RNP

Proposta de Projeto – Fase Piloto

GT-BAVi – Segunda Fase

Busca Avançada por Vídeos baseada em transcrição de áudio, metadados e anotação semântica

Eduardo Barrére (Universidade Federal de Juiz de Fora)

13 de Dezembro de 2016.

1. Visão geral

1.1. Descrição do produto/serviço resultante do piloto

O produto do piloto consiste numa arquitetura distribuída e escalonável que permita a transcrição de áudio (oriundos de vídeos ou não), a anotação semântica de textos (transcritos ou naturais) e a recomendação dos textos anotados semanticamente, visando à categorização desse conteúdo (conforme a DBpedia) e relacionamento com outros conteúdos digitais do mesmo serviço. Todas essas etapas podem estar encadeadas, ou serem chamadas de forma isolada, via API do piloto.

Neste cenário, será possível agregar valor a conteúdos digitais disponibilizados por serviços da RNP (videoaulas, vídeo@rnp, ICD etc.), através da ampliação dos termos de busca (quantidade e relevância) e relacionamento entre conteúdos do mesmo serviços, permitindo uma maior visibilidade a esses conteúdos.

Como produtos secundários, temos:

- Ampliação da base de conhecimento dos conteúdos digitais armazenados nos serviços da RNP (categorias e relacionamento entre conteúdos), permitindo a mineração desses dados para melhor entender o público que disponibiliza conteúdos nos serviços.
- Criar uma plataforma que permita a qualquer serviço da RNP chamar suas funcionalidades. Um exemplo de utilização do piloto, além dos conteúdos no formato de vídeo, seria o portal de notícias da RNP.
- Contribuir com o projeto DBpedia, através da colaboração no desenvolvimento do DBpedia Spotlight ou mesmo com a inserção de novos termos na wikipedia.
- Ampliação de um modelo em português para a transcrição de áudio, permitindo que a RNP possa compartilhá-lo com a sociedade.

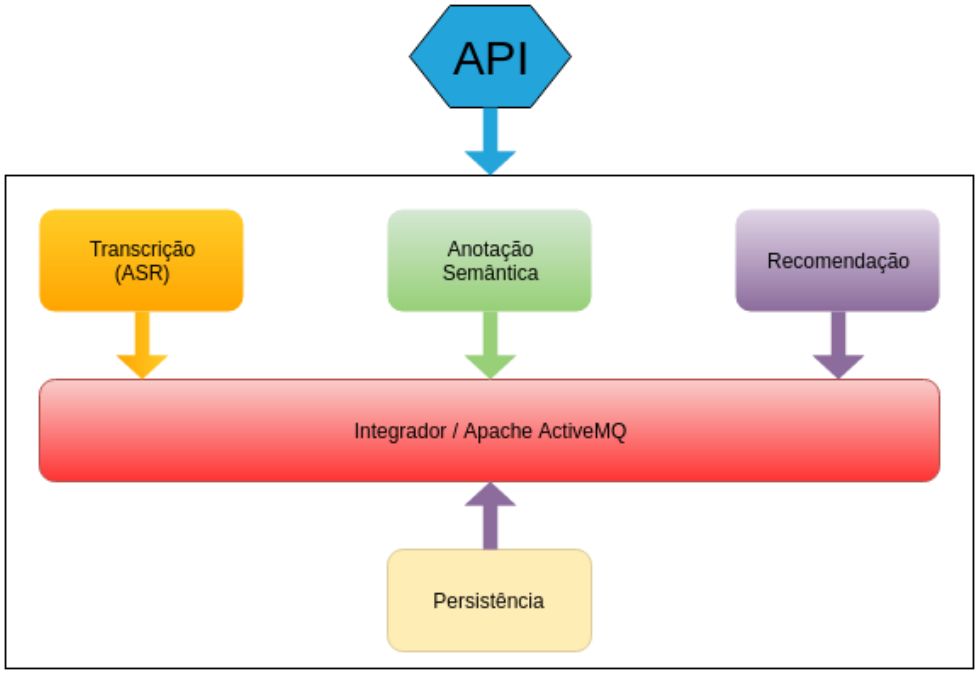
1.2. Identificação do público alvo

O público alvo é composto pelas pessoas que buscam por conteúdos armazenados pelos serviços de intercâmbio de conteúdos digitais da RNP, mas como o foco do projeto está em agregar valor a conteúdos armazenados por diversos serviços, também poderia ser considerados como público alvo, os próprios serviços de disponibilização de conteúdos digitais atuais e os que venham a ser ofertados futuramente, pois seus conteúdos estarão mais bem relacionados e com uma maior quantidade e qualidade de termos de busca.

