Tópicos em Ciência da Computação

Relatório de Estudos

Aplicação para a Recuperação de vídeos indexados por conceitos

Christian Danniel Paz Trillo
Orientadora: Dra. Renata Wassermann

São Paulo, Junho de 2004

Resumo

Procurar informação em vídeos compridos pode consumir muito tempo quando não se possui uma ferramenta apropriada para buscar o que se procura dentro dele. O enfoque do estudo apresentado é conhecer as ferramentas para conseguir uma recuperação eficiente de informação em vídeo digital. A aplicação protótipo a ser implementada utilizará uma entrevista com uma artista brasileira. O sistema deve permitir ao usuário entrar com consultas em linguagem natural (português do Brasil) e selecionar com base nelas o(s) pedaço(s) de vídeo que melhor responda a consulta. A recuperação utiliza uma ontologia de arte contemporânea que está sendo desenvolvida para minimizar o impacto dos problemas normalmente encontrados em Sistemas de Recuperação de Informação.

Conteúdo

1 Introdução

O relatório do estudo descreve a pesquisa feita durante o semestre, orientada ao desenvolvimento de uma aplicação de recuperação de vídeos para museus e espaços de exibição de arte. Nessa aplicação, as consultas serão feitas em linguagem natural e o suporte para a recuperação de informação será dado por uma ontologia do domínio. Os vídeos utilizados na aplicação são parte de uma entrevista feita com a artista brasileira Ana Teixeira. Os vídeos foram digitalizados, estão armazenados em formato MPEG, e são a base para o desenvolvimento da ontologia de arte contemporânea, que será aprimorada enquanto mais entrevistas são feitas. Uma parte importante da pesquisa se atenta em tornar o sistema de recuperação flexível de modo que possa ser adaptado a outras entrevistas, ou outros tipos de vídeos (cursos ou seminários, por exemplo) estendendo ou mudando a ontologia e o banco de dados de clips.

Foram revisadas técnicas de recuperação de informação e respectivas medidas de avaliação para as mesmas, sendo apresentadas na Seção 2. Na Seção 3 são mostradas algumas ferramentas existentes para o desenvolvimento e uso de ontologias. A descrição das componentes do sistema é feita na Seção 4. Uma revisão dos trabalhos relacionados à recuperação de vídeos e recuperação de informação baseada em ontologias é apresentada na Seção 5. Finalmente, as conclusões da pesquisa são apresentadas na Seção 6.

2 Recuperação de Informação

2.1 Definição

Um Sistema de Recuperação de Informação (RI) permite aos usuários procurar informação em um banco de dados através de consultas normalmente feitas através de palavras-chave[2]. Baseado na consulta, o sistema de RI recupera a informação que pode ser relevante para o usuário.

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 fazer com que a informação relevante deixe de ser utilizada só pela dificuldade de acessá-la.

2.2 Problemas na Recuperação de Informação

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 série de características, dentre as quais se destacam [23]:

- **Sinonímia**, quando existem diversos termos para descrever um mesmo objeto ou conceito, um sistema de recuperação de informação baseado em semelhança de palavras chave só recuperará aqueles documentos em que o objeto ou conceito esteja descrito com os mesmos termos com que a consulta foi elaborada. Por exemplo, se alguém estiver procurando algum lugar para se hospedar no Rio de Janeiro, poderia colocar em uma consulta "pousada no Rio de Janeiro", e só os documentos em que a palavra "pousada" estiver presente serão recuperados, mas naqueles documentos em que a informação de hospedagem esteja referenciada como "hotel" ou "pensão" por exemplo, não serão recuperados.

- **Polissemia**, uma palavra é polissêmica quando ela tem vários significados, expressando diversos conceitos. Neste caso a recuperação de informação por uma palavra polissêmica pode trazer documentos relacionados não só ao conceito procurado como também aos outros conceitos que são expressados pela mesma palavra. Neste caso por exemplo, a busca da palavra "árvore" no sentido da estrutura de dados pode fazer o sistema recuperar documentos relacionados a "árvore" no sentido do organismo biológico.

Estes problemas fazem com que os sistemas de recuperação de informação baseados em palavras-chave tenham um limite no seu desempenho conhecido como a barreira da palavra-chave[1] [23]. Um dos objetivos da pesquisa é organizar a informação sobre a qual a recuperação de informação vai ser feita para lidar com o problema

[1] Do inglês *keyword barrier*

da sinonímia. Polissemia é um problema que não será tratado pois a visão orientada ao domínio diminui seu impacto no sistema.

