Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Seleção de Anúncios para Veiculação Durante a Exibição de Vídeos na *Web*

KARLA SUGUIYAMA OKADA GOMES

Manaus - Amazonas

Fevereiro de 2010

KARLA SUGUIYAMA OKADA GOMES

Seleção de Anúncios para Veiculação Durante a Exibição de Vídeos na *Web*

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Orientador: Prof. Dr. Edleno Silva de Moura.

KARLA SUGUIYAMA OKADA GOMES

Seleção de Anúncios para Veiculação Durante a Exibição de Vídeos na *Web*

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Banca Examinadora

Prof. Dr. Edleno Silva de Moura - Orientador.
Departamento de Ciência da Computação – UFAM

Prof. Dr. João Marcos Bastos Cavalvanti, Ph.D.
Departamento de Ciência da Computação – UFAM

Prof. Dr. Adriano Alonso Veloso.
Departamento de Ciência da Computação – UFMG

Manaus - Amazonas

Fevereiro de 2010

Agradecimentos

Ao meu marido Marcelo pelo incentivo, carinho e compreensão constantes.

Ao meu filho Matteus pela alegria e conforto que me proporciona a cada sorriso.

Aos meus pais que são a base de tudo na minha vida.

Ao meu orientador, prof. Edleno Silva de Moura pela valiosíssima oportunidade, apoio e dedicação.

Aos professores Raimundo Barreto e Virgínia Brilhante.

A David Fernandes e Marco Cristo, cujas contribuições foram essenciais para os resultados deste trabalho.

A Klessius Berlt, Márcia Sampaio e a todos que contribuíram de alguma forma para a realização deste trabalho.

A Deus, por mais uma realização.

Resumo

O mercado de publicidade tem encontrado na *Web* uma das principais mídias para exposição de seus produtos e serviços para um público abrangente a custos relativamente baixos. A principal abordagem de publicidade na *Web* é a propaganda de busca cujos anúncios são selecionados com base nos termos de consultas feitas por usuários em máquinas de busca e são exibidos junto com as suas respostas, uma técnica não intrusiva conhecida como *keyword-targeted advertising* (propaganda direcionada baseada em palavra-chave).

O sucesso deste formato de publicidade, motivou grandes mediadores de informação a disseminá-lo em vários outros contextos, tais como páginas de conteúdo e páginas de serviços, levando ao surgimento da *content-targeted advertising* (propaganda direcionada baseada em conteúdo). O impacto da publicidade na *Web* é ainda maior se considerarmos o aumento expressivo de sua audiência, resultante da proliferação de material gerado pelos próprios usuários finais na chamada *Web* 2.0, tais como, a disseminação de *blogs*, redes sociais e *wikis*. Muitos *sites* têm-se destacado nesse âmbito, atingindo uma grande popularidade e tornando-se fontes promissoras para a publicidade, entre eles, os *sites* de compartilhamento de vídeos, nos quais os usuários podem disponibilizar seus próprios vídeos para outros usuários.

Neste trabalho procurou-se investigar alternativas para a seleção de anúncios a serem veiculados durante a exibição de vídeos postados na *Web*. Diferente de trabalhos anteriores, com o intuito de evitar o alto custo de processamento de imagens, buscou-se explorar metadados textuais relacionados aos vídeos disponibilizados pelos *sites* de compartilhamento destes, através de um estudo preliminar sobre a utilidade dos metadados como fonte de informação a ser usada na seleção de anúncios.

Através de uma coleção de vídeos e uma coleção real de propagandas, os metadados dos vídeos foram utilizados em experimentos com dois métodos de ordenação de propagandas: o vetorial e o vetorial com a aplicação de um modelo de importância de blocos que baseado em dados estatísticos, atribui peso a cada metadado visando estimar a importância da informação carregada pelo mesmo.

Para a avaliação dos resultados dos sistemas de seleção de propagandas estudados, foi criada uma coleção de referência contendo 81 vídeos. Cada vídeo foi assistido e

analisado para a determinação de quais produtos e/ou serviços poderiam ser sugeridos durante a veiculação do mesmo. Baseadas nessas informações, foram selecionadas e associadas manualmente propagandas consideradas relevantes ou não-relevantes para cada vídeo da coleção.

Os resultados experimentais obtidos revelaram que os metadados que discorrem mais sobre o conteúdo do vídeo, como a sua descrição, podem oferecer uma contribuição maior para a seleção de anúncios relevantes a serem mostrados durante à exibição do vídeo. Também pôde-se constatar que a aplicação dos pesos de acordo com o modelo de importância de blocos estudado, levou a resultados com um ganho de cerca de 7% em relação ao método vetorial sem a aplicação de pesos. Aspecto que deve ser considerado importante devido a possibilidade de um aumento da lucratividade do sistema de seleção de propagandas e devido ao impacto negativo que a veiculação de um anúncio não-relevante pode causar nos usuários.

Palavras-chaves: Propaganda Contextual, Modelo de Importância de Blocos, Fontes de Evidências Textuais, Metadados, Vídeos.

Abstract

The Internet has become one of the major media outlets for advertising markets, by exposing its products and services to large audiences at relatively low cost. The main approach of Web advertising is the search advertising whose ads are selected based on the keywords extracted from the user's search queries submitted to search engines and are matched against keywords associated with ads provided by advertisers, known as a non-intrusive technique called keyword-targeted advertising..

The success of keyword-targeted advertising has motivated information gatekeepers to disseminate their ad services over different contexts, such as, content pages and pages of services, leading to the emergence of content-targeted advertising which refers to the issue of matching ads to a web page that is browsed to. The impact of advertising on the Web is even greater if we consider the significant increase of their audience, resulting from the proliferation of the material generated by the users in the so-called Web 2.0, specially with the spread of blogs, social networking sites and wikis. Many websites have been highlighted in this context, achieving great popularity and becoming promising sources for advertising, for instance, the video sharing websites, where users can share digital media.

In this research, we were trying to investigate alternatives for advertisement selection that would run during the display of on-line videos. In order to avoid the high cost of image processing, we were aiming to explore textual metadata related to videos stored on video sharing websites, through a preliminary study on the usefulness of metadata as a source of information used in the selection of on-line advertisement.

While maintaining a video collection and a real ad collection, videos metadata were used in experiments with two ads ranking methods: the vector and the vector with the implementation of a block importance model which based on statistical data, gives a weight to each metadata to estimate the importance of the information carried.

In order to evaluate the output of the studied advertisement selection systems, a reference collection containing 81 videos was created. These videos were carefully analyzed in order to determine which products and/or services they could potentially

advertise. Based on the gathered information, advertisements were manually picked and thus potentially considered either relevant or irrelevant for their appropriate video contained in the collection.

The experimental results obtained showed that the metadata which rather describes video content information, such as its description, potentially offered a greater contribution to the selection of advertisement to be shown during its display. It could also be seen that the application of weights that worked according to the studied block importance model, provided gains of approximately 7% over the vector method that did not use the weights application model. This aspect must be considered important due to the possibility of increasing the profitability of the advertisement selection systems, and given the negative impact of non-relevant advertisement based on credibility and brand of advertisers.

Keywords: Content-targeted Advertising, Block Importance, Sources of Textual Evidences, Metadata, Videos.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O mercado de publicidade tem encontrado na *Web* uma das principais mídias para exposição de seus produtos e serviços para um público abrangente a custos relativamente baixos. Segundo [Giuffrida et al. 2008], estudos apontam que a assimilação imediata do público de um anúncio *on-line* como um *banner* estático é em torno de 40,00% comparado com os 41,00% de um comercial de televisão de 30 segundos.

Considerando a significativa diferença de custo de produção entre os dois tipos de anúncios, a publicidade na Internet apresenta uma relação de custo e benefício muito positiva que tem ocasionado um crescimento em ritmo acelerado desse segmento. Em 2008, isto se traduziu em um total de investimento de 23,4 bilhões de dólares, somente no mercado americano [IAB 2008], o que representou um aumento de 10,60 % em relação a 2007. Mais ainda, este foi o sexto ano consecutivo de uma expansão expressiva tanto em porcentuais quanto em valores monetários, como observado na Figura 1.