2. Definição do piloto

2.1. Arquitetura do piloto

A arquitetura do piloto está apresenta na Figura a seguir e está organizada em três blocos: Principal (API, Integrador, Persistência e Recomendação), Transcrição e Anotação Semântica.



Arquitetura Básica do piloto do GT-BAVi

Uma descrição sucinta dos componentes da arquitetura está presente na Tabela a seguir.

Módulo	**Máquina Virtual**	**Funcionalidade**	**Softwares**	**Observações**
API	1	Permite que os serviços da RNP solicitem as funcionalidades (transcrição e/ou anotação e/ou recomendação)	Desenvolvimento próprio em java.	----------
Integrador		Responsável por integrar, via troca de mensagens, todas as etapas do processo, desde a solicitação de processamento, até a entrega dos resultados obtidos.	ActiveMQ (livre) + desenvolvimento próprio	Permite a comunicação com processos na mesma MV ou em outras MV, possibilitando assim toda a escalabilidade prevista para o piloto.
Persistência		Armazena as solicitações da API e todos os resultados (intermediários e finais)	Blazegraph (livre)	----------
Recomendação		Associa ao texto anotado, categorias e conteúdos relacionados.	implementação própria	A base de conhecimento utilizada é a DBpedia
Transcrição	0 a n *	O ASR (Automatic Speech Recognition) é o módulo responsável por realizar a transcrição do áudio	Kaldi (livre)	Esta MV entrar em funcionamento, somente quando existirem recursos disponíveis no CDC. Cada MV pode utilizar tecnologias distintas (Kaldi, Coruja, etc.)
Anotação	0 a n *	A anotação semântica permite a busca por recursos, conforme suas categorias.	DBpedia Spotligth (livre) + implementação própria	Esta MV entrar em funcionamento, somente quando existirem recursos disponíveis no CDC. Cada MV pode utilizar tecnologias distintas (DBpedia Spotligth, implementações próprias, etc.)

2.2. Instituições participantes

Não teremos instituições parceiras, mas sim serviços parceiros. Nossa ideia inicial seria ter como parceiro algum serviço de conteúdo digital baseado em vídeo, principalmente o videoaulas@RNP, devido ao bom conhecimento da equipe sobre os padrões utilizados por este serviço, mas outros serviços como video@RNP e ICD podem também serem parceiros no piloto. Outro conteúdo interessante é o disponibilizado pelo portal de notícias da RNP, por ser baseado em texto natural (não transcrito). De qualquer forma, a RNP pode priorizar outros serviços, sem maiores prejuízos para o andamento do piloto.

Para que o serviço possa ser parceiro neste piloto, é necessário que ele faça chamadas à API do GT-BAVi, quando receber um novo conteúdo e depois para inserir o resultado do processamento em seu mecanismo de busca.

2.3. Objetivos e evoluções

O protótipo desenvolvido na fase precisa passar por melhorias para que possa ser testado pelos parceiros. Abaixo listamos as principais melhorias propostas pelo GT:

- **Desenvolvimento de interface de gestão do serviço:** é necessário que a RNP possua um *dashboard* para visualizar o fluxo de trabalho da API e uma interface administrativa para alterar os parâmetros do serviço, definir instâncias de processamento, etc. A interface administrativa permitirá ao usuário executar os módulos do servidor em servidores distintos, permitindo administrar os recursos disponíveis.
- **Melhorias na arquitetura do serviço:** é necessário implementar um modelo de autenticação de acesso ao serviço e permitir que algumas tarefas sejam agendadas. Assim, essas melhorias serão implementadas no serviço e estarão disponíveis via interface administrativa e pela API do serviço. Para autenticação, o GT propõe a implementação de uma autenticação baseada em tokens.
- **Melhoria no modelo de transcrição para português:** o modelo gratuito que o GT usou no começo do projeto (Coruja) possui uma taxa de erro muito alta e o GT começou a utilizar um modelo proprietário cedido gratuitamente durante um período de experiência. Este modelo gera uma taxa de erros muito pequena. No último mês do projeto, o GT começou a investir no treinamento de um modelo próprio, o que já possui uma taxa de erros melhor que a do Coruja. Porém, o modelo ainda não está adequado para produção. Assim, o GT propõe trabalhar na melhoria do modelo de transcrição, o que envolve a criação de uma base maior de treinamento, esforço em normalização do *corpus*, etc. Identificamos três vantagens em investir nesta tarefa: (1) dominar o processo de treinamento, o que permitirá à RNP e ao GT definir modelos específicos para alguns domínios de conhecimento; (2) possuir um modelo de transcrição sem a necessidade de pagamento de licença de uso; (3) distribuir uma melhor base de treinamento para português e com os dados da própria RNP (oriundos de vídeo-aulas, notícias do site da RNP, etc). Vale

ressaltar que o corpus mais conhecido para treinamento em língua portuguesa é o oferecida pelo projeto FalaBrasil, o qual possui poucas horas de áudio.

- **Melhoria no processo de anotação semântica**: a tarefa mais difícil do serviço proposto pelo GT é a anotação semântica. Levando em consideração que mesmo uma transcrição muito boa pode gerar um número de erros de palavras que pode dificultar o entendimento de algumas frases, a quantidade de ruídos pode inviabilizar o uso de algoritmos clássicos de anotação. Assim, o GT tem trabalhado em soluções de anotação de textos com elevado nível de ruídos para melhorar a acurácia da anotação. Esta é a tarefa que demanda maior esforço de pesquisa acadêmica.
- **Otimização dos módulos de anotação e recomendação**: os dois módulos possuem um tempo de processamento alto e é possível reduzir o processamento. O GT propõe o pré-processamento de alguns dados para reduzir o tempo de execução do módulo de anotação. Para o módulo de recomendação, o GT propõe utilizar os dados do DBpedia de forma off-line, realizando também um pré-processamento da base do DBpedia.
- **Geração de resultados detalhados dos vídeos processados:** atualmente o GT gera a transcrição e anotação de um vídeo completo. Contudo, é possível que o GT segmente o vídeo e entregue a transcrição e a anotação de parte do vídeo. Isso permitirá que os serviços de busca possam direcionar o usuário do serviço para um trecho específico do vídeo, permitindo o usuário assistir o trecho em que é falado alguma palavra ou o trecho onde um dado conceito é abordado.

Os resultados do piloto contribuirão diretamente no melhor entendimento e visibilidade dos serviços de disponibilização de conteúdos digitais da RNP, pois ao gerar o processamento dos conteúdos de um determinado serviço, será possível realizar um mapeamento mais detalhado do tipo de conteúdo disponibilizado e, principalmente, pelo fato de gerar o enriquecimento nos termos de busca e relacionamento entre conteúdos do mesmo serviço, será possível uma maior visibilidade desses conteúdos, seja pelo fato de estarem associados a um maior número de termos de busca, ou pelo fato de estarem relacionados a outros conteúdos de maior popularidade.

3. Macro cronograma de desenvolvimento do piloto

Listar e descrever todas as macro atividades que permitirão o alcance dos objetivos indicados na seção 2.3, dando foco especificamente ao **desenvolvimento tecnológico** necessário para a evolução do produto e melhorias em processos de gestão e uso dos resultados relacionados ao protótipo existente. Destacar em quais trimestres serão realizadas.