2.3 Medidas de Desempenho para Sistemas de RI

Existem algumas medidas padrão para avaliar o desempenho de Sistemas de RI [23]:

- **Precisão:** É a razão entre o número de documentos recuperados pelo sistema que são relevantes para a consulta e o número total de documentos recuperados.

$$P = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos recuperados}} \quad (1)$$

  Por exemplo, se os sistema recuperou 6 documentos para uma consulta, em que 3 de eles foram relevantes, a medida de precisão do sistema é de 0,5 ou 50%. A polissemia pode produzir baixas taxas de precisão no caso de alguns documentos irrelevantes serem recuperados.

- **Cobertura:** Podem existir muitos documentos no banco de dados que o usuário considera relevante, mas só alguns deles serão recuperados pelo sistema. A medida de cobertura é a razão do número de documentos relevantes dividido pelo número total de documentos relevantes no banco de dados.

$$R = \frac{\text{Número de documentos relevantes recuperados}}{\text{Número total de documentos relevantes no banco de dados}} \quad (2)$$

  A cobertura é difícil de medir pois requer o conhecimento do número total de documentos relevantes no banco de dados, que tem que ser determinado manualmente. Para fazer essa medida, normalmente se estabelece um limite superior baseado na seleção de documentos que deveriam ter sido recuperados mas não foram. Quanto maior for o esforço manual para achar esses documentos, mais precisa é a medida de cobertura. A sinonímia pode induzir uma baixa cobertura, pois é possível que não sejam recuperados alguns documentos relevantes que estão relacionados só a sinônimos das palavras utilizadas na consulta.

- **Tempo de Resposta:** É o tempo medido entre a submissão da consulta e a apresentação dos documentos recuperados pelo sistema.

As medidas de precisão e cobertura são independentes e a melhora de uma delas pode implicar uma perda na outra. Um sistema pode ter precisão de 100% se ele devolve só um documento do banco de dados que com certeza é relevante. Uma cobertura de 100% é obtida por um sistema que devolve a qualquer consulta todos os documentos do banco de dados. Portanto se utiliza uma medida que combina as duas, chamada de Medida-F (*F-Measure*)[30].

$$F = 2\frac{C\ P}{C + P} \tag{3}$$

em que a Medida-F é a meia harmônica da precisão e da cobertura. A vantagem do uso da Medida-F é que ela maximiza uma combinação de precisão e de cobertura.

2.4 Técnicas utilizadas em Recuperação de Informação

2.4.1 Indexação

Pode-se dizer que RI basicamente é composta de processos de indexação e de busca. Em geral, a indexação é feita na hora do cadastro dos documentos no banco de dados. A indexação tem que ser feita visando buscas rápidas, portanto usa-se um índice que consiste em uma lista de todas as palavras que ocorrem nos documentos no banco de dados [16]. Esse índice é conhecido como *arquivo invertido* ou dicionário, cada entrada do índice pode conter os seguintes campos (ver Figura 1):

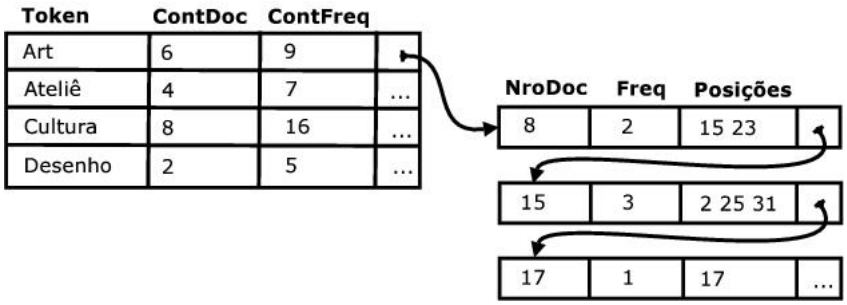

Figura 1: Estrutura de um arquivo invertido.

- ***Token*:** o *token* que representa a palavra, ou o radical da palavra indexada, o processo de obtenção do radical das palavras (remoção de afixos) será explicado posteriormente nesta Seção.
- **Contagem de documentos:** o número de documentos em que o token aparece.
- **Contagem de freqüência total:** o número de aparições do token na coleção. Indica o quão comum é o token na coleção de documentos.
- **Informação por documento:** uma lista de informações associadas às ocorrências do token em cada documento. Cada documento em que o token aparece tem uma entrada nessa lista, contendo as seguintes informações:

  - **Contagem de freqüência:** quantas vezes aparece o token no documento. Este valor indica, a grosso modo, se o documento realmente trata sobre esse conceito ou se o conceito é apenas nomeado.
  - **Posição:** contém uma lista das posições relativas nas que o token aparece no documento.