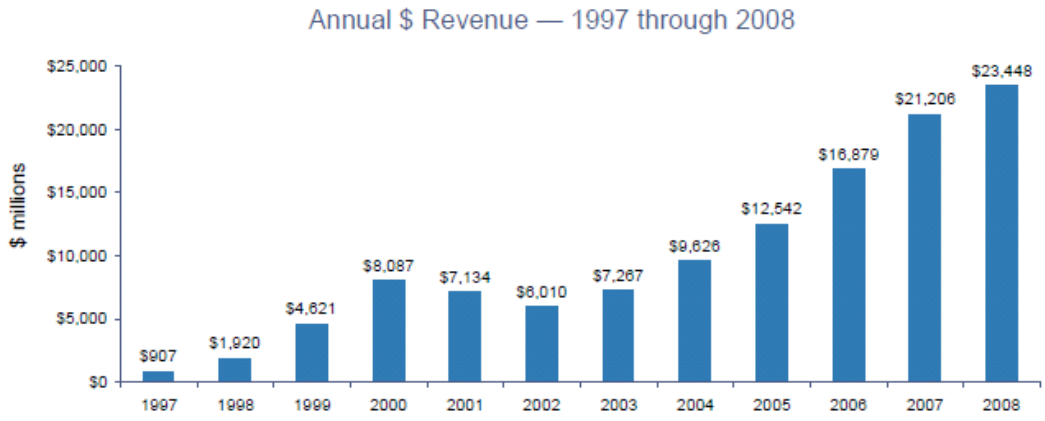

Figura 1: Publicidade *On-line* - Histórico de 1997 até 2008 [IAB 2008].

De fato, se considerarmos os primeiros catorze anos de existência da Internet, observamos que ela apresenta um crescimento, em termos de investimentos em

publicidade, muito mais elevado que outras mídias, como a Televisão aberta e a paga, durante o mesmo período de existência. Isto pode ser observado na Figura 2.

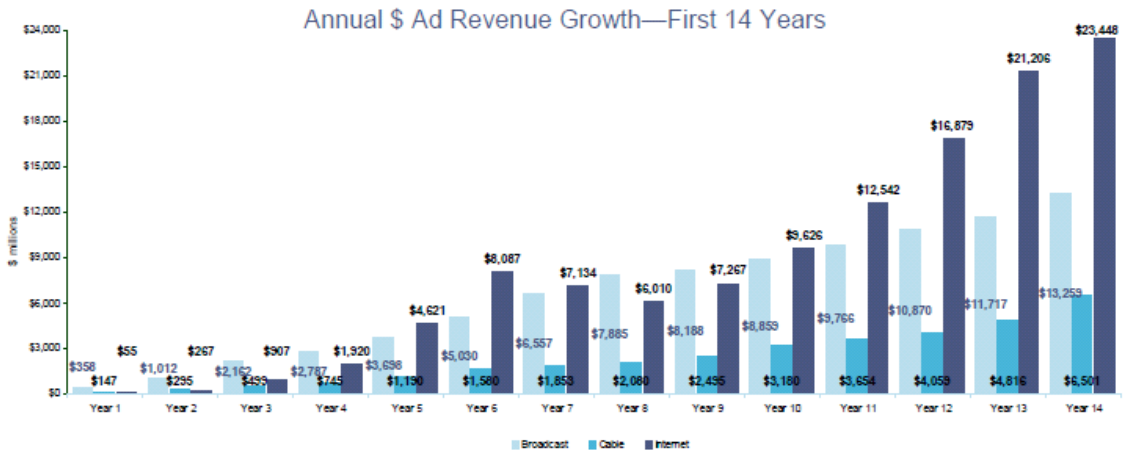

Figura 2: Estudo Comparativo do Crescimento do Investimento em Publicidade - 14 Primeiros Anos, com Inflação ajustada [IAB 2008].

Comparado com outros tipos de mídia, tal crescimento também pode ser associado ao fato dos anúncios *on-line* apresentarem características específicas e bastante interessantes para os anunciantes, tais como:

- Mensurabilidade do retorno: o desempenho de uma campanha publicitária na Internet pode ser mensurado precisamente comparado com outros veículos de comunicação. A relação entre o número de acessos recebidos e o número de exposições da campanha, já permite contabilizar o desempenho dos anúncios. Usualmente, dados são armazenados em detalhados arquivos de texto de *log* e podem ser processados e analisados em diferentes modelos.
- Exposição dinâmica: os anúncios em um *site* podem ser alterados a qualquer momento de forma bastante dinâmica. Basicamente, os editores dos *sites* precisam apenas estabelecer os espaços para os anúncios e os mesmos podem ser dinamicamente preenchidos durante a navegação dos usuários.

- Contextualidade: essa é uma das mais poderosas características dos anúncios *on-line*. Um anúncio pode ser dinamicamente inserido dependendo do contexto sendo exibido.
- Monitoramento do usuário: o comportamento *on-line* do usuário pode ser monitorado de diferentes maneiras. Uma abordagem típica é através do uso de *cookies* armazenados nos computadores dos usuários permitindo que dados acumulados das atividades dos usuários sejam posteriormente analisados.
- Direcionamento *um-para-um*: uma campanha publicitária *on-line* pode ser direcionada para cada usuário específico dependendo de seus interesses e necessidades.
- Volume de dados e disponibilidade: os arquivos de *log* produzidos por *web sites* de médio a grande porte crescem rapidamente chegando a produzir gigabytes de dados em um curto período de tempo. Isto fornece matéria-prima para análise estatística e modelagem complexa dos fenômenos observados em uma rede de publicidade. Em geral, a alta disponibilidade desses *logs* permite o desenvolvimento de modelos em tempo real.
- Teste dinâmico de modelos: a facilidade para modificar configurações de ambientes de teste de modelos é outra característica interessante para estatísticas e mineração de dados. Os modelos podem ser alterados várias vezes ao dia e a reação dos usuários a diferentes modelos pode ser mensurada em tempo real. Desta forma, é possível refinar e adaptar um modelo ainda em desenvolvimento.

Na Tabela 1 abaixo, é feita uma comparação da aplicabilidade das características citadas acima na Internet com outros veículos de comunicação tradicionais, como a televisão e a mídia impressa.

Característica	**Televisão**	**Mídia Impressa**	**Internet**
Mensurabilidade direta			X
Conteúdo dinâmico	X		X
Contextualidade	X	X	X
Precisão de dados			X

Controle dinâmico	X		X
Direcionamento *um-para-um*			X
Grande volume de dados para modelagem			X
Teste dinâmico de modelos			X

Tabela 1: Comparação das Características dos Anúncios em Diferentes Mídias [Giuffrida et al. 2008].

Atualmente, a abordagem de publicidade dominante na *Web* é a propaganda de busca, ou seja, aquela em que anúncios são exibidos junto com respostas fornecidas à consultas feitas pelos usuários em máquinas de busca.

A publicidade de busca em 2008 representou 45,00 % do total da receita de propaganda *on-line* no mercado americano [IAB 2008], maior que os 41,00 % reportados no ano de 2007. O sucesso deste formato de publicidade, levou grandes mediadores de informação como Google e Yahoo, a disseminá-lo em vários outros contextos, tais como páginas de conteúdo, páginas de serviços e vídeos. A principal função de sistemas de propaganda na *Web* é a seleção dos anúncios a serem exibidos em diferentes contextos.

O impacto da publicidade na Internet é ainda maior se considerarmos o aumento expressivo de sua audiência, resultante da proliferação de material gerado pelos próprios usuários na chamada Web 2.0. Entre o material produzido e disponibilizado por usuários finais, temos vários tipos de conteúdo de mídia incluindo notícias, entretenimento, *blogs*, redes sociais e *wikis*.

Diversos *sites* têm-se destacado nesse âmbito, atingindo uma grande popularidade e tornando-se fontes promissoras para a publicidade. Entre eles, pode-se citar o YouTube como um *site* de compartilhamento de vídeos, no qual, os usuários podem fazer *upload* e compartilhar seus próprios vídeos. Segundo Hua et al. [Hua et al. 2008], durante o mês de dezembro de 2007, somente nos Estados Unidos, usuários da Internet assistiram cerca de 10 bilhões de vídeos.

Este trabalho tem como objeto o estudo da publicidade direcionada baseada em conteúdo no contexto de serviços de vídeos digitais na *Web*. Em particular, pretendemos determinar como metadados relacionados a um certo vídeo podem ser explorados para aumentar a precisão de algoritmos de seleção de propagandas.

Diferente de outros trabalhos na literatura, pretendemos utilizar informações relacionadas à estrutura da página que disponibiliza o vídeo como evidência para a criação de melhores funções de seleção de propagandas. Para conseguir tais melhorias, iremos adaptar e aplicar a técnica proposta por Fernandes et al. [Fernandes et al. 2007] afim de descobrir a importância de diferentes blocos de informação da página relacionada ao vídeo e então utilizar tais medidas de importância em funções de seleção de propagandas.

1.1 Trabalhos Relacionados

O crescimento da publicidade *on-line* tem motivado pesquisas sobre os mais diversos desafios de engenharia e modelagem da publicidade de busca. Os sistemas de publicidade *on-line* precisam lidar com grandes volumes de dados e transações que envolvem bilhões de páginas, anúncios e consultas.

