e nos metadados, que possibilitem um pleno acesso aos vídeos através de mecanismos de busca. Um desses repositórios é o Videoaula@RNP, para o qual foi desenvolvido um projeto com foco no enriquecimento semântico dos vídeos educacionais lá armazenados, com base na transcrição de áudio, anotação semântica e análise de similaridade. Este artigo apresenta esta solução e os impactos gerados a partir da sua aplicação no repositório.*

1. Introdução

Dentre a grande quantidade e variedade de materiais didáticos produzidos atualmente, a mídia que tem grande destaque é o vídeo. Além das características pedagógicas do vídeo, os avanços da internet banda larga e de tecnologias de hardware e software têm facilitado o acesso e consumo do vídeo [de Oliveira et al. 2010a]. Neste cenário, se fez necessário o armazenamento dessas mídias, gerando assim um aumento significativo no volume desses repositórios [de Oliveira et al. 2010b]. Na área de educação, o crescente aumento desses repositórios, no entanto, é acompanhado pela dificuldade de encontrar esses conteúdos, visto que os dados de auto-descrição que acompanham as videoaulas são geralmente escassos e genéricos [Yang and Meinel 2014]. Os mecanismos de busca ainda são muito dependentes de informações textuais que descrevem a mídia [Coelho and de Souza 2014]. Esses dados são geralmente preenchidos pelos próprios criadores dos conteúdos, que usualmente não se preocupam com os metadados durante o envio do vídeo para o repositório. Logo, diversos vídeos se tornam praticamente inacessíveis. Esse problema é ainda mais evidente quando um aluno deseja encontrar um trecho específico de um vídeo, um exemplo ou explicação de aula. Recuperar essa informação de forma manual é possível, mas é um trabalho muito custoso e exaustivo [Dias et al. 2017].

DOI: 10.5753/cbie.sbie.2018.1263

Técnicas automáticas de anotação ou extração de texto, como o OCR (*Optical Character Recognition*), podem ser usadas para enriquecer metadados de vídeos educacionais [Gomes Jr et al. 2017]. De forma similar, a anotação semântica sobre textos transcritos pode ser usada para extrair informações de vídeos e é bem menos custosa que métodos manuais para extrair conteúdo [Lawrence 2008, Turnbull et al. 2008]. A existência de uma variedade de técnicas e recursos, tanto de transcrição quanto de anotação, permite um número grande de combinações, adaptações e treinamentos possíveis para diversos conteúdos e mídias. Isso abre possibilidades de soluções eficientes para recuperação de informação em repositórios, como os de vídeos educacionais da Rede Nacional de Ensino e Pesquisa, a RNP (Videoaula@RNP[1]). O uso dessas técnicas, contudo, não é trivial e a escolha do conjunto de técnicas adequadas para enriquecer metadados de vídeos pode variar para cada tipo de conteúdo.

O presente artigo apresenta um estudo de caso de uso de técnicas de enriquecimento semântico em repositórios de videoaulas em língua portuguesa. O objetivo de tal estudo é mostrar uma combinação de soluções para o tratamento deste problema. Como cenário, foi escolhido o repositório Videoaula@RNP, o qual possui material didático produzido por diferentes instituições de ensino superior do Brasil e que oferece vídeos educacionais sobre diversas áreas de conhecimento. Este trabalho apresenta os resultados dos experimentos, analisando a melhoria no conjunto de dados dos vídeos através da associação automática de identificadores semânticos para cada mídia. Ainda, é demonstrado como esses identificadores semânticos podem auxiliar no processo de identificar mídias com assuntos correlatos, o que pode auxiliar no processo de recomendação de conteúdo ou expansão de consultas.