Essa estrutura é comumente utilizada em Sistemas de RI de texto completo, i.e. que indexam cada documento por todo o conteúdo dele. Quando palavras-chave são manualmente associadas aos documentos, não é preciso ter as contagens de freqüência nem posição, mas um indicador de importância delas permite ordenar por relevância os resultados das consultas.

2.4.2 Processamento de Consultas

Existem diversos tipos de processamento de consultas para a recuperação de informação, como a busca booleana, a recuperação ordenada e a recuperação probabilística que serão descritas a seguir.

- **Busca booleana:** Neste tipo de busca o usuário procura informações no banco de dados através de consultas que conectam palavras mediante operadores lógicos (AND, OR, e NOT). Este é o tipo de busca freqüentemente utilizado nos motores de busca. O desempenho das buscas booleanas pode ser melhorado através da utilização de um *tesauro* para incrementar sinônimos à consulta, em um processo chamado de *expansão de consulta*. Apesar dessas melhoras, as buscas booleanas apresentam alguns problemas como conjuntos de resultados muito grandes, complexidade na formulação de consultas, conjuntos não ordenados de resultados e uma recuperação booleana (o documento é ou não é relevante, não existe um grau de relevância). Este tipo de busca é adequada para usuários profissionais, mas apresenta maior grau de dificuldade para usuários esporádicos ou inexperientes.

- **Recuperação ordenada:** A maioria de buscadores na Web baseiam-se em uma técnica que ordena os resultados da busca com base na distribuição de freqüência dos termos da consulta na coleção de documentos. Por exemplo, um documento que contém várias ocorrências de uma palavra que não é comum na coleção de documentos, provavelmente seja relevante para uma consulta que contenha tal palavra. Assim mesmo, palavras comuns são consideradas com menor importância na ordenação dos resultados.

  Esse tipo de busca é usualmente empregado em interfaces de busca em que se permite aos usuários entrar com consultas em uma linguagem irrestrita, isto é, sem operadores ou uma sintaxe formal. Para permitir isso, os motores removem

as palavras de parada[2] e realizam algumas operações sobre as outras palavras, sendo a mais comum dentre elas a remoção de afixos.

Neste enfoque os documentos e as consultas são tratados como vetores no espaço multi-dimensional definido pelos termos existentes na coleção, e em que a freqüência de aparição dos termos é o valor escalar associado ao vetor na dimensão do termo. Assim, o cálculo da similaridade[3] entre a consulta e cada documento dá um valor de *relevância* do documento para com a consulta. Finalmente, os documentos são ordenados de acordo com a medida de similaridade deles com a consulta.

- **Recuperação probabilística:** Esse enfoque tenta formalizar as idéias por trás da recuperação ordenada em termos da teoria de probabilidades. O fator a ser calculado é a probabilidade de um documento ser relevante para a consulta.

2.4.3 Remoção de Afixos

Freqüentemente se utiliza em Sistemas de Recuperação de Informação, um processo de remoção de afixos (*Stemming*) para identificar as radicais das palavras. Com isso, as variantes de palavras são associadas ao mesmo radical, e a quantidade de armazenamento requerida decresce, pois o número de palavras indexadas diminui. Por exemplo as palavras “cumprido”, “cumpriu” e “cumprirá” são associadas ao mesmo radical “cumpr” que, como no exemplo, não necessariamente é uma palavra válida, mas permite associar todas as palavras derivadas do mesmo radical. Com essa associação é possível fazer com que consultas por palavras derivadas do mesmo radical recuperem os mesmos resultados, o que normalmente é válido pois documentos relacionados com palavras derivadas são relevantes para as mesmas consultas.

Um algoritmo de *Stemming* simples aplica heurísticas para detectar os pontos dos quais os afixos podem ser extraídos. Essas heurísticas obedecem a regras simples, como a detecção de marcas nas palavras, que normalmente são associadas com afixos, por exemplo: “mente”, utilizado na formação de advérbios no português ou “inho” comumente usado para a formação de diminutivos dos substantivos. Essas regras são especificadas para o algoritmo de *Stemming* junto com as exceções delas, por exemplo a palavra “caminho” termina com as letras “inho” mas, nesse caso, “inho” não representa um sufixo, mas é parte da palavra.