Em [Attardi et al. 2004], os autores propõem um projeto de implementação de sistemas de publicidade direcionada em larga escala, baseado em um modelo de filtragem de informação. Em [The Yahoo! Research Team 2006] diversas restrições de engenharia são focadas apontando a eficiência e os custos computacionais como fatores cruciais para a escolha de algoritmos de casamento entre anúncios e páginas *Web*.

Outros aspectos da publicidade de busca também são pesquisados em diversos trabalhos, como o modelo de receitas e valorização das propagandas no processo de ordenação das mesmas em [Feng et al. 2007], a sugestão de termos de busca em [Gleich et al. 2004], a caracterização de tráfego para detectar fraudes [Eneva 2003] e a comparação de estratégias de ordenação [Hemant et al. 2002].

Muitos trabalhos sobre publicidade de busca enfatizam que o fator mais importante para o sucesso dessa área é a relevância dos anúncios que são selecionados para serem exibidos. Em [Ribeiro-Neto et al. 2005] são propostas 10 estratégias de ordenação de propagandas e é feita uma avaliação da eficiência das mesmas para a propaganda baseada em conteúdo.

Em [Lacerda et al. 2006], os autores propõem utilizar aprendizagem de máquina para encontrar boas funções de ordenação para a propaganda contextual. Um algoritmo

de programação genética é aplicado para selecionar a função de ordenação que maximiza a média de precisão em uma coleção de treino. A função de *ranking* resultante é uma combinação não linear de simples componentes como a freqüência de termos dos anúncios na página alvo.

A aprendizagem de máquina também é utilizada em [Ciaramita et al. 2008] para a propaganda contextual utilizando um conjunto de características que visam capturar associações semânticas entre os vocabulários dos anúncios e da página alvo.

No contexto de associação de propagandas a elementos multimídia, pode-se encontrar alguns trabalhos como [Mei et al. 2007] que apresenta o *VideoSense*. O mesmo consiste em um sistema de propaganda para serviços de vídeos *on-line* que associa automaticamente, para cada vídeo, anúncios em formato de vídeo considerados relevantes. Além disso, ele procura inserir tais anúncios em posições apropriadas dentro de cada vídeo de maneira menos intrusiva ao usuário.

Além do *VideoSense*, os mesmos autores propuseram o *ImageSense* [Mei et al. 2008]. Este é um sistema de propaganda contextual direcionado a imagens, cujos anúncios considerados relevantes, são inseridos em imagens em áreas determinadas não intrusivas. Os anúncios relevantes são selecionados com base não somente na relevância textual, mas também na similaridade visual com o conteúdo da imagem.

Basicamente, este trabalho difere dos demais pela análise do impacto de várias características textuais da página em que um vídeo é exibido na *Web* para determinar as propagandas relevantes a serem veiculadas durante a exibição do conteúdo do mesmo, evitando o alto custo de análise de processamento de imagens para tal.

Este trabalho também, utiliza a abordagem baseada em importância de blocos para ordenação de documentos proposta em [Fernandes et al. 2007] que faz uso da informação de estrutura dos documentos para melhorar a função de ordenação de resultados de uma busca em páginas *Web*.

1.2 Contribuições do Trabalho

As principais contribuições deste trabalho podem ser apontadas como sendo, primeiramente, a construção de uma coleção de vídeos com um conjunto de

propagandas consideradas relevantes para cada elemento da coleção. A formação desta coleção será citada posteriormente no Capítulo 3.

O estudo do impacto da aplicação do modelo de importância de blocos proposto em [Fernandes et al. 2007] no contexto da propaganda baseada em conteúdo de vídeos na *Web*, caracteriza outra contribuição deste trabalho. Assim como, o estudo da utilização de metadados de vídeos para determinar a veiculação de anúncios relevantes a serem mostrados durante a exibição do mesmo, excluindo a necessidade de processamento de imagens para tal.

1.3 Organização da Dissertação

Esta dissertação é dividida como segue. No Capítulo 2 são introduzidos os conceitos da propaganda contextual baseada em conteúdo e dos sistemas de publicidade *on-line*. Tais conceitos são necessários para o entendimento deste trabalho. O Capítulo 3 apresenta a formação da coleção de vídeos e propagandas relacionadas e mostra os experimentos realizados com a avaliação dos resultados obtidos. No Capítulo 4 são apresentadas as conclusões e as sugestões de trabalhos futuros que podem ser desenvolvidos a partir dos resultados desta dissertação.

Capítulo 2

Conceitos Básicos

Esse capítulo apresenta uma definição de todo o ambiente da propaganda direcionada baseada em conteúdo, do modelo de Recuperação de Informação aplicado aqui para seleção de propagandas e da métrica utilizada para avaliar a qualidade dos sistemas de seleção de propaganda.

2.1 Propaganda Direcionada Baseada em Conteúdo

A empresa Google foi a pioneira a introduzir o modelo de propaganda direcionada baseada em conteúdo em 2002 [Rappa 2004]. O conceito da técnica não-intrusiva da propaganda direcionada baseada em palavras-chave foi estendido para o conteúdo de páginas *Web*.

No modelo das palavras-chaves, também conhecido como *Sponsered Search* [Broder et al. 2007], os termos utilizados nas consultas dos usuários são relacionados a palavras-chave associadas aos anúncios [Ribeiro-Neto et al. 2005]. Um *ranking* dos anúncios é então computado considerando também, a quantia que o anunciante está disposto a pagar pela inserção de seu anúncio. Os anúncios do topo do *ranking* são exibidos na página dos resultados da busca juntamente com as respostas da consulta do usuário.

Análoga à propaganda direcionada baseada em palavras-chave, a propaganda baseada em conteúdo consiste na seleção dos anúncios a serem exibidos com base no conteúdo da mídia sendo vista, ao invés da consulta do usuário, como demonstrado na Figura 3. Uma vez que os anúncios mais relevantes e lucrativos são conhecidos, os mesmos são exibidos aos usuários agrupados em listas pagas e posicionados na página de exibição da mídia.

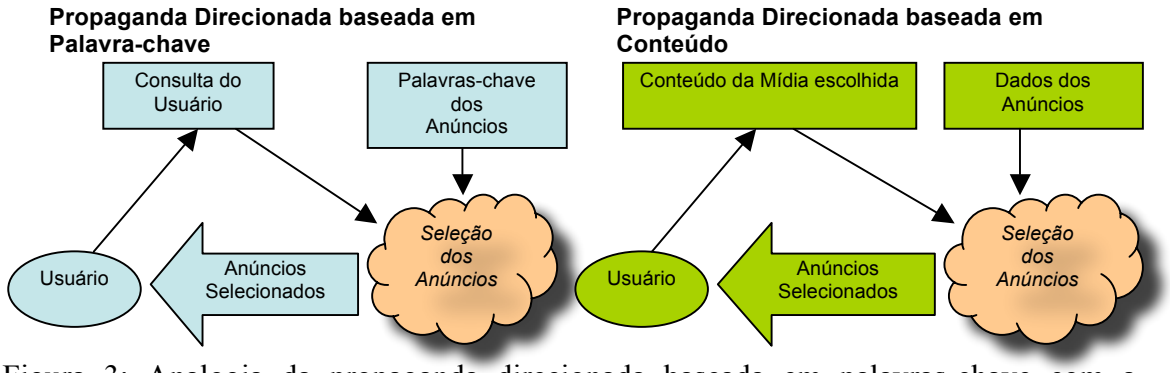

Figura 3: Analogia da propaganda direcionada baseada em palavras-chave com a baseada em conteúdo.

A propaganda baseada em conteúdo tem sido a abordagem contextual dominante de *marketing* na *Web* [Shields 2005]. Os sistemas para seleção de propagandas baseada em conteúdo atuam em um ambiente conhecido como rede de publicidade. Uma rede de publicidade é caracterizada por um padrão de relacionamento no qual todos os atores participantes são beneficiados [Cristo 2006], como ilustrado na Figura 4. Em geral, estas redes são compostas por quatro atores: o provedor do sistema de publicidade (do inglês *broker*), os anunciantes, os divulgadores (do inglês *publisher*) e os usuários.