2. Trabalhos Relacionados

Alguns pesquisadores desenvolveram métodos e técnicas para recuperação de informação de vídeos para geração de palavras-chave, metadados e identificação de assuntos correlatos. O uso de técnicas de reconhecimento automático de fala em conjunto com técnicas de anotação semântica foi explorado em [Yang and Meinel 2014, Grünewald and Meinel 2015, Zhao et al. 2015, Gomes Jr et al. 2017]. A diferença entre os trabalhos apresentados variam entre as técnicas de anotação, transcrição e o tipo de vídeo processado. Para o processo de recuperação de informação para anotação de mídias, destaca-se o uso de modelos de espaços vetoriais de [Raimond and Lowis 2012], o qual aplicou a técnica em programas de rádio, e o sistema de segmentação de vídeos para serem transcritos e anotados proposto por [Biswas et al. 2015]. Assim como em [Raimond and Lowis 2012], este estudo de caso aplicou um modelo de extração de tópicos. Contudo, este foi aplicado a vídeos educacionais e adaptado para associar identificadores de conceitos contidos nas principais áreas de conhecimento. Este trabalho também faz uso de segmentação automática para auxiliar no processo de anotação, assim como em [Biswas et al. 2015]. Contudo, os autores utilizam o método em conjunto com técnicas de OCR para auxiliar na identificação das palavras mais importantes de cada segmento, enquanto este trabalho faz uso do modelo de extração de tópicos para identificação dessas palavras baseado na sua frequência e relação com demais palavras.

Técnicas para auxiliar o processo de recomendação de conteúdos digi-

[1] http://www.videoaula.rnp.br/

tais estão presentes na literatura, como apresentadas em [Casagrande et al. 2015, Neves et al. 2016]. Em [Neves et al. 2016], os autores identificaram relacionamentos entre objetos de aprendizado utilizando o padrão SCORM (*Sharable Content Object Reference Model*). Embora este padrão seja propício para esse tipo de aplicação, muitos repositórios abertos não usam os padrões SCORM e LOM (*Learning Object Metadata*), como é o caso do repositório da RNP. Outros trabalhos fazem uso de dados ligados ou bases de conhecimento para calcular a similaridade entre recursos [Zhu and Iglesias 2017, Herrera et al. 2016, Cheniki et al. 2016]. O presente trabalho faz uso do grafo de conceitos da DBpedia para identificar vídeos com conteúdos correlatos.

3. Processo de enriquecimento semântico

Embora a produção de vídeos tenha sido facilitada pela possibilidade de aquisição de equipamentos mais baratos e de boas ferramentas para edição, percebe-se que a grande maioria das videoaulas disponibilizadas pelas instituições no Videoaula@RNP possuem um formato tradicional, onde o professor discursa sobre um assunto, sem a presença de muitos recursos audiovisuais que complementem o cenário. Técnicas de segmentação através de eventos e tomadas do vídeo foram utilizadas para facilitar a extração da informação relevante. No entanto, em sua grande maioria, as videoaulas possuem um conteúdo informativo uniforme, dificultando a identificação de eventos e a detecção de limite de tomadas. Esses tipos de vídeos não podem ser facilmente segmentados de acordo com um evento específico, cada parte do vídeo se mostra igualmente importante para o usuário [Taskiran et al. 2006]. Neste contexto, abordagens que fazem uso de reconhecimento de fala, como a apresentada em [Gomes Jr et al. 2017], podem gerar informação útil para a anotação. Em [Gravier et al. 2015] os autores afirmam que, na maioria dos casos, a fala, a linguagem e o áudio são importantes portadores de semântica nos conteúdos multimídia. Em particular, a linguagem é de extrema importância para a compreensão da mensagem.

Por esta razão, a abordagem adotada para este estudo de caso se baseou em 4 etapas (Figura 1). Na primeira etapa, os vídeos são coletados do Videoaula@RNP através de um *crawler* que recupera todos os vídeos e suas informações (título, descrição e palavras-chave). Em seguida, técnicas de reconhecimento automático de fala são utilizadas para gerar uma transcrição automática do áudio desse vídeo. O transcrito, então, é utilizado para associação de identificadores semânticos ao vídeo. Por fim, os vídeos são analisados com toda a base processada para que sejam identificados vídeos com assuntos correlatos.

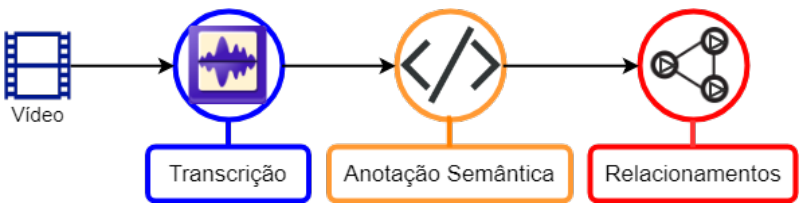

Figura 1. Etapas do processo de enriquecimento semântico

3.1. Etapa de reconhecimento automática de fala

Como a maior parte da informação contida em vídeos educacionais se encontram na fala do orador, técnicas de reconhecimento automático de fala podem ser utilizadas para extrair essa informação, gerando um texto automaticamente transcrito. Esse processo, contudo, é fortemente influenciado por algumas variáveis, como a presença de mais de um

falante, presença de música de fundo, presença de palavras ou fonemas desconhecidos ao longo do áudio, etc. Para minimizar esses problemas, é necessário um treinamento desses algoritmos com dados condizentes com o cenário de aplicação. Assim, embora existam diversos serviços na Web que podem ser utilizados para realizar transcrição automática de áudios, estes são treinados de forma generalista, o que pode inviabilizar seu uso.