A maioria desses algoritmos heurísticos removem só sufixos, pois sufixos mudam a categoria sintática da palavra, mas conservam o sentido da palavra em que são colocados. Um exemplo de algoritmo baseado em regras heurísticas, muito utilizado,

[2]Palavras de parada são palavras muito comuns nos textos da coleção, que os motores de busca evitam porque normalmente não ajudam em refinar a busca e só fazem com que a entrega dos resultados seja mais demorada.

[3]Assumir ortogonalidade das dimensões do vetor, isto é, os termos não são correlatos entre eles, simplifica o cálculo da similaridade que pode ser feito, do modo mais simples, com um produto interno entre os dois vetores.

é o algoritmo de Porter [28] que remove as terminações morfológicas mais comuns das palavras, e foi inicialmente criado para palavras em inglês. Posteriormente, Porter desenvolveu *Snowball*, uma linguagem para uso de cadeias que pode ser utilizado para criar algoritmos de *Stemming*, gerando programas em `Java`, ou `ANSI C`. O algoritmo de *Stemming* para português em *Snowball* está disponível na rede[29], junto com as versões já compiladas em `Java` e `ANSI C`.

Em 2001, foi desenvolvido um algoritmo de *Stemming* para português [27] composto por uma série de regras, aplicadas seqüencialmente. Cada regra especifica o sufixo que será removido, o menor tamanho do radical que pode ficar após a remoção do sufixo, o sufixo substituto do removido e as exceções à regra, sendo esses dois últimos opcionais. Existe uma implementação disponível do algoritmo em `ANSI C` e uma versão dele foi desenvolvida em `Java` como uma componente na nossa aplicação.

3 Organização da informação a ser recuperada

Em Sistemas de Recuperação de informação, a representação, armazenamento, organização e acesso à informação são importantes [2]. A organização dos elementos deve prover aos usuários um acesso fácil e rápido à informação de que eles precisam.

A representação, armazenamento e organização da informação sobre o domínio serão feitas em uma *ontologia*. Uma ontologia é definida em [35] 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.

3.1 Representação da Ontologia

Web Ontology Language [31](OWL) é a linguagem que será utilizada para a representação da ontologia na aplicação. OWL é a recomendação da W3C para descrição de ontologias, que estende o vocabulário de RDF (*Resource Definition Framework*) [8]. Foi criada para ser utilizada quando a informação a ser representada precisa ser processada por aplicações, e não só ser apresentada a personas. OWL pode ser usado para representar explicitamente os significados de termos em vocabulários e as relações entre eles.

OWL permite a representação das construções que serão utilizadas na implementação. A seguir são apresentadas as principais componentes a serem utilizadas na aplicação e as suas correspondentes construções em OWL:

- **Classe**: É um descritor de um conjunto de objetos, abstratos ou do mundo real agrupados por compartilhar algumas propriedades, por exemplo: *Obra de Arte*. Uma classe é representada pela construção *owl:Class*. A classe *owl:Thing* é a que agrupa todos os objetos.
- **Especialização**: Uma classe X é uma especialização de uma outra classe Y se todos os indivíduos que fazem parte de X pertencem a Y também, e se diz que X é uma sub-classe de Y. Com a relação de especialização podem se formar hierarquias de classes. Por exemplo, a classe *Obra Literária* é uma especialização de *Obra de Arte* pois todos os textos considerados obras literárias são obras de arte. A construção *rdf:subClassOf* é utilizada para denotar especialização em OWL. Todas as classes OWL são sub-classes de *owl:Thing*.

- **Relações**: Um conceito normalmente não aparece isolado em seu domínio, sempre aparece relacionado com outros conceitos ou valores. Essas relações são denotadas por um nome, seu domínio e sua imagem. Por exemplo, a relação *é feita por* tem como domínio a classe *Obra de Arte* e como imagem a classe *Artista*. *rdf:Property* é utilizada para expressar relações, mas para especificar que a relação associa dois indivíduos, pode-se utilizar *owl:ObjectProperty*, enquanto *owl:DatatypeProperty* permite associar um indivíduo com um valor, como um inteiro por exemplo.

- **Classes equivalentes**: Duas classes são equivalentes se agrupam os mesmos indivíduos. Esta construção serve para expressar termos alternativos (sinônimos), e em OWL é expressa por *owl:equivalentClass*.

3.2 Uso de ontologias em RI

Ontologias têm sido utilizadas freqüentemente em Sistemas de RI para melhorar as medidas de precisão e cobertura. Dois enfoques podem ser utilizados para obter essas melhoras: expansão de consultas por meio de conceitos semanticamente relacionados e medidas de distância conceitual.