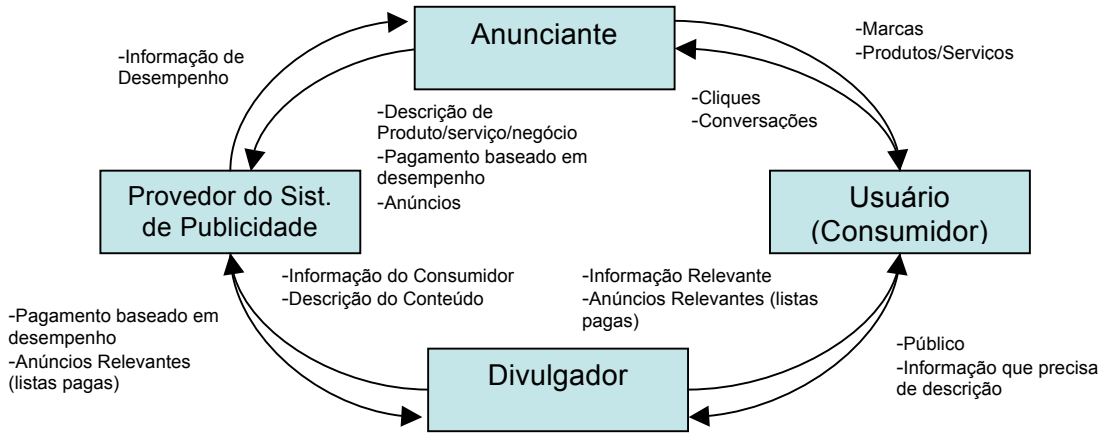

Figura 4: Rede de Publicidade e seus atores [Cristo 2006]

O provedor do sistema de publicidade é responsável pela manutenção da rede. Atua como um mediador entre os anunciantes e os divulgadores, determina quais anunciantes e quais divulgadores podem participar da rede e estabelece as políticas de publicação a serem seguidas. Por exemplo, os provedores não podem permitir conteúdo pornográfico, linguagem imprópria e violação de direitos autorais. Eles também

procuram evitar a participação de empresas que promovem ou lidam com assuntos ilegais, tais como, drogas e jogos de azar.

Os provedores também são responsáveis por fazer um sistema de leilão com a oferta de ferramentas (interfaces, base de dados, vocabulários controlados) que os anunciantes utilizam para descrever seus produtos e serviços. O corretor também é responsável pelos sistemas que serão usados para associar as palavras-chave/conteúdo aos anúncios e pelos sistemas de avaliação que permitem mensurar o desempenho dos divulgadores e anunciantes.

Os anunciantes participam da rede com a expectativa de que eles serão indicados pelos divulgadores, a potenciais usuários. Pelo ponto de vista dos anunciantes, potenciais usuários são aqueles interessados ou que possam se interessar em seus produtos ou serviços. Esse é o caso de muitos usuários que procuram por informações em diretórios e máquinas de busca ou conteúdos editoriais, navegando na *Web*.

Usualmente, as atividades dos anunciantes são organizadas em torno de campanhas que são definidas por um conjunto de anúncios com objetivos temáticos e temporais específicos. Eles pagam ao provedor de acordo com o tráfego fornecido pelos divulgadores e com base nos relatórios de desempenho que eles recebem, é possível ajustar suas campanhas dinamicamente, o que permite maximizar suas receitas e a qualidade dos sistemas.

Os divulgadores são os proprietários das páginas *Web* nas quais os anúncios são exibidos. Os mesmos estão interessados em valorizar suas páginas através da lealdade de seu público. Eles tipicamente, visam maximizar o retorno dos anúncios fornecendo uma experiência agradável aos usuários.

O último ator na rede é o usuário ou o consumidor. Os usuários estão interessados em receber informações relevantes dos divulgadores. Consequentemente, eles naturalmente são segmentados pela descrição de suas necessidades por meios de palavras-chave ou pela navegação nas páginas *Web*, cujos conteúdos são de seu interesse. Ocasionalmente, eles podem clicar em anúncios exibidos, acessar as páginas dos anunciantes e iniciar transações comerciais.

Os sistemas de propaganda *on-line* são uma extensão dos sistemas de *broadcasting*. As propagandas na televisão são sempre criteriosamente escolhidas para a programação com as quais elas são exibidas. Por exemplo, os comerciais de cervejas são sempre

exibidos em partidas de futebol, e comerciais de instituições financeiras são sempre exibidos com a programação de notícias financeiras. As decisões de associação das propagandas com o conteúdo da programação são, portanto, baseadas na experiência e intuição da rede transmissora e da agência de publicidade.

2.2 Sistema de Seleção de Propagandas em Serviços de Vídeo na Web

Um sistema de seleção de propagandas tem como objetivo a apresentação de uma lista de propagandas relacionadas ao conteúdo de uma mídia alvo. Espera-se que as propagandas associadas sejam relevantes para os usuários, adequadas e rentáveis para os anunciantes e divulgadores [Lacerda 2008]. Portanto, os fatores que contribuem para a ordem na qual as propagandas são exibidas, são primeiramente, a relação e adequação das propagandas ao conteúdo da mídia e a quantia que o anunciante está disposto a pagar pelos acessos a suas propagandas.

No contexto dos serviços de vídeo, consideramos que a função principal do sistema de seleção de propagandas baseada em conteúdo, é selecionar os k primeiros anúncios de uma coleção A, de acordo com a relevância em relação ao conteúdo de um vídeo alvo. Tal função pode ser comparada à de um sistema de Recuperação de Informação tradicional, cuja atividade principal é trazer os k primeiros documentos que satisfaçam a uma dada consulta [Cristo 2006]. A Figura 5 ilustra os principais componentes do contexto de propagandas em serviços de vídeos na *Web*.

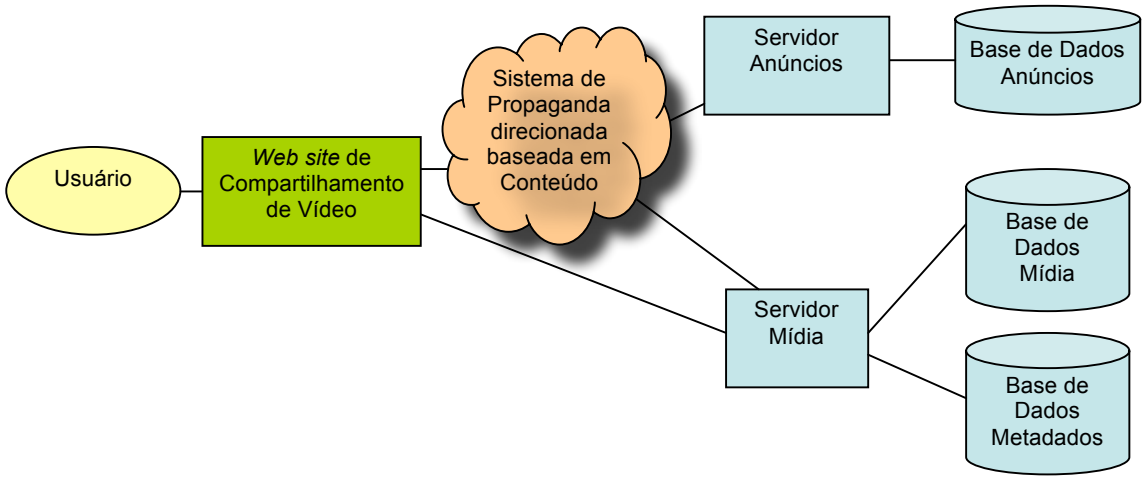

Figura 5: Sistema de Seleção de Propagandas em Serviços de Vídeos na Web.

Neste trabalho, consideramos que uma propaganda é composta de 4 partes estruturais: um título, uma descrição, palavras-chave e um apontador. Sendo estes, os componentes de propagandas comumente encontrados em sistemas comerciais. O apontador direciona o usuário para a página onde a transação pode ser iniciada a qual é chamada de *landing page* [Lacerda 2008]. Nessa página, o usuário pode também encontrar mais informação relacionada à propaganda ou à empresa, seus produtos e serviços.

Além das informações citadas acima, um conjunto de palavras-chave $K = \{k1, k2, \ldots, km\}$ é associado a cada propaganda. As palavras-chave podem ser compostas de uma ou mais palavras e são utilizadas pelos anunciantes para descrever os tópicos que devem existir em mídia alvo, as quais tal propaganda pode ser associada.

Para associar uma dada palavra-chave *k* a uma de suas propagandas, o anunciante precisa fazer uma oferta (um lance) para *k* em um sistema do tipo leilão. Quanto maior a oferta que o anunciante fizer pela palavra-chave *k*, maiores são as chances de que sua propaganda seja mostrada na lista de propagandas associadas a mídias nas quais o tópico *k* esteja presente. Os anunciantes pagam somente por ofertas que forem seguidas pelos usuários. Além disso, um anunciante pode associar várias propagandas ao mesmo produto ou serviço. Tal grupo de propagandas é conhecido como campanha.