Este cronograma deve atentar e estar alinhado as entregas pré-definidas, veja seção 6, em especial em conformidade e com nível de desenvolvimento compatível ao momento do projeto:

- Demonstração dos Resultados Parciais no Workshop da RNP (WRNP) – maio/2017

- Workshop de Disseminação dos Resultados para Instituições Clientes da RNP – setembro/2017
- Apresentação Final dos Resultados para o Comitê de Avaliação e RNP – outubro/2017

Macro atividades	**1º. Trim.**	**2º. Trim.**	**3º. Trim.**	**4º. Trim.**
1. Implantação da arquitetura em MVs separadas	X	X		
2. Teste de integração com serviços		X	X	
3. Testes de escalabilidade da arquitetura		X	X	
4. Desenvolvimento do DashBoard (interface de configuração e controle)	X	X	X	
5. Treinamento dos Modelos de acústicos e de linguagem	X	X	X	X
6. Desenvolvimento de algoritmos de anotação semântica		X	X	X
7. Otimização dos módulos da arquitetura	X	X	X	X
8. Elaboração de relatórios e participação em eventos da RNP	X	X	X	X

4. Recursos para o desenvolvimento do piloto

4.1. Recursos oferecidos pela RNP para execução do piloto

A RNP oferece alguns ambientes que podem ser utilizados para o desenvolvimento e testes do piloto como:

- Recursos virtualizados em Pontos de Presença (PoPs) da RNP (http://www.rnp.br/institucional/pontos-presenca)
- Ambiente PlanetLab (http://www.rnp.br/pesquisa-e-desenvolvimento/redes-experimentacao)
- Ambiente de experimentação em Internet do Futuro, conhecido como FIBRE (*Future Internet Research and Experimentation*) e que está disponível em http://www.rnp.br/pesquisa-e-desenvolvimento/internet-futuro.
- Laboratório de Gestão de Identidade https://gidlab.rnp.br/

Estes e outros recursos disponíveis no ambiente de produção da RNP (http://www.rnp.br) e considerados necessários ao desenvolvimento do piloto, poderão ser listados nesta proposta na seção 5.2 Recursos para o projeto.

4.2. Recursos virtualizados para o desenvolvimento do piloto

Para o desenvolvimento do piloto, a infraestrutura oferecida pelo CDC da RNP certamente será satisfatória. Os recursos necessários estão apresentados na Tabela a seguir.

Máquina Virtual	Quantidade de MVs	Memória	Processador	Disco	Outros Recursos
Integrador, Persistência e Recomendação	1	16GB	>= 2 núcleos	1TB	------
Transcrição	1 a n *	>= 16GB	>= 2 núcleos	500GB	------
Anotação	1 a n *	>= 32GB	>= 2 núcleos	500GB	------
Teste e treinamento do Modelo Acústico e de Linguagem.	1	>= 16GB	>= 2 núcleos	1TB	Se possível, uma GPU

* Para os testes iniciais, será necessário habilitar pelo menos uma MV de Transcrição e Anotação, mesmo que em horários pré-agendados. A instanciação de mais MVs poderá ocorrer somente nos horários disponíveis.

4.3. Pessoal

Descrever o nome completo e a função de cada membro da equipe e respectivos valores em R$ (bruto)

Nome Completo	Função	Valor mensal*	Data de Início	Data de término	Valor total
Eduardo Barrére	Coordenador geral	R$ 2.100,00	01/01/2017	31/12/2017	R$ 25.200,00
Jairo Francisco de Souza	Coordenador adjunto	R$ 1.500,00	01/01/2017	31/12/2017	R$ 18.000,00
Laura Lima Dias	Assistente 3	R$ 1.300,00	01/01/2017	31/12/2017	R$ 15.600,00
Marcelo Machado	Assistente 3	R$ 1.300,00	01/01/2017	31/12/2017	R$ 15.600,00
José Eduardo de Carvalho Silva	Assistente 3	R$ 1.300,00	01/01/2017	31/12/2017	R$ 15.600,00
Nicolas Ferranti	Estagiário	R$ 680,00	01/01/2017	31/12/2017	R$ 8.160,00
Marcos Valadão Gualberto Ferreira	Estagiário	R$ 680,00	01/01/2017	31/12/2017	R$ 8.160,00
Jorão Gomes Junior	Estagiário	R$ 680,00	01/01/2017	31/12/2017	R$ 8.160,00
João Paulo Radd Pires da Silva	Estagiário	R$ 680,00	01/01/2017	31/12/2017	R$ 8.160,00
<a definir>**	Estagiário	R$ 680,00	01/01/2017	31/12/2017	R$ 8.160,00
	Total	R$ 10.900,00			R$ 130.800,00