O primeiro enfoque consiste em resolver a consulta do usuário considerando os termos relacionados aos que fazem parte da consulta, assim, alguns documentos que não contenham nenhum dos termos presentes na consulta podem ser recuperados. Nesse enfoque, dois caminhos podem ser considerados. O primeiro, consiste em realizar as inferências em tempo de indexação, isto é, relacionar os documentos no índice não só aos termos diretamente associados, mas também aos inferidos pelas relações de equivalência e de especialização. No outro pode-se realizar as inferências em tempo de consulta, adicionando os termos relacionados à consulta antes de fazê-la na coleção.

O segundo enfoque utiliza a distância conceitual para medir a similaridade entre os termos da consulta e do documento. Uma ontologia pode ser vista como um grafo dirigido, em que os nós são os conceitos na ontologia, e uma aresta dirigida entre dois nós representa uma relação entre os mesmos. A distância conceitual entre dois nós é o comprimento do caminho mais curto entre eles no grafo dirigido.

3.3 Ferramentas

Para o desenvolvimento da ontologia, está sendo utilizado Protégé-2000 [10, 1], um editor de ontologias desenvolvido em Stanford, que apresenta um ambiente extensível para manipulação de ontologias. Protégé provê uma extensão [21] para exportar ontologias em OWL, e permite criar as construções OWL a serem utilizadas na aplicação de maneira visual. A Figura 2 mostra parte do estado atual da ontologia da arte contemporânea sendo desenvolvida, dentro do ambiente do Protégé. A hierarquia apresentada é a que está a baixo da classe *Obra de Arte*. A extensão para OWL de Protégé manipula as ontologias através de Jena[22] que será usado também

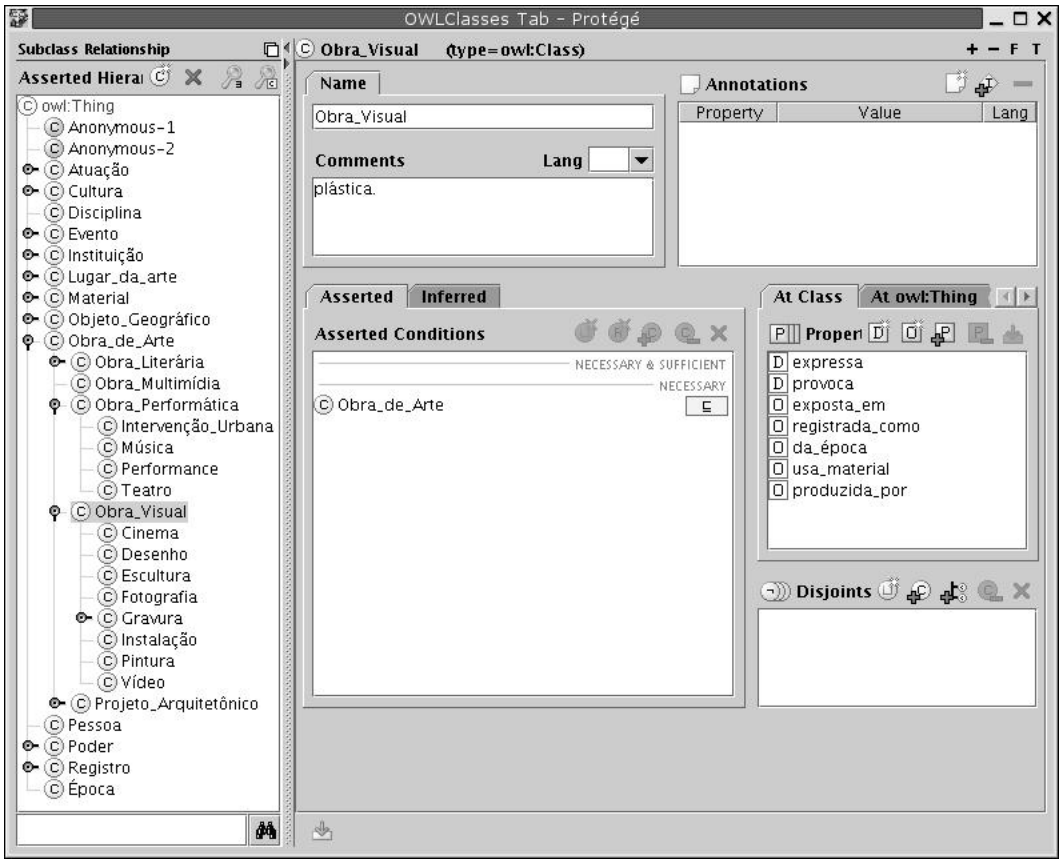

Figura 2: Estado atual da ontologia.

na aplicação para navegar por dentro da ontologia. Jena oferece uma linguagem de consulta, RDQL, para fazer consultas orientadas a dados (similar a SQL) no modelo oferecido pela ontologia.

