Proposta para Grupo de Trabalho

GT-BAVi – Busca Avançada por Vídeos baseada em transcrição de áudio, metadados e anotação semântica

Prof. Dr. Eduardo Barrére

Universidade Federal de Juiz de Fora - UFJF

Julho de 2015

1. Título

GT-BAVi: Busca Avançada por Vídeos baseada em transcrição de áudio, metadados e anotação semântica

2. Coordenador

Coordenador

Prof. Dr. Eduardo Barrére
Departamento de Ciência da Computação - DCC
Universidade Federal de Juiz de Fora – UFJF
Lattes: http://lattes.cnpq.br/0735298552666402

Coordenador Adjunto

Prof. Dr. Jairo Francisco de Souza
Departamento de Ciência da Computação - DCC
Universidade Federal de Juiz de Fora – UFJF
Lattes: http://lattes.cnpq.br/4516605108233899

3. Resumo

O compartilhamento de vídeos é algo cada vez mais comum nos dias atuais, seja pela facilidade na sua produção ou disponibilização. Um dos principais motivos para o sucesso dos vídeos na internet é que esta mídia é um importante recurso para o aprendizado ou lazer, pois agregada elementos visuais, textuais e auditivos. Acompanhando esse cenário, a RNP disponibiliza diversos serviços que têm como elemento principal o vídeo. O desafio, para todos que querem disponibilizar vídeos, é como facilitar e ampliar o processo de busca em seus acervos. Através da extração de informações de áudio e metadados, gerando a anotação semântica de transcrições automáticas, este projeto pretende realizar a indexação de vídeos disponíveis nos Serviços da RNP, ampliando as possibilidades de busca e recomendação de vídeos.

4. Abstract

Sharing videos is something increasingly common these days, especially due the facility to produce or make available a video. A major reason for the success of videos on the Internet is that this media is an important resource for learning and leisure, because aggregate visuals, text and audio. Following this scenario, RNP offers several services whose main element is the video. The challenge, not only for RNP, but for all who want to display videos, is to facilitate and expand the search process in their collections. By extracting information from audio and metadata, generating semantic annotation of automatic transcriptions, this project aims to carry out indexing of videos available in the Services of RNP, which expands the possibilities of search and recommendation.

5. Parcerias

O projeto será desenvolvido em sua integridade na Universidade Federal de Juiz de Fora (UFJF), pelos laboratórios de pesquisa:

- Laboratório de Aplicações e Inovação em Computação – LapIC (Multimídia, IPTV e videoaulas)

- Núcleo de Engenharia do Conhecimento – NEnC (Recuperação de Informação, Resolução de Identidade e Representação do Conhecimento)

5.1. Realizações e Competências

Os grupos de pesquisa LApIC/UFJF e NEnC/UFJF possuem um histórico de trabalhos desenvolvidos nas áreas de multimídia e recuperação da informação baseada em análise semântica. Recentemente, várias ações desses grupos estão voltadas para a produção e uso de vídeos educacionais e recuperação de vídeos em repositórios, respectivamente.

5.2.1. Projetos em colaboração

Alguns dos projetos desenvolvidos pelos grupos de pesquisa nos últimos anos:

2007 – 2009 GT-EDAD - Grupo de Trabalho em Educação a Distância

Descrição: O projeto visa colocar o ambiente distribuído do servidor multimídia RIO, como um serviço operacional da RNP. Disponibilizar através do servidor as videoaulas do curso de Tecnologia em Sistemas de Computação do consórcio CEDERJ.
Atuação: Desenvolvimento do RIOComposer
Financiamento: RNP

2008 – 2010 GingaFrEvo & GingaRAP: Evolução do Middleware Ginga para Múltiplas Plataformas (Componentização) & Ferramentas para Desenvolvimento e Distribuição de Aplicações Declarativas

Descrição: o projeto é subdividido em dois subprojetos: 1) GingaRAP: suporte a autoria de aplicações. 2) GingaFrEvo: evolução da tecnologia Ginga.
Proponente: PUC-Rio. Parceiros: UFScar, UFPB, UFRN, UFES, UFMA, UFJF, ...
Atuação: Desenvolvimento do módulo de análise de audiência em TV Digital como parte do GingaFrEvo. Financiamento: CTIC/RNP