* O limite total mensal não deve ser superior a R$ 11.000,00 (valor bruto antes dos impostos)

** Estagiário para auxiliar no processo de treinamento dos modelos de transcrição (acústico e de linguagem).

5. Cronograma e entregas pré-definidas

Os relatórios de planejamento, relatórios técnicos, relatórios de acompanhamento e demais entregas listadas a seguir são pré-definidas e fazem parte integrante desta proposta e devem ser entregues pela equipe deste Grupo de Trabalho à Gerência de Grupos de Trabalho, conforme cronograma indicado nesta seção. Também deverão ser realizadas entregas referentes a documentação, participação em eventos presenciais (WRNP, Workshop de Disseminação do GT e Workshop de Apresentação de Resultados) entre outros que compõem o desenvolvimento do projeto.

Os modelos destes relatórios e demais entregas serão compartilhados com o coordenador do GT na ocasião da reunião de boas vindas em data a ser agendada com os projetos selecionados para fase 2.

5.1. Relatórios

Os relatórios são entregas do projeto (Relatórios de Planejamento – RP e Relatórios Técnicos), na articulação com os grupos de outras organizações envolvidos no mesmo tema. O acompanhamento dos resultados parciais é realizado a partir dos relatórios trimestrais de acompanhamento (Relatório de Acompanhamento – RA) e na apresentação e discussão do tema no Workshop RNP (WRNP) e na transferência de conhecimento feita à RNP.

As responsabilidades da coordenação do projeto por parte dos contratados englobam a gestão do projeto do GT, incluindo a utilização da Wiki da RNP para disponibilização de informações sobre ações, atividades e tarefas, assim como de indicadores de progresso e status.

Além disso, todo o código fonte deve ser mantido atualizado pela equipe de desenvolvimento diretamente no ambiente de desenvolvimento colaborativo a ser indicado e disponibilizado pela RNP.

Os relatórios são agrupados em três tipos:

5.1.1. Relatórios de Planejamento (RPs)

RP4: Planejamento de Recursos Virtualizados

Estimativa de demanda, especificação detalhada de máquinas virtuais necessárias ao desenvolvimento e implantação do piloto, com as respectivas justificativas de dimensionamento.

RP5: **Planejamento da estrutura de pacotes de trabalho de desenvolvimento tecnológico e do cronograma de entregas destes pacotes**

Estrutura de pacotes de trabalho a serem realizadas ao longo do GT que descrevem os principais grupos de atividades que são necessários para desenvolvimento deste projeto. Um pacote de trabalho é um grupo de atividades que não deve durar mais do que 3 meses de execução. Cada pacote de trabalho deve ter uma data de entrega associada e o cronograma de marcos é a distribuição das datas de entrega de cada pacote de trabalho ao longo dos 12 meses de projeto.

RP6: Planejamento do Workshop RNP (WRNP)

Descrição da demonstração a ser realizada, equipamentos necessários, lista de integrantes do GT que irão participar, texto e demais documentos para divulgação no evento (O WRNP ocorrerá junto com o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC 2015, em Vitória, Espírito Santo).

RP7: Relatório de planejamento do Workshop de Disseminação do GT

Descrição das atividades previstas para a reunião de encerramento do projeto piloto. Este workshop pode ser apenas com os usuários participantes do piloto ou reunindo outros usuários interessantes para disseminação do resultado do projeto. Ex.: apresentações sobre o projeto e das experiências dos usuários do piloto, tutorial

ministrado durante o workshop, além de documentação, manuais e códigos-fonte a serem disponibilizados.

RP8: Relatório de planejamento para inclusão no portfólio da RNP

Definição de como será a inclusão do produto no portfólio da RNP, detalhando onde será disponibilizado o produto ou o código, onde é possível encontrar mais informações sobre o produto online (página do produto na RNP ou em site do próprio), como será disseminado (por exemplo: via um curso na grade da ESR ou via manuais de usuário e tutoriais abertos) e seu modelo de sustentabilidade.

5.2. Relatórios Técnicos (RTs)

Os relatórios técnicos devem refletir os resultados das atividades realizadas pelo GT para alcançar o seu objetivo de implantação de um piloto.