4 Descrição das componentes do Sistema

Os módulos que estão sendo desenvolvidos para o Sistema são mostrados na Figura 3, e são explicados nas seguintes sub-seções.

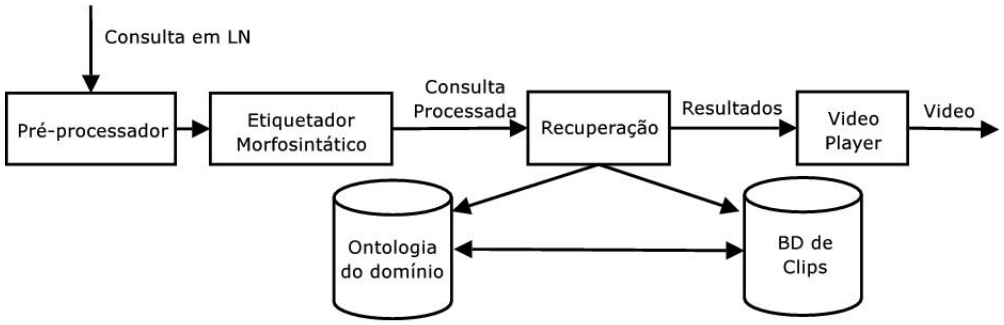

Figura 3: Processamento de uma consulta no sistema.

4.1 Pré-Processador

A consulta feita pelo usuário será processada para detectar e corrigir palavras que possam ter erros ortográficos. A biblioteca a ser utilizada para a correção dos erros ortográficos é o `Jazzy`[15], que é uma ferramenta de software livre para correção de erros, que permite especificar um arquivo de palavras que contém as palavras que ele deve aceitar como válidas. O arquivo de palavras a ser utilizado, é gerado a partir do dicionário eletrônico `br.ispell` [34] e um dicionário orientado ao domínio automaticamente extraído da ontologia para o formato suportado pelo `Jazzy`. O Pré-Processador aplicará a seguir um processo de remoção de afixos para a língua portuguesa detalhado na Seção 2.4.3.

4.2 Etiquetador Morfo-Sintático

A consulta pré-processada será passada para este módulo que identifica as componentes sintáticas da consulta relevantes para a recuperação dos vídeos. Palavras de pergunta, tais como *quem*, *como* serão etiquetadas, de modo com que o objetivo da consulta seja conhecido. Por exemplo, uma pergunta *quem* provavelmente se refere a pessoas. As palavras que são parte do domínio, ou seja extraídas da ontologia, também serão etiquetadas.

4.3 Ontologia do domínio

A ontologia do domínio foi explicada na Seção 3.

4.4 Banco de dados de Clips

Os clips estão armazenados em formato MPEG. Um arquivo XML contém as informações sobre os clips e determina os relacionamentos entre os clips e os conceitos relacionados na ontologia. Os relacionamentos serão processados para gerar um arquivo invertido de índice otimizado para seu uso no processo de recuperação. Além disso, os clips poderiam estar relacionados com recursos (fotografias ou imagens) que são mostrados para o usuário em um tempo especificado dentro do reprodutor de vídeo, usualmente quando o entrevistado faz menção ao recurso.

4.4.1 Processo de recuperação

Será efetuado um processo de recuperação ordenado dos clips, com a finalidade de mostrar os vídeos relacionados à consulta na ordem da relevância que eles têm na mesma. Esse processo de recuperação será melhorado com expansão de consultas e medida de distância conceitual.

4.4.2 Reprodutor de Vídeo

A aplicação mostrará os resultados da consulta, e deve permitir ao usuário visualizar os clips recuperados. O reprodutor de vídeo já foi implementado, e utiliza *Java Media Framework*[32](JMF) para exibir os clips e dar funcionalidade de reproduzir, por pausa e parar o clip, assim como botões de navegação para ver os resultados.

O banco de dados de clips, o processo de recuperação e o reprodutor de vídeo estão implementados atualmente para suportar recuperação de vídeos por palavras-chave em uma aplicação desenvolvida em Java. Esta aplicação, por ser baseada em palavras-chave, não trata com sinonímia. Além disso, ela não faz pré-processamento (remoção de afixos), portanto uma consulta pela palavra *artistas* traz resultados diferentes aos trazidos por uma consulta pela palavra *artista*, por exemplo.