Para treinar o transcritor automático, é necessário gerar dois modelos: (1) o modelo acústico, responsável por decodificar os fonemas do áudio, e (2) o modelo de língua, responsável por representar a distribuição de probabilidade das sequências de palavras do cenário de discurso no qual será aplicado. Para treinar acústicos é necessário utilizar um *corpus* com áudios com trechos de fala e suas respectivas transcrições. Assim, a maior dificuldade no processo de treinamento é ter acesso a uma base de dados suficientemente completa de treinamento para a aplicação-alvo. Neste trabalho, para o treinamento do modelo acústico foram reunidas videoaulas de diversas áreas, extraídas do repositório do Coursera e das bases disponibilizadas gratuitamente pelo projeto VoxForge e pelo Laboratório de Processamento de Sinais da UFPA. Os arquivos somam aproximadamente 58 horas de áudio, distribuídos segundo a Tabela 1. O treinamento foi realizado com o *toolkit* Kaldi[2], utilizando redes neurais profundas.

Tabela 1. Bases de áudio utilizadas no treinamento do transcritor automático

Base	Quantidade (arquivos)	Tempo (horas)
Coursera	582	55
VoxForge	700	0,76
LaPS	700	0,9

Para que houvesse uma maior quantidade de dados de treinamento para o modelo acústico, foram utilizadas técnicas de *data augmentation* para sistemas de reconhecimento automático de fala [Ko et al. 2015]. Técnicas como essa, permitem criar um modelo acústico mais robusto, sem o custo de coletar mais dados de treinamento, pois diversos novos áudios podem ser gerados. Essas técnicas baseiam-se na geração de arquivos de áudio utilizando os arquivos já existentes para inserir variações, como inserção de ruídos, gerando novos dados para o treinamento. No *corpus* de treinamento para o modelo acústico, três técnicas foram aplicadas. A primeira trata-se de alterar a velocidade do áudio, gerando novos áudios com 90% e 110% da velocidade original do áudio, triplicando os dados. Em seguida, foram gerados mais dados utilizando sinais de áudio com ruídos, duplicando seu tamanho; e, por fim, foi utilizada uma técnica para adicionar reverberações nos sinais existentes [Ko et al. 2017], também duplicando os resultados. Ao final do processo, ocorreu um aumento em 12 vezes do tamanho da base de dados disponível para o treinamento. Para o treinamento de modelo de língua foram utilizadas as transcrições dos corpus de fala da Tabela 1 e as bases CETEN[3], OGI[4] e LapsFolha[5], as quais somam mais de 30 milhões de palavras, como descriminado na Tabela 2.

[2]http://kaldi-asr.org/
[3]https://www.linguateca.pt/cetenfolha/index_info.html
[4]https://catalog.ldc.upenn.edu/LDC94S17
[5]http://labvis.ufpa.br/falabrasil/

Tabela 2. Treinamento de modelo de língua

Base	Coursera	VoxForge	LaPS	CETEN	LapsFolha	OGI
Quantidade de palavras	487894	11081	7093	26500035	2764232	33881

Para demonstrar os resultados do treinamento, foi comparada a taxa de erro por palavra (*Word Error Rate – WER*) da transcrição em relação aos serviços ASR de grandes empresas como Google e Microsoft, utilizando as bases de áudio de fala do LaPS, VoxForge e das videoaulas como teste. Como apresenta a Tabela 3, o modelo treinado conseguiu obter resultados equivalentes a esses serviços nos casos de teste. Vale ressaltar que o resultado pode ser melhorado através de novos treinamentos, com maior número de horas de áudio. Contudo, como será discutido nas próximas seções, esta é uma taxa suficiente para gerar resultados satisfatórios no processo final.