RT4: Mapeamento de componentes e licenças de software

Descrição detalhada de cada componente (novo ou de reuso) que compõe a arquitetura do piloto, bem como sua respectiva licença de software. O entregável desse relatório deverá ser uma página na wiki onde as licenças e componentes podem ser incrementados ao longo do projeto.

RT5: Plano de testes do piloto

Descrição detalhada dos testes a serem realizados para a avaliação do piloto, indicando os procedimentos, resultados esperados e cronograma.

RT6: Avaliação dos resultados do piloto

Descrição dos resultados obtidos nos testes descritos no RT4, contendo avaliação, relato dos problemas encontrados e das soluções implementadas.

RT7: Recomendações para a implantação

Descrição da proposta de implantação, identificando o público alvo; descrição e dimensionamento da infraestrutura necessária para a implantação dos resultados; arquitetura proposta; definição dos processos de monitoração e gerenciamento do serviço; estimativa e perfil dos recursos humanos para a gerência e operação dos resultados.

5.2.1. Relatórios de Acompanhamento (RA)

RA5 a RA9: Relatórios de acompanhamento

Relato do progresso das atividades que foram planejadas no período.

RWRNP: Relatório de participação no WRNP

Relato da experiência da participação no WRNP, como sugestões e considerações dos visitantes ao trabalho do GT.

5.3. Site de divulgação do Grupo de Trabalho

O site para divulgação do GT é: https://sites.google.com/a/ice.ufjf.br/gt-bavi/

5.3.1. Atualização do site do GT

Deverá ser atualizado o site do GT com as informações relevantes do projeto na fase piloto, para disseminação do trabalho. O site do projeto deverá citar o apoio da RNP, com referência ao site da RNP. Deve-se disponibilizar o site do projeto também em inglês.

5.4. Participação no Workshop da RNP (WRNP)

5.4.1. Apresentação em sessão técnica e demonstração do protótipo

A discrição da RNP, deverá ser realizada uma apresentação e uma demonstração técnica da proposta do GT durante o Workshop da RNP (WRNP) nos dias 15 e 16/05 em Belém, PA, que acontece em conjunto com o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2016).

5.5. Workshop de Disseminação do GT

5.5.1. Realização do Workshop de Disseminação do GT

O GT deve organizar um workshop para a disseminação dos resultados do GT para potenciais interessados em absorver os produtos/serviços desenvolvidos durante o piloto, focando nos aspectos técnicos explorados durante o piloto e nos diferentes casos de uso da solução desenvolvida.

5.6. Entrega dos produtos desenvolvidos durante o piloto

5.6.1. Piloto desenvolvido

Fontes, executáveis, scripts, arquivos de configuração etc.

5.6.2. Documentação do piloto

Documentação técnica, manuais de instalação, manuais do usuário etc.

5.7. Avaliação do piloto

5.7.1. Apresentação dos resultados do GT

Deverá ser realizada uma apresentação para um comitê de avaliação dos GTs, com ênfase no piloto desenvolvido e no produto/serviço a ser disponibilizado para os usuários da RNP. A partir dessa avaliação, serão selecionados os GTs que poderão ser recomendados para possível modelagem de serviço/produto para oferta da RNP.

Cronograma de entregas pré-definidas

27/01/2017

- RP4: Planejamento de Recursos Virtualizados
- RP5: Planejamento da estrutura de pacotes de trabalho de desenvolvimento tecnológico e do cronograma de entregas destes pacotes

24/02/2017

- RT4: Relatório de mapeamento de componentes e licenças de software
- RT5: Plano de testes do piloto

31/03/2017

- Site do GT atualizado
- Iniciar a implantação do piloto[1]
- RP6: Relatório de planejamento do WRNP (demonstração, material e viagens)
- RA5: Relatório de acompanhamento trimestral jan/fev/mar

28/04/2017

- Entrega do **código-fonte** da versão implantada no piloto (códigos-fonte, executáveis, *scripts*, arquivos de configuração etc.), incluindo o sistema e as ferramentas de suporte à operação;
- Entrega de **documentação** (manuais de instalação e administração, manuais de usuário etc.).

15/05/2017 a 16/05/2017

- WRNP: Apresentação em sessão técnica e demonstração dos resultados parciais do piloto no Workshop RNP nos dias 15 e 16/05 em Belém, PA.