A seleção de propagandas baseada em conteúdo no contexto de serviços de vídeos compartilhados por usuários na *Web* é uma atividade ainda não muito estudada na literatura. O foco desta dissertação concentra-se na principal função de um sistema de seleção de propagandas que é gerar o *ranking* das propagandas a serem exibidas. Para realizar tal função, um modelo de Recuperação de Informação que considera a estrutura de documentos foi aplicado e adaptado aos metadados de vídeos.

2.3 Modelo de Recuperação de Informação utilizando Informação de Estrutura

Fernandes et al. [Fernandes et al. 2007] sugerem utilizar a estrutura de blocos das páginas *Web* para melhorar o *ranking* de máquinas de busca na *Web*. Neste modelo, uma página é vista como sendo um conjunto de blocos não sobrepostos, representados por uma tupla *(l,c)*, onde *l* é o rótulo do bloco e *c* é o conteúdo do mesmo. Duas páginas

são consideradas estruturalmente equivalentes se compartilham os mesmos blocos, ou seja, possuem a mesma quantidade de blocos e seus blocos possuem os mesmos rótulos em ambas as páginas.

No contexto desta dissertação, ao invés de termos a estrutura das páginas *Web*, teremos a estrutura dos metadados dos vídeos, utilizando as definições especificadas neste modelo para fazer o *ranking* das propagandas.

O modelo em questão segmenta as páginas *Web* em blocos, definindo o que chamamos de Classe. Uma Classe é um conjunto de blocos $\{b_{l,p}, b_{l,z}\ldots\}$ que pertencem a páginas distintas (estruturalmente equivalentes) e possuem o mesmo rótulo ou nome. No nosso contexto, cada campo dos metadados dos vídeos, será equivalente a uma Classe, como descrito na Tabela 2 abaixo.

Fernandes et al. 2007	**Adaptação do Modelo**
Estrutura das Páginas Web	Estrutura dos metadados dos Vídeos
Classes	Campos

Tabela 2: Mapeamento dos itens do método proposto em [Fernandes et al. 2007].

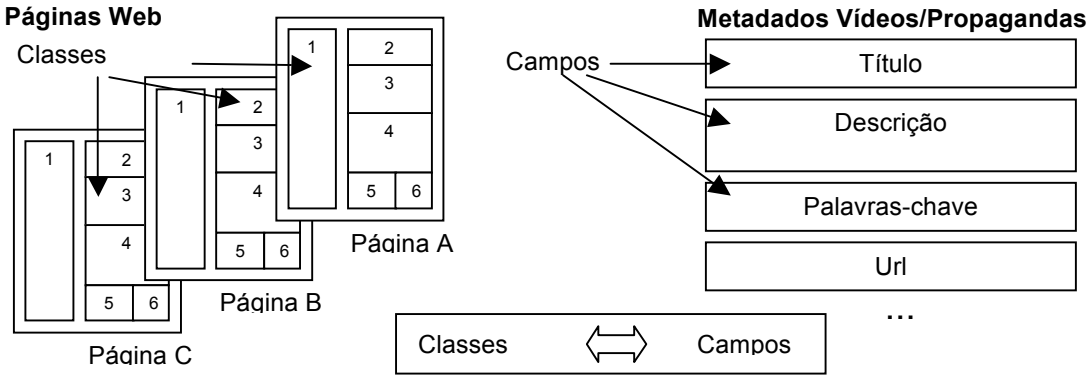

Figura 6: Adaptação dos itens do modelo de RI proposto em [Fernandes et al. 2007] para o contexto de Seleção de Propagandas em Serviços de Vídeos na *Web*.

Os sistemas de recuperação de informação estimam um peso para cada termo em cada documento de uma coleção, criando uma matriz termo-documento. Esta matriz passa a ser usada para calcular similaridades entre consultas e documentos. A maioria dos sistemas de recuperação de informação existentes, assume que todas as ocorrências

de um termo *t* em um documento *d* têm o mesmo valor durante o cálculo do peso de *t* em *d*.

A idéia principal deste modelo, é baseada na hipótese de que o valor de cada ocorrência de um termo pode variar dependendo de sua localização dentro do documento. Por exemplo, em uma página *Web*, a ocorrência de um termo no título pode ser mais importante para a estimativa do peso do termo, do que a ocorrência do mesmo termo no menu desta página.

O modelo propõe melhorias para o cálculo dos pesos de cada termo *t* em um dado documento *d* considerando a localização de cada ocorrência de *t* em *d*. Quando comparado com o método de *ranking* que considera os documentos como unidades monolíticas, este modelo de *ranking* baseado em blocos, obtém melhorias na qualidade dos resultados de busca em *web sites* que possuem estruturas heterogêneas. Adicionalmente, este método não incrementa o custo de processamento de consultas quando comparado com sistemas que não usam informação da estrutura dos documentos.

O modelo, o qual será citado neste trabalho como modelo de importância de blocos, baseia-se em estatísticas sobre a ocorrência de termos em documentos estruturados em uma coleção, para calcular valores de importância de Classes de blocos. Foram adotadas idéias similares às propostas no modelo de espaço vetorial [Baeza-Yates et al. 1999] para calcular o peso da ocorrência de cada termo em cada bloco. Então, esses pesos foram usados para calcular a importância das Classes. Novos conceitos para calcular tal importância, derivados do modelo de espaço vetorial, são introduzidos abaixo.

2.3.1 ICF *(Inverse Class Frequency)*

A *ICF* é uma medida da quantidade de informação que agrega uma ocorrência de um termo *t* na Classe *C*. Essa medida possui a seguinte definição:

Dada uma Classe $C = \{b_1, ..., b_{Bc}\}$ contendo B_C elementos e um termo *t* que ocorre pelo menos em um bloco de *C*, a *ICF* de um termo *t* em *C* é definida como:

$$ICF(t,C) = \log \frac{B_C}{B_{(t,C)}}$$

onde $B_{(t, C)}$ é o número de blocos de C em que t ocorre. Ressaltando que a *ICF* é similar ao conceito de *IDF* (*Inverse Document Frequency*), mas considera cada Classe como uma "coleção de documentos" separada. Como na *IDF*, a intuição por trás dos valores de *ICF* é quantificar a significância da ocorrência de um termo no bloco de uma dada Classe.

2.3.2 ICF Médio da Classe – AICF(C)

O *ICF* médio de uma Classe *C*, *AICF(C)*, é o valor da média de *ICF* de todos os termos que ocorrem em *C* e é dada pela fórmula:

$$AICF(C) = \frac{\sum_{t \in C} ICF(t, C)}{V_C}$$

onde V_C é o tamanho do vocabulário da classe *C*, o número de termos distintos que ocorrem pelo menos uma vez em *C*. Como a medida *IDF*, *ICF* é uma medida da quantidade de informação que agrega uma ocorrência de *t* na Classe *C*. Recebe um valor alto para termos que são raros em uma Classe e baixo para termos que são comuns. Se todos os blocos de uma Classe possuem conteúdo muito similar, a *ICF* dos termos e da Classe será baixa.

Portanto, quando calculamos a *AICF(C)*, obtemos uma medida de quão frequente é o conteúdo de diferentes blocos na Classe. Assim, Classes cujos blocos possuem conteúdo muito repetitivo (*AICF* baixa) são menos importantes e Classes cujos blocos possuem conteúdo mais diversificado (*AICF* alta), são mais importantes e provavelmente serão mais relevantes para identificar o tópico principal da página.

2.3.3 Distribuição Média dos Termos de uma Classe – *Class Spread*

A distribuição média de um termo pela Classe, ou *Class Spread*, é outro conceito introduzido para medir a importância de uma Classe. Esta métrica é baseada na heurística de que blocos que possuem termos em comum com outros blocos tendem a ser relacionados com o tópico principal da página.

Primeiro, tem-se o *numBlockOcurr(t, p)*, que é a quantidade de blocos na página *p* em que o termo *t* ocorre, e *numBlocks(p)*, que é a quantidade de blocos em uma página *p*. Então, a distribuição de um termo *t* em uma página *p*, *termSpread(t, p)*, é dada por:

$$termSpread(t,p)=\frac{numBlockOcurr(t,p)}{numBlocks(p)}$$

Em seguida, pode-se calcular a distribuição média dos termos do bloco b, que contém *numTerms(b)* termos distintos, na página P(b):

$$blockSpread(b)=\sum_{t\in b}\frac{termSpread(t,P(b))}{numTerms(b)}$$

A distribuição média dos termos de um bloco indica o quanto o conteúdo de um bloco é relacionado com o conteúdo dos outros blocos da página e será usado para calcular o grau de distribuição do conteúdo de uma Classe na coleção.

Então, pode-se calcular a distribuição média dos termos de uma Classe, a *classSpread(b),* dada por:

$$classSpread(C)=\sum_{b\in C}\frac{blockSpread(b)}{N_C}$$

sendo N_C o número de blocos de uma classe C.