Tabela 3. Comparação de taxas de erro

Base de Dados	Word Error Rate (%)		
	Modelo treinado	Google	Microsoft
LaPS	7,93	11,1	16,0
VoxForge	14,31	13,6	16,0
Videoaulas	39,09	35,9	44,7

3.2. Etapa de anotação semântica

Após realizar a etapa de transcrição, o próximo passo realizado é a etapa de anotação semântica. Nessa etapa, são associadas *tags* ao vídeo, as quais representam o seu conteúdo. Para essa etapa, existem duas abordagens principais que geralmente são utilizadas: uma baseada na extração de tópicos e outra no reconhecimento de entidades. A abordagem baseada na extração de tópicos visa atribuir *tags* que melhor representam os assuntos que o texto aborda. No reconhecimento de entidades, os termos principais do texto são selecionados e a estes são atribuídas as *tags* que melhor os representam. Neste trabalho, chamamos as *tags* de anotações semânticas, uma vez que será atribuída a cada *tag* um identificador de um recurso da ontologia da DBpedia[6]. Conforme análise feita em [Gomes Jr et al. 2017], abordagens baseadas na extração de tópicos geram melhores resultados quando são utilizadas em transcritos automáticos, uma vez que esse tipo de texto está sujeito à presença de palavras que não foram efetivamente pronunciadas no texto (ruído). Esse fenômeno ocorre independente do transcritor utilizado, visto que o processo de transcrição automática faz uso de modelos (acústico e de língua) probabilísticos para predizer a palavra dita. Assim, por exemplo, palavras ausentes no treinamento, ao serem pronunciadas pelo falante, serão aproximadas para palavras conhecidas pelo transcritor.

Neste processo, foi utilizado para anotação o modelo de tópicos eTVSM (*enhanced topic-based vector space model*) [Raimond and Lowis 2012], que fornece um modelo representativo da linguagem utilizando um espaço vetorial multidimensional, onde cada dimensão representa uma categoria presente na ontologia da DBpedia. O modelo de tópicos foi especializado para associar apenas identificadores de recursos que representam

[6]http://dbpedia.org/

conceitos. Para isso, foi criado um método de filtragem nos dados da DBpedia, de forma a identificar recursos que são instâncias de conceitos e retirá-las do treinamento do modelo. Com isso, houve uma redução de 63848 entidades, aproximadamente 10% no volume de dados da DBpedia a serem processados e uma maior especialização do modelo.

3.3. Etapa de identificação de conteúdos correlatos

Após associar as *tags* semânticas que identificam o conteúdo do vídeo, o mesmo é comparado com o restante do repositório processado para identificar vídeos com conteúdos similares. Para cada vídeo v_i dentre os mais correlatos a *v*, é inserido no banco uma relação de *v* para v_i. Vale ressaltar que a relação não é simétrica. Estas relações permitem que, ao pesquisar por um dado conteúdo, o usuário possa receber como sugestão vídeos similares em conteúdo, mesmo que possuam *tags* distintas associadas a ele.

Para calcular as similaridades entre os vídeos, foi utilizada a abordagem descrita em [Dias et al. 2017], a qual faz uso do grafo de categorias da DBpedia. Como foram processados vídeos em língua portuguesa, foi utilizada a DBpediaPT[7] em conjunto com a base em inglês. Para cada recurso anotado, o recurso correspondente na língua inglesa é identificado, uma vez que o volume de dados na DBpedia em inglês é muito maior que o repositório em português [Dias et al. 2017], o que permite alcançar um resultado melhor.

Após obter os correspondentes de todas as anotações de um vídeo v, é realizada uma busca no grafo para recuperar as categorias c_i associadas a este recurso no grafo. Em seguida, é gerado o conjunto que representa a semântica do vídeo v, o qual contém todas as categorias mais amplas e mais específicas relacionadas a cada categoria c_i no grafo. Como exemplo, com o recurso dbpedia:Home, temos a categoria category:Home. Essa categoria possui ligação com categorias mais abrangentes, como category::uthenics, e mais específicas, como category:Domestic_life. Após essa busca, é calculada a similaridade $s(v, v_i)$ entre cada vídeo v_i em relação a v utilizando o Coeficiente de Sorensen-Dice para o par de conjuntos à partir de v e v_i. É atribuído como relacionado a v os k vídeos cujos conjuntos possuem os maiores valores de similaridade a v, tal que $sim(v, v_i) > \tau$. O valor τ representa um limiar inferior; τ e k são parametrizáveis.