30/06/2017

- RWRNP: Relatório de participação no WRNP
- RA6: Relatório de acompanhamento trimestral abr/mai/jun

28/07/2017

- RP7: Relatório de planejamento do Workshop de Disseminação do GT

25/08/2017

- RT6: Avaliação dos Resultados do Piloto
- RT7: Recomendações para a implantação do serviço/produto

Entre 01/09/2017 a 30/09/2017 (data a definir)

- Realização do Workshop de Disseminação do GT (data a definir)

Entre 01/10/2017 a 31/10/2017 (data a definir)

[1] Início das atividades planejadas no RP6.

- Apresentação Final dos Resultados para o comitê de avaliação
- RA7: Relatório de acompanhamento trimestral ago/set/out

24/11/2017

- RP8: Relatório de planejamento para inclusão no portfólio da RNP
- Atualização do RT4: Relatório de mapeamento de componentes e licenças de software
- Entrega final do **código-fonte** e **documentação**

15/12/2017

- RA8: Relatório de acompanhamento out/nov/dez

6. Referências

Amaral, R. P. B. 2011. Indexação de Programas Noticiosos (Doctoral dissertation, INSTITUTO SUPERIOR TÉCNICO).

Aquino, M. C. 2007. Hipertexto 2.0, folksonomia e memória coletiva: um estudo das tags na organização da web. E-Compós, Brasília, 9.

Asghar, M. N., Hussain, F., & Manton, R. 2014. Video indexing: a survey. Framework, 3(01).

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. 2007. Dbpedia: A nucleus for a web of open data (pp. 722-735). Springer Berlin Heidelberg.

Awad, A., Polyvyanyy, A., & Weske, M. (2008). Semantic querying of business process models. In IEEE International Conference on Enterprise Distributed Object Computing Conference (EDOC 2008) (pp. 85–94).

Baeza-Yates, R. & Ribeiro-Neto, B. 2011. Modern Information Retrieval - the concepts and technology behind search, Second edition. Pearson Education Ltd., Harlow, England.

Becker, J., & Kuropka, D. 2003. Topic-based vector space model. In Proceedings of the 6th International Conference on Business Information Systems (pp. 7–12).

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C. & Rose, R. 2007. Automatic speech recognition and speech variability: A review. Speech Communication, 49(10), 763-786.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. 2014. Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 56, 85-100.

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

Cordon, O., Moya, F. & Zarco, C. 2004. Fuzzy logic and multi-objective evolutionary algorithms as soft computing tools for persistent query learning in text retrieval environments. In Proceedings of the IEEE international conference on fuzzy systems(pp. 571–576).

Croft, W. B., Metzler, D., & Strohman, T. 2010. Search engines: Information retrieval in practice (Vol. 283). Reading: Addison-Wesley.

Dahl, G. E. 2015. Deep learning approaches to problems in speech recognition, computational chemistry, and natural language text processing (Doctoral dissertation, University of Toronto).

Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. 2013. Improving efficiency and accuracy in multilingual entity extraction. InProceedings of the 9th International Conference on Semantic Systems (pp. 121-124). ACM.

Deléglise, P., Estève, Y., Meignier, S., & Merlin, T. 2005. The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news. In Interspeech (pp. 1653-1656).

D. Milne and I. H. Witten. Learning to link with wikipedia. In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM'08, pages 509–518, New York, NY, USA, 2008. ACM.

Ênio dos Santos Silva, "Desenvolvimento de um reconhecedor automático de voz com suporte a grandes vocabularios para o português brasileiro" Tech. Rep., 2005.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago - a core of semantic knowledge. In 16th international World Wide Web conference, 2007

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-specific keyphrase extraction.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005, July). Text classification: a recent overview. In Proceedings of the 9th WSEAS International Conference on Computers (pp. 1-6). World Scientific and Engineering Academy and Society (WSEAS).

Gravier, G., Bonastre, J. F., Geoffrois, E., Galliano, S., McTait, K., & Choukri, K. 2004. The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In LREC.