Para exemplificar o conceito base da *Class Spread*, pode-se citar que a maioria dos termos encontrados no título de uma notícia, normalmente, possui uma alta incidência no texto do corpo da mesma. Desta forma, a distribuição média dos termos de uma Classe que contém títulos de notícias, seria considerada alta.

2.3.4 Importância de uma Classe

Uma vez tendo calculado os valores de *Class Spread* e *AICF* de uma Classe *C*, pode-se calcular a importância da Classe, através do produto dessas duas medidas:

$$class\,\mathrm{Im}\,por\,\tan ce(C) = classSpread(C)\ x\ AICF(C)$$

2.4 Métrica de Avaliação

Segundo Bucley et al. [Bucley et al. 2004] a questão de qual métrica utilizar para avaliar sistemas de recuperação de informação tem recebido muita atenção na literatura. Diferentes métricas de avaliação possuem diferentes propriedades em relação a quão próximas estão dos critérios de satisfação do usuário, quão fáceis são de interpretar, quão significantes são os valores de média e quanto poder possuem para discriminar os resultados obtidos. As métricas de avaliação usualmente mais utilizadas são de alguma maneira, derivadas da precisão e revocação. A precisão é a proporção de documentos retornados que são relevantes e a revocação é a proporção de documentos relevantes que são retornados.

Bucley et al. [Bucley et al. 2004] propõem a métrica de avaliação "bpref" introduzida como uma métrica robusta em relação à informações de relevância incompletas, cuja principal idéia é medir a efetividade de um sistema com base apenas em documentos que foram avaliados. Diferente de métricas como *R-precision*, MAP e P(10) que são completamente determinadas pelos *ranks* de documentos relevantes no conjunto de resultados e não fazem distinção entre documentos que são explicitamente avaliados como não-relevantes e documentos que são assumidos como não-relevantes porque não foram avaliados.

Para um tópico com R documentos relevantes onde *r* é um documento relevante e *n* é um membro dos R primeiros documentos avaliados como não-relevante retornados por um sistema, a fórmula de bpref é dada por:

$$bpref = \frac{1}{R}\sum_{r} 1 - \frac{|\,n\ retornado\ antes\ que\ r\,|}{R}$$

Quando o número de documentos relevantes é muito pequeno, a fórmula acima não

é indicada pelo fato de a avaliação ficar restrita a poucos pares de documentos. Por esta razão, sugere-se a variante da métrica bpref chamada de bpref-10, a qual garante o uso de pelo menos 10 pares de documentos:

$$bpref - 10 = \frac{1}{R} \sum_{r} 1 - \frac{|\, n\ retornado\ antes\ que\ r\,|}{10+R}$$

onde *n* é um membro dos 10+R primeiros documentos avaliados como não-relevante retornados pelo sistema.

Capítulo 3

Experimentos e Discussão dos Resultados

Neste capítulo são relatados os experimentos de ordenação de propagandas baseada no conteúdo de metadados de vídeos juntamente com a apresentação dos seus resultados. Tendo em mãos uma coleção de vídeos e uma coleção de propagandas, experimentos foram realizados em diferentes cenários, aplicando-se o método vetorial, definindo o grupo 1 e aplicando-se o método vetorial com o modelo de importância de blocos, definindo o grupo 2 de experimentos. São expostos os objetivos, configurações e limitações de cada cenário de experimentação, assim como, os critérios de formação das coleções em questão.

Relata-se também, o processo de constituição de uma base de vídeos e propagandas referencial, composta por propagandas consideradas relevantes ou não-relevantes para cada elemento da coleção de vídeos. As propagandas foram avaliadas e consideradas relevantes ou não para veiculação durante a reprodução de um determinado vídeo.

Utilizando a base de vídeos e propagandas associadas citada acima como referência, os resultados obtidos nos experimentos de ordenação de propagandas foram avaliados e seus valores são apresentados, revelando quais condições podem ser vistas como mais favoráveis neste contexto.

3.1 Ambiente de Experimentação

3.1.1 Coleção de Vídeos

Primeiramente, foi formada uma coleção com 81 vídeos coletados a partir do sistema de compartilhamento de vídeos YouTube. Procurou-se obter apenas vídeos cujos metadados ou pelo menos a maior parte do conteúdo deles estivessem descritos no idioma português. Tal restrição foi aplicada devido à base de propagandas também apresentar-se em português, como será visto posteriormente.

As informações sobre cada vídeo incluem título, descrição, palavras-chave, categoria e comentários sobre o mesmo, como demonstrado na Tabela 3 abaixo. Ao compartilhar um vídeo no *site* YouTube, o *Uploader* (usuário dono do vídeo) pode

associar livremente um título, uma descrição e palavras-chave para descrever o conteúdo do vídeo. Ele também precisa associar uma categoria, a qual deve ser escolhida dentre um conjunto de categorias pré-definidas pelo sistema. Os comentários são fornecidos por outros usuários do YouTube e por serem preenchidos de forma colaborativa, neste campo existe uma mistura de idiomas (português, inglês, espanhol). Apenas 68,00% dos vídeos da coleção possuem esta informação.

Os vídeos pertencentes à coleção abrangem os mais diversos temas, tais como, esportes, entretenimento, saúde, relacionamentos afetivos, culinária, gastronomia, turismo, viagens, entre outros e cada vídeo possui em média, uma duração de cerca de 3 minutos e 30 segundos.

Campo do vídeo	Preenchimento	Idioma	Quantidade média de termos distintos
Identificador	100,00%	-	-
Título	100,00%	Português	4
Descrição	100,00%	Português	41
Palavras-chave	100,00%	Português	7
Uploader	100,00%	-	-
Categoria	100,00%	Inglês	1
Comentários	68,00%	Português, Inglês, Espanhol	236

Tabela 3: Descrição dos metadados dos vídeos da coleção utilizada nos experimentos.

3.1.2 Coleção de Propagandas

A coleção é formada por 93.972 propagandas no idioma português. As mesmas foram obtidas a partir de uma coleção real de anúncios e estão agrupadas em 2.029 campanhas diferentes, abrangendo alguns segmentos de produtos e serviços, entre eles, cosméticos, material esportivo, imóveis, eletrônicos, pacotes turísticos, serviços de *buffet* e organização de eventos, consultorias, escolas, faculdades e outros. As propagandas foram fornecidas por 1.744 anunciantes.

As propagandas são compostas por identificador, identificador da campanha a qual pertencem, título, apontador para a página do anunciante referente ao produto em questão, palavras-chave e descrição, como demonstrado na Tabela 4 abaixo. Neste

conjunto de propagandas, os anunciantes associaram em média apenas 1 termo como palavra-chave para cada propaganda e atribuíram uma descrição com cerca de 60 termos distintos. Todos os campos das propagandas estão 100% preenchidos.

Campo da propaganda	Quantidade média de termos distintos
Identificador	-
Identificador da Campanha	-
Título	6
URL	-
Palavras-chave	1
Descrição	60

Tabela 4: Descrição dos metadados das propagandas utilizadas nos experimentos.

3.1.3 Base de Referência (vídeos e propagandas associadas)

A partir da coleção de vídeos e da coleção de propagandas, foi construída uma base de referência formada por propagandas consideradas relevantes para veiculação durante a reprodução de cada vídeo da coleção. Para realizar tal atividade, todos os vídeos foram assistidos através do *site* YouTube com o objetivo de descobrir quais produtos e/ou serviços poderiam sugerir a venda durante a exibição do conteúdo dos vídeos.

Com base em tais associações, foram elaboradas consultas. Para cada vídeo, foram geradas em média 5 consultas com termos relacionados ao conteúdo assistido do mesmo. As consultas foram executadas na coleção de propagandas, aplicando-se o método vetorial.

As propagandas retornadas para cada consulta foram avaliadas como sendo relevantes ou não relevantes, de acordo com a sugestão de venda que o conteúdo do vídeo correspondente poderia induzir. Para cada consulta, em média, foram avaliados 50 resultados. Após a avaliação dos resultados, obteve-se então uma coleção de referência que serve para avaliar sistemas que selecionam propagandas para serem veiculadas para cada vídeo estudado. A Tabela 5 apresenta estatísticas relacionadas à essa coleção formada de vídeos e suas respectivas propagandas consideradas relevantes ou não relevantes.

Quantidade total de vídeos	Média de duração (minutos)	Média de propagandas relevantes por vídeo	Média de propagandas não relevantes por video
81	3,5	112	124

Tabela 5: Descrição geral da base de referência, vídeos e propagandas associadas.