4. Estudo de caso

O serviço Videoaula@RNP é um repositório de vídeos criado pela RNP para disponibilizar videoaulas criadas por instituições associadas. Nesse portal, as próprias instituições submetem diretamente os conteúdos que estarão disponíveis para o acesso público, sendo responsáveis pelo envio do vídeo e de meta-informações, como título, descrição e palavras-chave. Estas informações são utilizadas pelo buscador do repositório para responder as consultas feitas pelos usuários. Porém, nem sempre essas informações são preenchidas de maneira que facilite a identificação do conteúdo do vídeo ou sua busca.

Para entender o cenário, foram identificadas e analisados os metadados de mais de 800 videoaulas disponíveis no serviço. Dessa análise obteve-se os resultados disponíveis na Tabela 4, onde é possível perceber que a média de *tokens* (palavras-chaves) por vídeo é baixa, logo, para um vídeo ser encontrado pelo usuário, este tem poucas possibilidades de palavras-chave para encontrá-lo. Além disso, a maior parte desses vídeos possuem *tokens* "gerais", isto é, palavras que não servem para identificar o conteúdo daquele

[7]http://pt.dbpedia.org/

vídeo especificamente, como, por exemplo, os *tokens* "aula" e "videoaula". Sendo assim, mesmo tendo um número considerável de videoaulas cadastradas na plataforma, existe uma carência de informações que ajudem a identificar cada uma delas, o que dificulta o processo de busca. Neste cenário, um usuário dificilmente encontraria um conteúdo específico de uma disciplina, sem ter que acessar diversos vídeos e pesquisar manualmente pelo trecho que lhe interessa. Esta falta de informações não é uma falha do serviço em si, mas do processo manual de inserção de vídeos no repositório, no qual o usuário não se atenta para a necessidade de inserir *tokens* pertinentes para que seu vídeo seja encontrado.

Tabela 4. Valores obtidos após a análise das videoaulas no Videoaula@RNP

Informação	Valores coletados
Número de videoaulas	858
Total de *tokens*	2225
Total de *tokens* distintos	849
Média de *tokens* por vídeo	2.59 ± 1.34
Videoaulas com *tokens* "gerais"	604
Videoaulas que não possuíam *tokens*	2

Para analisar o resultado do processo de enriquecimento sobre o repositório, foi criado um *dataset* com 93 vídeos selecionados aleatoriamente (11% do repositório), totalizando 3604 minutos de vídeos (aproximadamente 60 horas). Esses vídeos abordam diversos tópicos em áreas de conhecimento distintas. Para o processamento desse *dataset*, o sistema foi configurado para atribuir no máximo cinco relacionamentos para cada vídeo. O processamento total do *dataset* foi paralelizado em 4 máquinas de transcrição e levou aproximadamente 26 horas, utilizando máquinas com processadores Intel Xeon, 32 GB RAM e sistema operacional Ubuntu 16.04. O tempo de processamento final de cada etapa está presente na Tabela 5. A transcrição foi responsável por mais da metade do tempo de processamento total do sistema. Além desse processo, a anotação semântica consumiu quase a totalidade do restante do tempo.

Após o processamento dos vídeos, foi feita uma avaliação manual do *dataset*, dividido igualmente entre 4 avaliadores da equipe. Cada avaliador foi instruído a assistir todo o vídeo e avaliar os recursos gerados para o vídeo; verificar quantos recursos estavam corretos, quantos já existiam como *tags* na base; e quantos novos termos relevantes foram encontrados. No total, foram atribuídas 5 *tags* por vídeo, onde em média 3,3 termos corretos foram obtidos por vídeo. De todas as *tags* associadas, apenas 20 dos 311 termos relevantes encontrados já estavam presentes no repositório da RNP, o que comprova o aumento de *tags* de busca para os vídeos da base.