Um objetivo da proposta é comparar os resultados obtidos pela aplicação baseada em palavras-chave e a aplicação baseada na ontologia, para avaliar os benefícios da utilização de ontologias.

5 Trabalhos relacionados

Ontologias têm sido comumente utilizadas em Recuperação de Informação. Em OntoSeek [12] foi proposto o uso de ontologias para incrementar precisão e cobertura em domínios estreitos, como catálogos de produtos, por exemplo. OntoSeek usou uma linguagem limitada para representar conceitos e uma ontologia abrangente, WordNet [25], para casamento de conceitos. Khan propôs, em [18], a aplicação das idéias de distância conceitual em recuperação de dados de áudio utilizando um mecanismo de expansão de consultas que trata com consultas de usuário em linguagem natural no domínio de esportes.

Um trabalho de recuperação automática de vídeo baseada em conteúdo, apresentando um enfoque estatístico para navegação em documentos multimídia, foi apresentado na Universidade de Cambridge[5]. Outros trabalhos utilizam MPEG-7[4] para prover informação mais detalhada sobre o conteúdo dos dados de áudio e vídeo, aproveitando dessa informação para a recuperação [3]. O grupo de pesquisa ISIS da Universidade de Amsterdam propôs um método interativo para recuperação de vídeo, guiando a interação do usuário com informação do domínio extraída da ontologia.

Em Janeiro de 2001, o Museu de Arte Krannert em Champaign, Illinois, USA, apresentou uma exibição no trabalho do escultor Jacques Lipchitz (1891-1973). Além das esculturas, desenhos e pinturas, a exibição apresentou um software desenvolvido por Bruce Bassett e Histor Systems titulado “Conversartion with Jacques Lipchitz: A Breakthrough in Interactivity”. Este sistema foi desenvolvido desde os anos 1970 e, na sua primeira versão, trabalhava com fitas VHS que eram manualmente selecionadas e mostradas dependendo da questão formulada. Na nova versão apresentada no museu esse trabalho pioneiro foi melhorado utilizando fragmentos de vídeo digitalizado. Entretanto, devido ao trabalho ter sido criado como parte de uma iniciativa comercial, a informação disponível sobre a sua implementação é limitada[20, 4].

[4]MPEG-7 é a recomendação para descrição e busca de conteúdo de áudio e vídeo

6 Conclusões

Os diversos problemas da recuperação de informação podem ser tratados para melhorar os resultados obtidos pelas buscas. Melhoras como o uso de ontologias para tratar o problema da sinonímia (considerando os conceitos relacionados), a remoção de afixos para diminuir a complexidade das linguagens de consulta e o tamanho dos índices gerados, devem ser aplicados de maneira conjunta para obter os melhores resultados.

A implementação dos métodos de busca de recuperação ordenada e recuperação probabilística se faz necessária para conhecer o comportamento de ambos modelos de busca quando utilizados com uma ontologia. Da mesma forma o mecanismo de expansão de consulta será implementado para tempo de indexação e tempo de consulta, comparando qual dos dois traz melhores resultados. As medidas padrão de recuperação de informação serão utilizadas para fazer as análises comparativas das diversas configurações a serem implementadas.

A indexação será feita manualmente, mas testes devem ser feitos posteriormente com indexação baseada no texto completo, seja este digitado manualmente ou extraído automaticamente mediante reconhecimento automático de fala. A ontologia nesse sentido pode ser utilizada também para detectar e corrigir erros no texto reconhecido. Essas variantes na indexação devem ser implementadas em uma aplicação de administração que permita gerar aplicações, baseada nos clips e na ontologia do domínio, permitindo cumprir o objetivo de gerar um marco de trabalho flexível.

Referências

[1] Protégé–2000, 2004. http://protege.stanford.edu.

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

[3] W. Bailer, H. Mayer, H. Neuschmied, W. Haas, M. Lux, and W. Klieber. Content-based video retrieval and summarization using MPEG-7. In *Proceedings of the Internet Imaging V*, pages 1–12, San Jose, CA, USA, January 2004.

[4] B. Bassett and Histor Systems. Conversation with Jacques Lipchitz: A breakthrough in interactivity, 2001. A description of the system is available in: http://www.conversationwithjacqueslipchitz.org/.

[5] M. G. Brown, J. T. Foote, Gareth J. F. Jones, K. Sparck-Jones, and S. J. Young. Automatic content-based retrieval of broadcast news. In *Proceedings of the Third ACM Multimedia Conference*, pages 35–43, 1995.

