Universidade de São Paulo
Instituto de Matemática e Estatística
Mestrado em Ciência da Computação
Tópicos em Ciência da Computação

Plano de Estudos: Aplicação para a recuperação de vídeos indexados por conceitos

Christian Danniel Paz Trillo
cpaz@ime.usp.br

Orientadora: Dra. Renata Wassermann

7 de Abril de 2004

1 Introdução

A pesquisa a ser desenvolvida durante o semestre, é orientada ao desenvolvimento de uma aplicação de recuperação de informação baseada em consultas em linguagem natural. A recuperação de informação será feita de um banco de dados de vídeos, os quais são fragmentos de uma entrevista com a artista brasileira Ana Teixeira feita e editada por Paula Braga.

O projeto nasceu baseado na idéia de um trabalho do Bruce Bassett, "Conversation with Jacques Lipchitz: A Break-through in Interactivity" [10]. Bassett contava com mais de trezentas horas de entrevista com o artista já falecido Jacques Lipchitz, feitas durante seu trabalho na televisão, com as quais construiu um programa interativo que permite simular uma entrevista com Lipchitz, em que o usuário faz perguntas para ele em linguagem natural pelo teclado, e a cada uma delas, a resposta mais adequada do banco de dados de entrevistas é mostrada em forma de um clip de Lipchitz falando. Este trabalho foi principalmente comercial, e é fechado, o qual não permite conhecer a sua implementação e faz interessante pesquisar como é que um trabalho desse tipo pode ser implementado.

2 Recuperação de Informação

O problema da recuperação de informação se apresenta quando existe uma grande quantidade de informação para a qual é requerido um acesso rápido e esse acesso está se tornando cada vez mais difícil. Isso pode ocasionar que a informação relevante deixe de ser usada só pela dificuldade de acessá-la. Um sistema de recuperação de informação permite procurar pela informação mediante consultas, normalmente feitas por meio de palavras chave que são procuradas dentro do banco de dados contendo a informação [15].

Os sistemas de recuperação de informação baseados em busca por palavra chave são limitados na sua capacidade de distinguir textos relevantes e irrelevantes, isto devido a uma serie de características, entre as quais as mais importantes são [12]:

- Sinonímia.
- Polissemia.

Estes problemas fazem com que os sistemas de recuperação de informação baseados em palavras chaves tenham um limite no seu desempenho conhecido como a barreira da palavra chave[1] [12]. Um dos objetivos da pesquisa é organizar a informação sobre a qual a recuperação de informação vai ser feita para lidar com o problema da sinonímia e algumas formas de polissemia.

3 Organização da informação a ser recuperada

Recuperação de informação tem a ver com representação, armazenamento, organização e acesso à informação [2], a organização dos elementos deve prover aos usuários um acesso fácil e rápido à informação que eles precisam.

Neste projeto a representação, armazenamento e organização da informação sobre o domínio serão feitas em uma *ontologia*. Uma ontologia é definida em [19] como:

> Uma especificação explícita de uma conceitualização, pode tomar uma variedade de formas, mas necessariamente incluirá um vocabulário de termos e alguma especificação do seu significado. Isto inclui definições e uma indicação de como os conceitos são inter-relacionados o qual coletivamente impõe uma estrutura no domínio e restringe as possíveis interpretações dos termos.

A implementação seguirá o formato padrão `OWL` (*Web Ontology Language*) [13, 16]. `OWL` é uma extensão de vocabulário sobre `RDF` (*Resource Definition Framework*) [21],

[1] Do inglês *keyword barrier*

e é uma linguagem padrão feita para escrever ontologias. Foi criada para ser usada quando a informação a ser representada precisa ser processada por aplicações, e não só ser apresentada a humanos. OWL pode ser usado para representar explicitamente os significados de termos em vocabulários e as relações entre eles.

4 O problema a resolver: Converse com a artista Ana Teixeira

Baseado no fato de ter o vídeo de uma entrevista com a artista brasileira Ana Teixeira em formato digital, o problema consiste em permitir a um usuário ver as partes da entrevista que sejam do seu interesse sem ter de navegar por dentro do vídeo inteiro. Uma idéia similar foi apresentada por Bruce Bassett no seu trabalho "Conversation with Jacques Lipchitz: A Break-through in Interactivity" [10], no qual o usuário podia fazer perguntas pelo teclado para recuperar as partes da entrevista do seu interesse.

Neste trabalho, para resolver o problema, será desenvolvida uma aplicação de recuperação de informação, baseada em consultas em linguagem natural, para evitar os problemas ocasionados pela sinonímia. A cada consulta feita para a qual exista informação associada contida na entrevista, o fragmento de vídeo (ou os fragmentos) contendo a informação associada será mostrado ao usuário.

Para isto, o vídeo foi fragmentado e será cadastrado em um banco de dados junto com as informações sobre cada um dos fragmentos, indicando os conceitos aos quais está relacionado cada um deles. Esses conceitos junto com os seus relacionamentos serão cadastrados em uma ontologia. As consultas serão pré-processadas para evitar erros ortográficos [2] e as respostas às consultas serão procuradas dentro do banco de dados levando em consideração os possíveis relacionamentos dos termos usados na consulta segundo a ontologia.

A aplicação para visualização dos vídeos já está implementada em Java. Atualmente cada fragmento de vídeo está relacionado a um conjunto de palavras chave, e a recuperação de informação é feita através de consultas de palavras chave, onde é necessário que a consulta seja feita pela mesma palavra, sem levar em consideração sinonímia, variações de número, gênero ou erros ortográficos.