Tabela 5. Distribuição do tempo do processamento, em minutos

Etapa	Tempo total	Tempo médio	Porcentagem
Transcrição	901	9,7	56,2%
Anotação Semântica	698	7,51	43,5%
Relacionamentos	4,65	0,05	0,3%

Vale ressaltar que o experimento mede o quanto o sistema consegue associar aos vídeos as *tags* esperadas pelos especialistas. Em muitos casos, contudo, não podemos

considerar como erradas as *tags* associadas automaticamente. Essa é uma característica do modelo de anotação adotado, o qual tende a associar *tags* mais gerais em detrimento de *tags* mais específicas. Por outro lado, o humano tende a escolher como relevante *tags* mais específicas. Por exemplo, um vídeo em que um especialista escolheu a *tag* "Algoritmo de Dijkstra" como relevante teve a *tag* "Teoria dos Grafos" associada a ele. De fato, o termo "Teoria dos Grafos" compartilha diversas categorias em comum com o termo "Algoritmo de Dijkstra" na ontologia da DBpedia. Para o algoritmo ser mais tolerante aos erros no processo de transcrição, é importante que não seja guiado pelas palavras contidas no transcrito individualmente. Por isso, o modelo de tópicos tende a ser mais adequado neste cenário, mesmo tendo uma maior probabilidade de associar *tags* mais gerais.

No processo de identificação de vídeos com conteúdos correlatos, o algoritmo foi parametrizado para registar os cinco vídeos mais similares a cada vídeo processado, ou seja, $k = 5$. O limiar inferior para relacionamento foi definido como 30%, ou seja, $\tau = 0,3$. Esses valores foram definidos empiricamente, após experimentos no laboratório. Como resultado, foram criados relacionamentos em 48 dos vídeos processados. Como os vídeos que compõem a base foram selecionados aleatoriamente, não houve uma similaridade de mais de 30% para nenhum desses vídeos. Contudo, foram identificados alguns grupos de vídeos relacionados após o experimento. Três subgrafos fortemente conectados e isolados foram identificados. No primeiro grupo, estão os vídeos relacionados à temática de Banco de Dados, contendo oito vídeos. No segundo grupo, tem-se vídeos sobre Filosofia e Ética com 14 vídeos sobre essa temática. No terceiro grupo, foram identificados oito vídeos sobre TICs. Ainda, foram identificados dois subgrafos com menos vídeos e fracamente conectados. No primeiro grupo, estão seis vídeos da área de Matemática Financeira e, no segundo grupo, estão três vídeos sobre Computação Gráfica. Vale ressaltar que esses vídeos não possuíam praticamente nenhum metadado sobre seu conteúdo que tenha sido informado pelo usuário ao inserí-los no repositório.

Com o estudo de caso, mostra-se que foi possível extrair informação desses vídeos e, a partir disso, compreender o conteúdo existente no repositório. Os métodos aplicados permitem que algoritmos de busca façam uso dos relacionamentos e *tags* automáticas para melhorar seus resultados e que sistemas de recomendação de conteúdo possam oferecer ao aluno os recursos didáticos adequados para o seu interesse.

5. Conclusão

Nesse trabalho, foi descrito um processo para enriquecimento semântico de vídeos baseado em transcrição, anotação semântica e geração de relacionamentos por similaridade de conteúdo. O estudo de caso foi realizado no repositório de videoaulas da RNP, o Videoaula@RNP. O estudo de caso mostrou que é possível enriquecer semanticamente um repositório de videoaulas, mesmo que esses vídeos não possuam informação adicional associada a ele pelo usuário. A falta de informações é um problema em grandes repositórios de vídeos, principalmente quando o usuário não se vê motivado a escolher corretamente as *tags* de descrição do vídeo. Neste trabalho, foi escolhido o cenário de videoaulas, pois são vídeos com características que podem ser exploradas pelas técnicas escolhidas. Conforme apresentando em [Gomes Jr et al. 2017], o fala do narrador é a principal fonte de informações na maior parte dos vídeos educacionais. Além disso, muitas videoaulas ainda são gravadas em um único plano e sem uma definição clara de tomadas, o que dificulta o uso de técnicas de segmentação de vídeos e, em alguns casos, de reconhecimento

ótico de caracteres. Assim, o uso de técnicas de reconhecimento automático da fala é adequada como fonte de informações para o processo de enriquecimento semântico. Por outro lado, o uso de transcritos automáticos são passíveis de erros podendo ter um erro em torno de 40% em sistemas de grandes empresas, gerando a necessidade da escolha correta de métodos para anotação semântica que sejam mais tolerantes à presença desses erros de transcrição. Por fim, a análise de similaridade entre vídeos é beneficiada pelo uso de grandes bases de conhecimento abertas e generalistas, como a DBpedia.