[6] 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.

[7] World Wide Web Consortium. Semantic Web, 2001.

[8] World Wide Web Consortium. RDF Resource Definition Framework, 2004. http://www.w3.org/RDF/.

[9] Itaú Cultural. Enciclopédia de artes visuais, 2003.

[10] J. Gennari, M. Musen, R. Fergerson, W. Grosso, M. Crubézy, H. Eriksson, N. Noy, and S.Tu. The evolution of Protégé–2000: An environment for knowledge–based systems development. *International Journal of Human-Computer Studies*, 58(1):89–123, 2003.

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

[12] N. Guarino, C. Masolo, and G. Vetere. Ontoseek: Content-based access to the web. *IEEE Intelligent Systems*, 14(3):70–80, May 1999.

[13] P. Harpring. User's guide to the AAT Data Releases. Technical report, Getty Vocabulary Program, 2001. This technical report is available in: http://www.getty.edu/research/tools/vocabulary.

[14] E. Hyvönen, A. Styrman, and S. Saarela. Ontology-based image retrieval. In *Towards the semantic web and web services, Proceedings of XML Finland 2002 Conference*, pages 15–27, Finland, 2002. An electronic version can be found in: http://www.cs.helsinki.fi/u/eahyvone/publications/yomuseum.pdf.

[15] M. Idzelis and A. Roy. Jazzy - Java Spell Check API, 2004. http://sourceforge.net/projects/jazzy.

[16] P. Jackson and I. Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction, & Categorization*. John Benjamins Pub Co, 1st edition, 2002.

[17] 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. An electronic version can be found in: citeseer.nj.nec.com/jing94association.html.

[18] L. Khan. *Ontology-based Information Selection*. PhD thesis, Department of Computer Science, University of Southern California, 2000.

[19] 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`.

[20] G. 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`.

[21] H. Knublauch, M. Musen, and A. Rector. Editing description logics ontologies with the Protégé OWL plugin. In *International Workshop on Description Logics*, Whistler, BC, Canada, 2004.

[22] HP Labs. Jena 2.1 - A Semantic Web Framework, 2004. http://www.hpl.hp.com/news/2004/jan-mar/jena2.1.html.

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

[24] D. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview. Technical report, World Wide Web Consortium, 2004. This technical report can be found in: http://www.w3.org/TR/2004/REC-owl-features-20040210/.

[25] G. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[26] National Institute of Standards Technology. Text REtrieval Conference, 2004.

[27] V. Orengo and C. Huyck. A stemming algorithm for the Portuguese language. In *Proceedings of the 8th International Symposium on String Processing and Information Retrieval(SPIRE) 2001*, pages 186–193, 2001.

[28] M. F. Porter. An algorithm for suffix stripping. In *ACM SIGIR Conference on R&D in Information Retrieval*, pages 318–327, 1980. A plain text version of the article and implementations of the proposed algorithm can be found in: http://www.tartarus.org/˜martin/PorterStemmer/.

[29] M. F. Porter. Snowball, 2004. http://snowball.tartarus.org/index.php.

[30] 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`.

[31] M. Smith, C. Welthy, and D. McGuiness. OWL Web Ontology Language Guide. Technical report, World Wide Web Consortium, 2004. This technical report can be found in: http://www.w3.org/TR/2004/REC-owl-guide-20040210/.

[32] Sun Systems. Java Media Framework API, 2004. http://java.sun.com/products/java-media/jmf/.

[33] Y. Tzitzikas. *Collaborative Ontology-based Information Indexing and Retrieval*. PhD thesis, Department of Computer Science, University of Crete, September 2002.

[34] R. Ueda. Ispell Dictionary for Brazilian Portuguese: br.ispell, 2002. http://www.ime.usp.br/ ueda/br.ispell/.

[35] 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.

[36] M. Worring, A. Bagdanov, J. v. Gemert, J-M. Geusebroek, M. Hoang, A.Th. Schreiber, C.G.M. Snoek, J. Vendrig, J. Wielemaker, and A.W.M. Smeulders. Interactive indexing and retrieval of multimedia content. In *Proceedings of the 29th Annual Conference on Current Trends in Theory and Practice of Informatics (SOFSEM), volume 2540 of Lecture Notes in Computer Science*, pages 135–148, Milovy, Czech Republic, 2002. Springer-Verlag.

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

[38] H. Yang, L. Chaisorn, Y. Zhao, S. Neo, and T. Chua. VideoQA: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641. ACM Press, 2003.