3.2 Experimentos – Métodos de Ordenação de Respostas

Foram realizados dois grupos de experimentos para avaliar formas alternativas de selecionar propagandas para serem veiculadas durante a exibição dos vídeos. No primeiro, aplicou-se o método de ordenação de respostas vetorial e no segundo, o método vetorial com o modelo de importância de blocos (vide capítulo 2).

3.2.1 Grupo 1: Método Vetorial

O conteúdo dos metadados dos vídeos foi utilizado para gerar consultas submetidas à base de propagandas. Para cada vídeo, foram executadas consultas contendo:

- Somente o Título;
- Somente a Descrição;
- Somente Palavras-chave;
- Somente a Categoria;
- Somente os Comentários;
- Todos os campos citados acima;
- Todo o conteúdo da página *Web* do vídeo (excluindo os itens de *Markup language*);
- Combinação dos campos Titulo + Palavras-chave;
- Combinação dos campos Titulo + Palavras-chave + Descrição;
- Combinação dos campos Titulo + Palavras-chave + Descrição + Comentários.

Além dos campos descritos acima, também criou-se um campo contendo informação sobre o *Uploader* (usuário que enviou o vídeo para ser armazenado no *site*). A informação de *Uploader* foi introduzida com a hipótese de que geralmente (não obrigatoriamente) os vídeos compartilhados pelo mesmo *Uploader* tendem a tratar de um único (ou de poucos) domínio de interesse, podendo então revelar informação que ajude na caracterização do conteúdo do vídeo e, conseqüentemente, na seleção de propagandas a serem veiculadas para quem está assistindo o vídeo.

A informação sobre o *Uploader* foi gerada da seguinte maneira: para cada vídeo foram extraídos do sistema YouTube mais vídeos do mesmo *Uploader,* solicitando-se os 50 últimos vídeos postados da mesma categoria do vídeo correspondente e os 50 últimos vídeos postados de todas as categorias existentes misturadas. Possibilitando então, realizar as seguintes consultas ao sistema de seleção de propagandas:

- Vídeos do mesmo *Uploader* e Mesma Categoria:
    - Todos os campos do vídeo original + Titulo dos vídeos postados pelo *Uploader*;
    - Todos os campos do vídeo original + Palavras-chave dos vídeos postados pelo *Uploader*;
    - Todos os campos do vídeo original + Descrição dos vídeos postados pelo *Uploader*;
    - Todos os campos do vídeo original + (Título + Palavras-chave + Descrição dos vídeos postados pelo *Uploader*).

- Vídeos do mesmo *Uploader* e Categorias Misturadas:
    - Todos os campos do vídeo original + Titulo dos vídeos do *Uploader*;
    - Todos os campos do vídeo original + Palavras-chave dos vídeos do *Uploader*;
    - Todos os campos do vídeo original + Descrição dos vídeos do *Uploader*;
    - Todos os campos do vídeo original + (Título + Palavras-chave + Descrição dos vídeos do *Uploader*);

Antes de serem executadas, todas as consultas foram submetidas ao processo de retirada de *stopwords*, palavras com alta freqüência na coleção de documentos não capazes de diferenciar um documento do outro [Baeza-Yates et al. 1999].

3.2.2 Grupo 2: Método Vetorial com o Modelo de Importância de Blocos

Primeiramente, os vídeos foram formatados de maneira que cada campo passou a ser tratado como uma Classe de blocos e o método de importância de blocos foi aplicado, conforme descrito no Capítulo 2, para calcular os pesos de cada campo. Os resultados obtidos são apresentados nas Tabelas 6, 7 e 8 abaixo.

Pesos SPREAD	
Campo do Vídeo	Peso
Titulo	2,176258
Descrição	1,387521
Palavras-chave	1,914037
Categoria	1,048780
Comentários	0,764052

Tabela 6: Método Importância de Blocos. Pesos SPREAD para a coleção de vídeos.

Pesos AICF	
Campo Vídeo	Peso
Titulo	4,303633
Descrição	4,209950
Palavras-chave	4,341201
Categoria	2,920043
Comentários	4,165699

Tabela 7: Método Importância de Blocos. Pesos AICF para a coleção de vídeos.

Pesos SPREAD x AICF	
Campo Vídeo	Peso
Titulo	9,365816
Descrição	5,841394
Palavras-chave	8,309219
Categoria	3,062483
Comentários	3,182811

Tabela 8: Método Importância de Blocos. Pesos SPREAD x AICF para a coleção de vídeos.

Após a determinação dos pesos dos campos, para cada vídeo foram gerados 3 grupos de consultas:

- Com todos os campos ponderados de acordo com os valores SPREAD;
- Com todos os campos ponderados de acordo com os valores AICF;
- Com todos os campos ponderados de acordo com os valores SPREAD x AICF.

Os pesos AICF e SPREAD foram recentemente utilizados na literatura [Figueiredo et al. 2009] como métricas para avaliar a qualidade da informação disponível em um campo, avaliando o poder descritivo e discriminativo de características textuais presentes em algumas aplicações *Web* 2.0.

Segundos os autores, a capacidade de discriminação entre um objeto e os demais objetos da coleção, característica capturada pela métrica AICF; e a acurácia da descrição do conteúdo de um objeto, característica capturada pela métrica SPREAD, são propriedades desejadas em evidências usadas em sistemas de recuperação de informação. Segundo os autores, quanto maior os valores de AICF e SPREAD combinados, maior a utilidade esperada do campo em sistemas de busca.

Como pode ser observado através das Tabelas 6, 7 e 8, os pesos obtidos com a coleção de vídeos em estudo, indicam que os campos mais promissores a serem utilizados pelo sistema de seleção de propagandas, são o título e as palavras-chave, seguidos pela descrição do vídeo. Segundo essas métricas, a atribuição de pesos maiores para esses três metadados deve resultar em melhora na qualidade de um sistema de seleção de anúncios.

3.3 Resultados Experimentais e Avaliação

Para avaliar os resultados obtidos foi utilizada a métrica de bpref-10, por ser apropriada para situações de julgamentos de relevância incompletos, cuja idéia principal é medir a efetividade de um sistema com base somente nos documentos que foram avaliados. A métrica bpref-10 utiliza uma função do número de vezes que documentos avaliados como não-relevantes são retornados antes de documentos avaliados como relevantes (vide Capítulo 2).

Após a execução das consultas, a base de referência foi utilizada para avaliar os resultados obtidos, apresentando os valores expostos nas Tabelas 9, 10, 11 e 12, abaixo.

Grupo 1: Método Vetorial		
Consulta – Campos do Vídeo	Qtde Vídeos	Bpref-10
Campo Categoria	81	0,0120
Campo Título	81	0,0930
Conteúdo da Página do Vídeo	81	0,1379
Campo Palavras-chave	81	0,1568
Campo Comentários	55	0,1690
Campo Descrição	81	0,2051

Tabela 9: Resultados do Método Vetorial.

Através dos resultados apresentados pela Tabela 9 pode-se notar que o campo Categoria por apresentar seu conteúdo no idioma inglês obteve o pior resultado. A consulta com o conteúdo da página do vídeo possui muitos elementos, tais como informações sobre direitos autorais, frases promocionais e outras informações que fizeram piorar a qualidade da ordenação de respostas provida pelo sistema de ordenação de propagandas. Enquanto os campos título, palavras-chave e comentários dos vídeos não obtiveram resultados melhores que a descrição atribuída aos mesmos. Fato que diverge da indicação dos pesos expostos nas Tabelas 6, 7 e 8, cujos valores apontam os campos título e palavras-chave como as melhores evidências para serem utilizadas em sistemas de busca.

Com o intuito de verificar a indicação de qualidade dos campos revelada pelos valores dos pesos, resolveu-se não apenas utilizar os melhores campos indicados isoladamente mas sim, fazer combinações dos mesmos e utilizá-los em novos experimentos, cujos resultados, expostos na Tabela 10 abaixo, apontam a consulta contendo todos os campos dos vídeos como o melhor resultado.

Grupo 1: Método Vetorial – Combinação de Campos		
Consulta – Campos do Vídeo	Qtde Vídeos	Bpref-10
Titulo + Palavras-chave	81	0,1714
Titulo + Palavras-chave + Descrição	81	0,2583
Titulo + Palavras-chave + Descrição + Comentários	81	0,2719
Titulo + Palavras-chave + Descrição + Comentários + Categoria (Todos os campos)	81	0,2732

Tabela 10: Resultados do Método Vetorial com Combinações de Campos.

Nos experimentos acrescentando outros vídeos do mesmo *Uploader* (Tabelas 11 e 12), os resultados obtidos não foram melhores que os obtidos nos experimentos da Tabela 10 utilizando todos os campos do vídeos. Apenas pode-se notar que o filtro por categoria pode ajudar a melhorar ligeiramente os resultados.