2010 – 2012 Enriquecendo dados com tecnologias semânticas

Descrição: Colaboração com pesquisadores da Universidade Livre de Berlin e da UNIRIO para criação de uma versão em português do DBPedia. Criação de aplicações que utilizam dados ligados, entre elas, dados governamentais. Criação de infraestrutura para disponibilização de dados governamentais abertos e ligados.
Situação: Concluído; Natureza: Pesquisa.
Atuação: Coordenação do projeto
Financiamento: CNPq

2014-atual GT-IpeTeVê - Serviço de Televisão IP de Alcance Global

Descrição: O objetivo geral do GT-IpeTeVê é a interconexão gerenciada de instituições ligadas à RNP ao testbed global de IPTV/IPv6 da UIT, o I3GT, para o oferecimento de um conjunto inicial de serviços de infraestrutura desenvolvidos pelo GT, que culminem na disponibilização de conteúdo multimídia de interesse da RNP sob a forma de serviço IPTV globalmente acessível.
Situação: Em andamento; Natureza: Pesquisa.
Atuação: Colaboração no desenvolvimento das Interfaces e funcionalidades dos diversos subsistemas envolvidos. Financiamento: RNP

5.2. Experiência da Equipe

Prof. Dr. Eduardo Barrére (UFJF) http://lattes.cnpq.br/0735298552666402

- eduardo.barrere@ice.ufjf.br

Possui graduação em Bacharelado em Ciência da Computação - UFSCar (1996), mestrado em Ciência da Computação - UFSCar (1998) e Doutorado em Engenharia de Sistemas e Computação - COPPE/UFRJ (2007). Professor adjunto IV da Universidade Federal de Juiz de Fora. Responsável pelo Laboratório de Aplicações e Inovação em Computação (LApIC) da UFJF. Desenvolve pesquisas na área de Multimídia, com atividades também nas áreas de redes de computadores e TV Digital. Avaliador da SERES / MEC. Professor permanente dos programas de pósgraduação da UFJF em Educação Matemática (Profissionalizante) e Ciência da Computação (Acadêmico).

Prof. Dr. Jairo Francisco de Souza (UFJF) http://lattes.cnpq.br/4516605108233899

- jairo.souza@ice.ufjf.br

Possui graduação em Ciência da Computação pela Universidade Federal de Juiz de Fora (2004), mestrado em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (2007) e doutorado em Informática pela PUC-RJ, ambos na linha de Banco de Dados. É professor do Departamento de Ciência da Computação da Universidade Federal de Juiz de Fora e participa de pesquisas sobre Recuperação de Informação, Resolução de Identidade e Representação do Conhecimento.

6. Duração do projeto

12 meses.

7. Sumário executivo

7.1 Introdução

É de conhecimento comum o grande volume de *tablets* e *smartphones* vendidos no Brasil nos últimos anos. É possível também agregar a essa informação o fato das novas tecnologias de acesso à internet para dispositivos móveis (3G e 4G) estarem cada vez mais consolidadas e baratas, mesmo o Brasil estando atrasado na adoção dessas tecnologias e não ter preços tão convidativos.

Algumas situações interessantes ocorrem a partir dessa tupla, dispositivos móveis e internet móvel, como o destaque da população brasileira na quantidade de usuários e no uso de diversos aplicativos, como Facebook, Instagram e WhatsApp. Em comum entre essas aplicações, a troca de mensagens, fotos e vídeos. Outro serviço muito utilizado no Brasil é o YouTube, onde são postados, por usuário do mundo todo, cerca

de 300 horas de vídeo por minuto[1]. Em pesquisa realizada no ano de 2013[2], o YouTube conta com mais de 40 milhões de espectadores no país, sendo que mais de 40% deles se interessam por vídeos que contenham entrevistas, documentários ou tutoriais (Barrére, 2014).