Além das características já implementadas é preciso implementar:

- A ontologia associada à entrevista com a artista implementado como uma ontologia em OWL. A coleta de informação para a ontologia será feita com ajuda

[2] Conhecido pelo termo em inglês como *Misspelling*.

da especialista do domínio, Paula Braga e se tem alguns repositórios [6, 4] de onde a informação do domínio pode ser extraida. Para a elaboração da ontologia se deve usar um editor de ontologias que gere e leia no formato `OWL` como o `Protegé`[1].

- O pré-processador, para evitar erros ortográficos, usando alguma ferramenta para este fim, como `Jazzy`[7].
- Um analisador morfo-sintático das consultas feitas pelo usuário, para a extrair do objetivo da consulta e das palavras chaves.
- O algoritmo de recuperação do(s) fragmento(s) de vídeo que satisfazem a consulta feita, que deve acessar a ontologia usando alguma biblioteca como `Jena`[11].

O processamento de uma consulta está representado na Figura 1.

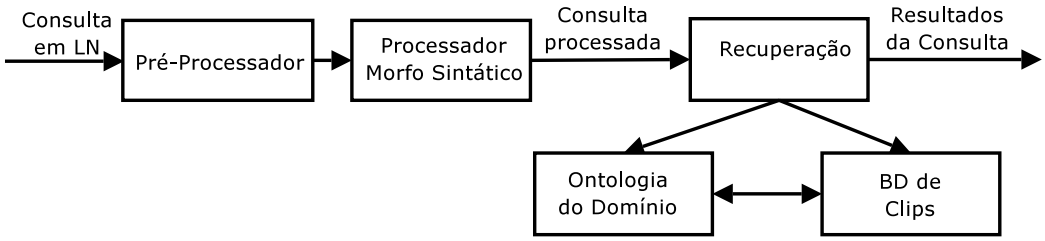

Figura 1: Processamento de uma consulta no sistema.

5 Plano de estudo

O objetivo da pesquisa é conhecer as ferramentas ou bibliotecas que podem ser usadas para implementar o pré-processador, e o analisador morfo-sintático, assim como os editores de ontologias que suportem o formato `OWL` e a própria linguagem `OWL`.

Além disso, é necessário pesquisar as medidas usadas para avaliar os sistemas de recuperação de informação, para fazer análise comparativa do desempenho da implementação atual, baseada em palavras chaves e a implementação a ser desenvolvida.

Referências

[1] Protegé-2000, 2004. `http://protege.stanford.edu`.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman, 1999.

[3] J. Carbonell, D. Harman, E. Hovy, S. Maiorano, J. Prange, and K. Sparck-Jones. Vision statement to guide research in question answering and text summarization. Technical report, Language Thecnologies Institute , Carnegie Mellon University, Pittsburgh, 2000.

[4] Itaú Cultural. Enciclopédia de artes visuais, 2003. Disponível em http://www.itaucultural.org.br/aplicexternas/enciclopedia/artesvisuais2%003/home/index.cfm.

[5] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220, 1993. http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html.

[6] Patricia Harpring. User's guide to the AAT Data Releases. Technical report, Getty Vocabulary Program, 2001. http://www.getty.edu/research/tools/vocabulary.

[7] M. Idzelis and A. Roy. Jazzy - Java Spell Check API, 2003. http://sourceforge.net/projects/jazzy.

[8] Y. Jing and W. Bruce Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO-94, 4th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 146–160, New York, US, 1994. citeseer.nj.nec.com/jing94association.html.

[9] H. Kim and J. Seo. A reliable indexing method for a practical QA system. In *Proceedings of the 19th International Conference on Computational Linguistics*, 2002. O artigo em formato digital pode ser achado em: http://www.isi.edu/~cyl/wsqa-coling2002/program.html.

[10] Greg Kline. High-tech sculptor has the answers. *The News-Gazette Online*, Outubro 2001. Publicado na web no 4 de Outubro de 2001, disponível em: http://www.news-gazette.com/story.cfm?Number=10249.

[11] HP Labs. Jena 2 - A Semantic Web Framework, 2004. http://www.hpl.hp.com/semweb/jena2.htm.

[12] Michael Mauldin. *Conceptual Information Retrieval A case study in adaptive partial parsing*. Kluwer Academic Publishers, 1991.

[13] D. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. Technical report, W3C, 2004. http://www.w3.org/TR/2004/REC-owl-features-20040210/.

[14] National Institute of Standards Technology. Text REtrieval Conference, 2004. http://trec.nist.gov/.

[15] C. J. Van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979. A versão eletrónica do livro pode ser achada em: http://citeseer.ist.psu.edu/vanrijsbergen79information.html.

[16] M. Smith, C. Welthy, and D. McGuiness. OWL Web Ontology Language Guide. Technical report, W3C, 2004. http://www.w3.org/TR/2004/REC-owl-guide-20040210/.

[17] Sun Systems. Java Media Framework API, 2003. http://java.sun.com/products/java-media/jmf/.

[18] Ricardo Ueda. Dicionário br.ispell, 2002. http://www.ime.usp.br/~ueda/br.ispell/.

[19] M. Uschold and R. Jasper. A framework for understanding and classifying ontology applications. In R. Benjamins, B. Chandrasekaran, A. Gomez Perez, N. Guarino, and M. Uschold, editors, *Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods*, pages 11.1–11.12, Sweden, 1999. CEUR publications.

[20] W3C. Semantic Web, 2001. http://www.w3.org/2001/sw/.

[21] W3C. RDF Resource Definition Framework, 2004. http://www.w3.org/RDF/.

[22] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.

[23] Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. VideoQA: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641. ACM Press, 2003.