Dos resultados apresentados, destaca-se o desenvolvimento do modelo de transcrição, que mesmo tendo sido treinado com uma base de treinamento muito menor que a usada por grandes nomes do mercado, utilizou técnicas que permitiu resultados satisfatórios para a base de teste, com uma taxa de erros de palavras de 44,08%. O uso da abordagem de anotação semântica nestes transcritos gerou mais de 300 termos corretos para a base de avaliação construída com 11% do volume do repositório. Na etapa de relacionamento, percorrendo a antologia da DBpedia, foi possível um enriquecimento extra com geração de categorias e criação de relacionamento entre as mídias da base, que até então eram desconhecidas tais relações entre os conteúdos dos vídeos. Dos vídeos relacionados, foram encontrados 5 subgrafos de vídeos abordando áreas díspares como Banco de Dados, Filosofia e Ética, Matemática Financeira, TICs e Computação Gráfica.

Como trabalhos futuros, pretende-se avaliar o quanto o esforço necessário para especializar os métodos de reconhecimento automático de fala, anotação semântica e cálculo de similaridade pode influenciar em repositórios de domínio específico. Embora a especialização possa permitir uma melhor acurácia, o esforço necessário para treinamento dessas técnicas pode ser um fator decisor na escolha dos métodos aplicados.

Referências

Biswas, A., Gandhi, A., and Deshmukh, O. (2015). Mmtoc: A multimodal method for table of content creation in educational videos. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 621–630.

Casagrande, M. F. R., Kozima, G., and Willrich, R. (2015). A recommendation technique based on metadata for digital repositories oriented to learning. *Brazilian Journal of Computers in Education*, 23(02):70.

Cheniki, N., Belkhir, A., Sam, Y., and Messai, N. (2016). Lods: A linked open data based similarity measure. In *25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pages 229–234. IEEE.

Coelho, S. A. and de Souza, J. F. (2014). Anotação semântica de transcritos para indexação e busca de vídeos. In *Conferência Ibero Americana WWW/INTERNET*.

de Oliveira, F. K., Santana, J. R., and de Oliveira Pontes, M. G. (2010a). O vídeo como ferramenta educacional a partir de múltiplas plataformas. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 1.

de Oliveira, F. K., Santana, J. R., and de Oliveira Pontes, M. G. (2010b). O vídeo pela internet como ferramenta educacional. In *Anais do Workshop de Informática na Escola*, volume 1, pages 1389–1392.

Dias, L. L., Barbosa, J. S., Barrére, E., and de Souza, J. F. (2017). An approach to identify similarity among educational resources using external knowledge bases. *Brazilian Journal of Computers in Education*, 25(02):18.

Gomes Jr, J., Souza, J., and Barrére, E. (2017). Comparativo entre fontes de dados para anotação automática de videoaulas. In *Simpósio Brasileiro de Informática na Educação*, volume 28, page 1127.

Gravier, G., Jones, G. F., Larson, M., and Ordelman, R. (2015). Overview of the 2015 workshop on speech, language and audio in multimedia. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1347–1348.

Grünewald, F. and Meinel, C. (2015). Implementation and evaluation of digital e-lecture annotation in learning groups to foster active learning. *IEEE Transactions on Learning Technologies*, 8(3):286–298.

Herrera, J. E. T., Casanova, M. A., Nunes, B. P., Lopes, G. R., and Leme, L. (2016). Dbpedia profiler tool: profiling the connectivity of entity pairs in dbpedia. In *International Workshop on Intelligent Exploration of Semantic Data.*

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *Acoustics, Speech and Signal Processing, 2017 IEEE International Conference on*, pages 5220–5224.

Lawrence, R. (2008). *Fundamentals of speech recognition*. Pearson Education India.

Neves, D. E., Ishitani, L., and Brandão, W. C. (2016). Methodology for recommendation and aggregation of learning objects in scorm. *Brazilian Journal of Computers in Education*, 24(01):11.

Raimond, Y. and Lowis, C. (2012). Automated interlinking of speech radio archives. *LDOW*, 937.

Taskiran, C. M., Pizlo, Z., Amir, A., Ponceleon, D., and Delp, E. J. (2006). Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia*, 8(4):775–791.

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476.

Yang, H. and Meinel, C. (2014). Content based lecture video retrieval using speech and video text information. *IEEE Transactions on Learning Technologies*, 7(2):142–154.

Zhao, B., Xu, S., Lin, S., Luo, X., and Duan, L. (2015). A new visual navigation system for exploring biomedical open educational resource (oer) videos. *Journal of the American Medical Informatics Association*, 23(e1):e34–e41.

Zhu, G. and Iglesias, C. A. (2017). Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85.