Outros repositórios de vídeos também mostram um grande crescimento, mas todos eles apresentam um desafio em comum, como facilitar a busca por vídeos? Cada um adota estratégias diferentes, mas que basicamente envolve: cadastro de informações básicas (título, palavras-chave etc), metadados (principalmente no caso de Objetos de Aprendizagem), legendas, entre outras soluções.

Não distante desse cenário, a RNP também oferece diversos serviços relacionados a vídeo, essencialmente os serviços de Disponibilização de Conteúdos Digitais. Entre eles se destacam os serviços de Vídeo Sob Demanda e Videoaula@RNP, pois são baseado na disponibilização de vídeos previamente armazenados. Nesta vertente, alguns projetos fomentados pela RNP também poderão, em um futuro próximo e caso venham a se tornar serviços, lidar com vídeos armazenados (GT-VOA e GT-IpêTeVê).

Nesse cenário, a RNP também passa a ter os mesmos desafios dos repositórios de vídeo existentes na internet, essencialmente capacidade de armazenamento, banda de internet e processamento dos servidores para consumo e como facilitar a busca dos usuários no repositório. Uma das características dos serviços da RNP que armazenam vídeos, em relação aos repositórios de vídeo na internet, é o menor volume de vídeos inseridos por unidade de tempo.

O GT-BAVi pretende atuar não só nas informações básicas de um vídeo (título etc.) e metadados, mas também na transcrição do áudio e posterior análise semântica do mesmo. Isto será feito para ampliar as possibilidades do usuário encontrar um vídeo que contemple sua busca ou para que mecanismos de recomendação possam gerar uma indicação mais adequada para aquele usuário. Sendo assim, o projeto é transversal a alguns serviços já existentes na RNP, ou seja, ele pretende colaborar com o processo de busca dos vídeos armazenados nesses serviços.

Esta proposta de grupo de trabalho é fundamentada nas competências técnicas da equipe proponente e em demanda observada durante diversas reuniões da RNP, principalmente na reunião SIG Conteúdos Digitais, ocorrida em novembro de 2014, na cidade de Salvador, na qual diversos usuários de TVs Universitárias abordaram o desafio de disponibilizar vídeos “de forma adequada” para que posteriormente fossem “encontrados” e utilizados pelas demais TVs participantes. Desafio também enfrentado pelos mantenedores do Videoaula@RNP, tornar os conteúdos disponibilizados mais “atrativos” e fáceis de serem encontrados.

7.2 Fundamentação

Para recomendar arquivos de vídeo, é necessário garantir que estes vídeos estejam corretamente recuperados ou classificados através da identificação de contextos ou conceitos principais apresentados no vídeo. Para muitos buscadores, a recuperação de vídeo é realizada utilizando metadados, texto encontrado no contexto do vídeo (ao redor

[1] http://www.youtube.com/yt/press/pt-BR/statistics.html

[2] http://www.proxxima.com.br/home/negocios/2013/07/25/Brasil-e-um-dos-paises-que-mais-cresce-emconsumo-de-videos-online.html

do arquivo), entre outras informações textuais (Sack & Waitelonis, 2010; Turnbull et al, 2008). Alguns sistemas buscam categorizar vídeos através de técnicas de reconhecimento de objetos e ações (Maybury, 2012). Somar estas técnicas com técnicas baseadas em texto pode trazer resultados relevantes.

Para facilitar a indexação de vídeos, é comum a seleção e marcação de palavraschave, de forma manual que identificam e descrevem o conteúdo (Stamou & Kollias, 2005). 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 (Raimond & Lowis, 2012), 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 (Specia & Motta, 2007) para tratar esse problema sem muito esforço manual.

Outra técnica usada para melhorar a busca de vídeos é a utilização de textos transcritos, permitindo que o vídeo (ou áudio) possa ser recuperado através de palavras faladas ao longo da sua reprodução (Natarajan et al, 2012). Ainda que a utilização da transcrição, uma vez que é realizada de forma manual, é muito custosa e raramente é utilizada em sistemas de busca (Natarajan et al, 2012).

Considerando a problemática acima, é necessário o uso de técnicas próprias para indexação de grandes volumes de vídeos e áudios para permitir seu uso em grandes repositórios, sejam eles de objetos de aprendizagem ou não.