Gravier, G., Jones, G. F., Larson, M., & Ordelman, R. 2015. Overview of the 2015 Workshop on Speech, Language and Audio in Multimedia. In Proceedings of the 23rd Annual ACM Conference on Multimedia Conference (pp. 1347-1348). ACM.

Gupta, Y., Saini, A., & Saxena, A. K. 2015. A new fuzzy logic based ranking function for efficient Information Retrieval system. Expert Systems with Applications, 42(3), 1223-1234.

Habibian, A., Mensink, T., & Snoek, C. G. 2015. Discovering Semantic Vocabularies for Cross-Media Retrieval. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (pp. 131-138). ACM.

Handschuh, S., Staab, S., & Maedche, A. 2001. CREAM: creating relational metadata with a component-based, ontology-driven annotation framework. In Proceedings of the 1st international conference on Knowledge capture (pp. 76-83). ACM.

Jiang, Yu-Gang, et al. 2013. "High-level event recognition in unconstrained videos."International Journal of Multimedia Information Retrieval 2.2: 73-101.

Kawahara, T., Lee, A., & Shikano, K. 2001. Julius - an open source real-time large vocabulary recognition engine. INTERSPEECH.

Lowin, C.; Raimond, Y.; Tweed, J. 2012. Automated Semantic Tagging of Speech Audio, WWW2012 demos track:Lyon – France.

L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feb 1989.

Ma, H., Ay, S. A., Zimmermann, R., & Kim, S. H. 2014. Large-scale geo-tagged video indexing and queries. GeoInformatica, 18(4), 671-697.

Maynard, D., & Hare, J. 2015. Entity-Based Opinion Mining from Text and Multimedia. In Advances in Social Media Analysis (pp. 65-86). Springer International Publishing.

Meinedo, H. 2008. Audio pre-processing and speech recognition for broadcast news. Universidade Técnica de Lisboa, Diss.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. 2011. DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems (pp. 1-8). ACM.

Mercier, A., & Beigbeder, M. 2005. Fuzzy proximity ranking with boolean queries. InProceedings of the 14th text retrieval conference (TREC 2005) (pp. 433–442).

Oliveira, R., Batista, P., Neto, N., & Klautau, A. 2011. Recursos para desenvolvimento de aplicativos com suporte a reconhecimento de voz para desktop e sistemas embarcados. 12o Fórum Internacional de Software Livre.

Peter Willett, (2006),"The Porter stemming algorithm: then and now", Program, Vol. 40 Iss 3 pp. 219 – 223.

Polyvyanyy, A. 2007. Evaluation of a novel information retrieval model: eTVSM. Master's thesis, Hasso Plattner Institut.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., & Goranov, M. 2003. Towards semantic web information extraction. In Human Language Technologies Workshop at the 2nd International Semantic Web Conference (ISWC2003) (Vol. 20).

Raimond, Y., & Lowis, C. 2012. Automated interlinking of speech radio archives. LDOW, 937 of CEUR Workshop Proceedings, London, UK, 2012.CEUR-WS.org.

RAMALHO, J. C. 2000. Anotação Estrutural de Documentos e sua Semântica, PhD Thesis, Departamento de Informática, Escola de Engenharia, Universidade do Minho, Braga, Portugal.

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

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In KDD, pages 457–466.

Seymore, K., Chen, S., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., ... & Thayer, E. 1998. The 1997 CMU Sphinx-3 English broadcast news transcription system. In DARPA Broadcast News Transcription and Understanding Workshop.

Souza, R. R. 2006. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. Perspectivas em ciência da informação, 11(2), 161-173.

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

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

Vinciarelli, A. (2005). Noisy text categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(12), 1882-1895.

Yang, Y. (1995, July). Noise reduction in a statistical approach to text categorization. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 256-263). ACM.

Yap, K. H., & Wu, K. (2005). A soft relevance framework in content-based image retrieval systems. Circuits and Systems for Video Technology, IEEE Transactions on, 15(12), 1557-1568.

Ynoguti, C. A. (1999). Reconhecimento de fala contínua usando modelos ocultos de Markov (Doctoral dissertation, Universidade Estadual de Campinas).

Young, S. 2008. HMMs and related speech recognition technologies, Springer Handbook of Speech Processing. Springer-Verlag, Berlin Heidelberg, 539–557.