Para tratar esse problema, foi desenvolvida a prova de conceito de um framework (Coelho & Souza, 2014) para ser utilizado na 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, (5) fornecer interfaces para integração com ferramentas de Question Answering (Yao & van Durme, 2014). A arquitetura projetada para atingir os objetivos anteriormente descritos pode ser visualizada a seguir:

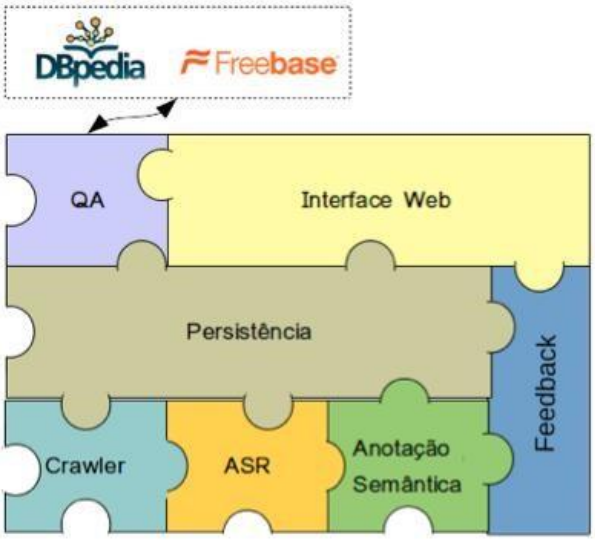

Figura 1. Arquitetura do módulo de recomendação por transcrição de áudio

O submódulo 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 submódulo é responsável por notificar o submódulo de ASR sobre a ocorrência de novos arquivos, os quais serão processados para geração dos seus transcritos. Este submódulo é responsável também pela aplicação de filtros no áudio ou vídeo e conversão de formatos. No contexto deste projeto, este submódulo faria a comunicação com os Serviços da RNP que armazenam vídeos.

O submódulo 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 do ASR, é possível configurar o uso de ferramentas externas. Inicialmente, o submódulo de ASR realiza transcrições em idioma português utilizando a ferramenta Coruja JLapsAPI. Na fase de préprocessamento, o áudio é extraído do vídeo e segmentado em trechos de menos de dois minutos.

A transcrição de áudio é um processo que pode gerar textos com muitos ruídos. Dessa forma, não é viável o uso de técnicas clássicas de recuperação de informação, fazendo-se necessário desenvolver técnicas para identificar contexto baseado em textos com ruídos. O principal problema encontrado na transcrição é a presença de erros causados por homofonia, ou seja, palavras que possuem sons parecidos e que podem ter sido transcritas erradas.

O submódulo de anotação semântica tem como objetivo associar automaticamente tags para os vídeos. O submódulo recebe como entrada as transcrições geradas pelo ASR. O submódulo de anotação semântica necessita avaliar a transcrição, identificar o contexto do áudio e escolher qual o conjunto de tags que melhor descrevem o vídeo, considerando a característica imprecisa do áudio transcrito. O submódulo permite que o desenvolvedor acrescente diferentes estratégias na escolha de tags. Porém, quatro estratégias estão disponíveis para serem utilizadas: Spotlight Lucene (Mendes et al, 2011), Spotlight JDBM (Daiber et al, 2013), eTVSM (Raimond & Lowis, 2012) e matrizes de coocorrência. É permitido que o desenvolvedor defina o uso de mais de uma estratégia ao mesmo tempo. Neste caso, o submódulo é responsável por combinar os resultados das diferentes estratégias. Novas estratégias podem ser introduzidas neste submódulo e serão desenvolvidas ao longo do projeto.

O submódulo de feedback é utilizado por usuários especialistas para avaliar manualmente o resultado da classificação dos vídeos de forma a criar uma base para retroalimentação dos modelos utilizados para classificação.

Ao ser utilizado um banco de dados de grafos 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 submódulo de busca permite que métodos de busca possam ser inseridos na aplicação. O submódulo recebe a consulta

do usuário e possui acesso ao submódulo 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 palavrachave 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, em que 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?”.

No caso específico desta proposta, o submódulo de Interface Web poderá ser uma interface propriamente dita ou somente um serviço web (*web service*) a ser integrado aos demais serviços da RNP, estritamente na parte de busca e juntamente com os demais componentes de interface de cada serviço.

7.2 Objetivos

O objetivo principal do projeto é ampliar e qualificar o tipo de informação a ser utilizada dos vídeos armazenados nos serviços da RNP, com a finalidade de facilitar a busca dos usuários e ampliar a visibilidade dos mesmos.

Como objetivos secundários, temos:

- Fomentar o uso dos serviços da RNP baseados em armazenamento de vídeo
- Detectar, criar e ampliar ontologias relacionadas às temáticas dos vídeos da RNP (Souza, 2014)
- Fornecer base de conhecimento para posterior análise e classificação dos vídeos armazenados nos repositórios de cada tipo de serviço, permitindo assim ampliar futuras ações estratégicas (novos usuários de serviço, divulgação mais direcionada etc.).

7.3 Potenciais Impactos

Como potenciais impactos, temos:

- Políticas mais eficazes para ampliar o número de usuários desses serviços
- Melhorar, ampliando as possibilidades de sucesso, a busca por vídeos e em consequência disso o consumo desses serviços
- Contribuir com as pesquisas na área de transcrição de vídeos e anotação semântica

8. Recursos financeiros

8.1. Equipamentos e softwares

Descrição	Quantidade	Valor total
Notebook 14" (Core i7 - 8GB - 500GB) - R$ 4.087,00	1	4.087,00
Desktop s/ monitor (Core i7 - 8GB - 500GB) - R$ 3.600,00	3	10.800,00

Monitor LED 23.8" - R$ 800,00	3	2.400,00
Servidor s/ monitor (rack) – R$ 6.500,00	1	6.500,00
Valor Total (de acordo com o Anexo 3)		23.787,00

9. Ambiente para testes do protótipo

O protótipo será testado, inicialmente, usando a infraestrutura provida pelo projeto (servidor) no LApIC/UFJF, utilizando para tal finalidade videoaulas elaboradas no RIOComposer, aplicações multimídia geradas no Cacuriá e vídeos armazenados no serviço Vídeo sob Demanda da RNP, caso seja possível a disponibilização. Numa fase seguinte, o protótipo poderá ser testado de forma integrada aos serviços da RNP, sem modificar os conteúdos de cada serviço, mas sim tendo acesso aos vídeos disponíveis por eles e gerando uma camada complementar de busca.

10. Referências

Barrére, E. Videoaulas: aspectos técnicos, pedagógicos, aplicações e bricolagem. In: Maria Augusta Silveira Netto Nunes, Elizabeth Matos Rocha. (Org.). Anais da Jornada de Atualização em Informática na Educação. 1ed. Dourados: EaD-UFGD, 2014, v. 1, p. 70-105.

Coelho, S. A., Souza, J. F. Anotação Semântica de Transcritos para Indexação e Busca de Vídeos. In: Conferência Ibero Americana WWW/INTERNET, 2014, Porto, Portugal. 12ª Conferência Ibero Americana WWW/INTERNET. IADIS, 2014. v.1. p.51 – 58.

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

Maybury, M. T. Section 2: Video Extraction , in Multimedia Information Extraction: Advances in Video, Audio, and Imagery Analysis for Search, Data Mining, Surveillance, and Authoring, pages 113–117. John Wiley and Sons, Inc., 2012.

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

Natarajan, P., Macrostie, E., Prasad, R. & Watson, J. Analysis of Multimodal Natural Language Content in Broadcast Video, pages 175–184. John Wiley and Sons, Inc., 2012.

Raimond, Y. & Lowis, C. 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.

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

Souza, J. F., Siqueira, S. W. M., Melo, R. N. & Lucena, C. J. P. Análise de abordagens populacionais para meta-alinhamento de ontologias. iSys: Revista Brasileira de Sistemas de Informação, v. 7, p. 75-97, 2014.

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

Yao, X. & van Durme, B. Information extraction over structured data: Question answering with freebase. In Proceedings of ACL, 2014.