Grupo 1: Método Vetorial – *Uploader* – Vídeos da Mesma Categoria		
Consulta – Campos do Vídeo	Qtde Vídeos	Bpref-10
Todos os campos do vídeo original + Título	81	0,1434
Todos os campos do vídeo original + Palavras-chave	81	0,1542
Todos os campos do vídeo original + Descrição	81	0,1460
Todos os campos do vídeo original + (Titulo + Descrição + Palavras-chave)	81	0,1454

Tabela 11: Resultados do Método Vetorial incluindo vídeos do mesmo *Uploader* (mesma categoria).

Grupo 1: Método Vetorial – *Uploader* – Vídeos de Categorias Misturadas		
Consulta – Campos do Vídeo	Qtde Vídeos	Bpref-10
Todos os campos do vídeo original + Título	81	0,1285
Todos os campos do vídeo original + Palavras-chave	81	0,1568
Todos os campos do vídeo original + Descrição	81	0,1118
Todos os campos do vídeo original + (Titulo + Descrição + Palavras-chave)	81	0,1180

Tabela 12: Resultados do Método Vetorial incluindo vídeos do mesmo *Uploader* (categorias misturadas).

Os resultados apresentados na Tabela 13 abaixo, apresentaram valores muito próximos quando se compara os tipos de peso aplicados. Todos tiveram um desempenho ligeiramente melhor que os obtidos nos experimentos das tabelas 9, 10, 11 e 12, ou seja, os experimentos do grupo 2, que utilizam importância de blocos apresentaram resultados superiores.

Grupo 2: Método Vetorial – Importância de Blocos			
Consulta - Vídeo	Qtde Vídeos	Bpref10	Pesos
Todos os Campos	81	0,2907	AICF
Todos os Campos	81	0,2943	SPREAD
Todos os Campos	81	0,2942	SPREAD x AICF

Tabela 13: Resultados do Método Vetorial com o modelo de Importância de Blocos.

O teste estatístico realizado indicou que as diferenças entre todos os métodos baseados em peso e o vetorial é estatisticamente significante com p-value <0.05 em todos os casos. Por outro lado, a diferença entre as três variações utilizando peso não foi significante.

Capítulo 4

Conclusões e Trabalhos Futuros

Neste trabalho procurou-se investigar alternativas para a seleção de anúncios a serem mostrados durante a exibição de vídeos postados na *Web*. O trabalho foi desenvolvido por meio de um estudo de caso que utilizou vídeos coletados do *site* YouTube. Evitando o alto custo de processamento de imagens, buscou-se explorar metadados textuais relacionados a vídeos disponibilizados pelo *site* YouTube.

Para a avaliação dos resultados dos sistemas de seleção de propagandas estudados, foi criada uma coleção de referência contendo 81 vídeos. O conteúdo de cada vídeo foi assistido e analisado para a determinação de quais produtos e/ou serviços poderiam ser sugeridos durante a veiculação do mesmo. Baseadas nessas informações, foram selecionadas e associadas manualmente propagandas consideradas relevantes e não relevantes para cada vídeo da coleção.

Além da montagem da coleção de propagandas, foi realizado um estudo preliminar sobre a utilidade dos metadados como fonte de informação a ser usada na seleção de anúncios a serem veiculados durante a exibição de vídeos. Os metadados dos vídeos foram utilizados em experimentos com dois métodos de ordenação de propagandas: o vetorial e o vetorial com a aplicação do modelo de importância de blocos proposto em [Fernandes et al. 2007], que atribui um peso a cada metadado visando estimar a importância da informação carregada pelo mesmo como fonte de informação.

Com os resultados dos experimentos acima citados chegou-se a conclusão que o método vetorial com o modelo de importância de blocos apresentou um ganho de 7% no desempenho do sistema de ordenação das propagandas em relação ao vetorial sem aplicação de pesos.

Quanto à aplicação dos metadados dos vídeos para a seleção de propagandas, pode-se notar que campos que discorrem mais sobre o conteúdo do vídeo, como a descrição e os comentários (quando existentes), apresentam uma maior contribuição em relação aos campos que apresentam termos ou frases isoladas, como o titulo, palavras-chave e categoria do vídeo. Entretanto, pode-se concluir que a utilização de cada campo separadamente não chega a ser melhor que a utilização de todos os campos estudados juntos e que este resultado pode ser melhorado ponderando-se os campos de acordo com o modelo de importância de blocos aplicado neste trabalho.

Após a realização deste estudo preliminar, como complemento deste trabalho, sugere-se a expansão da coleção de referência, com o acréscimo de mais vídeos e mais avaliações de anúncios relevantes ou não relevantes para a veiculação durante à exibição dos mesmos.

Outra atividade sugerida é a investigação do cálculo e aplicação de pesos também na base de propagandas e a análise do impacto no retorno do método vetorial para o sistema de seleção de propagandas.

Sugere-se também, a exploração de um recurso não utilizado neste trabalho que é a informação sobre vídeos relacionados, disponível na página *Web* de cada vídeo a ser estudado, para determinar o domínio de interesse do vídeo em questão. Seguindo a mesma linha de estudo, sugere-se buscar outras formas de aplicação das informações relacionadas ao *Uploader* do vídeo, além das utilizadas neste trabalho.

Referências Bibliográficas

Attardi, G., Esuli, A., Simi, M., "Best Bets, Thousands of Queries in Search of a Client," *Proceedings of the 13th International Conference on World Wide Web*, Alternate Track Papers and Posters, ACM Press, 2004.

Baeza-Yates, R., Ribeiro-Neto, B., *Modern Information Retrieval*, New York, ACM Press, 1999.

Broder, A., Fontoura, M., Josifovski, V., Riedel, L., A semantic approach to contextual advertising. In SIGIR'07. ACM Press, 2007.

Buckley, C., and E. M. Voorhees. Retrieval evaluation with incomplete information. SIGIR'04, (27), 2004.

Ciaramita, M., Murdock, V., Plachouras, V., Semantic associations for contextual advertising. IJECR 9(1), 2008.

Cristo, M.; Ribeiro-Neto, B., Sobre Publicidade Direcionada Baseada em Conteúdo. Tese de Doutorado em Ciências da Computação. Universidade Federal de Minas Gerais. 2006.

Eneva, E., Detecting invalid clicks in online paid search listings: a problem description for the use of unlabeled data. In Tom Fawcett and Nina Mishra, editors, Workshop on the Continuum from Labeled to Unlabeled Data, 20th International Conference on Machine Learning, Washington DC, USA, August 2003. AAAI Press.

Feng, J., Bhargava, H., Pennock, D., "Implementing Sponsored Search in Web Search Engines: Computational Evaluation of Alternative Mechanisms," *Informs Journal on Computing*, Vol. 19, No 1:134-148, 2007.

Figueiredo, F., Belém, F., Pinto, H., Almeida, J., Gonçalves, M. A., Fernandes, D., Moura, E., Cristo, M., Evidence of quality of textual features on the web 2.0. In: Conference on Information and Knowledge Management (CIKM), 2009, Hong Kong. Proceeding of the 18th ACM conference on Information and knowledge management, 2009. v. 1. p. 909-918.

Aun, F., Two Large Ad Networks Embrace Behavioral Targeting. ClickZ Experts, July of 2008. http://www.clickz.com/3630287

Fernandes, D., de Moura, E. S., Ribeiro-Neto, B., da Silva, A. S., and Gonçalves, M. A. (2007). Computing block importance for searching on web sites. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 165–174, New York, NY, USA. ACM.

Gleich D., and Zhukov, L., SVD based term suggestion and ranking system. In Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04), pages 391–394, Washington, DC, USA, 2004. IEEE Computer Society.

Giuffrida, G., Cantone, V., and Tribulato, G., An apriori based approach to improve on-line advertising performance. In C. Soares, Y. Peng, J. Meng, Z.-H. Zhou, and T. Washio, editors, Applications of Data Mining in E-Business and Finance, pages 53–63. IOS Press, 2008.

Bhargava, H., Feng, J., Paid placement strategies for internet search engines. In Proceedings of the 11th international conference on World Wide Web, pages 117_123, New York, NY, USA, 2002.

Hua, X.-S., Mei, T., and Li, S., "When multimedia advertising meets the new internet era," in Proceedings of IEEE International Workshop on Multimedia Signal Processing, 2008, pp.1–5.

IAB; PRICE WATER HOUSE COOPERS. IAB internet advertising revenue report. 2008.

Krol, C., Zeroing in on content-targeted ads. BtoB Online, February 2005. Available at http://www.btobonline.com/article.cms?articleId=23413.

Lacerda, A., M. Cristo, M.A. Goncalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto, "Learning to Advertise," *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, pp. 549-556, 2006.

Lacerda, A., Ziviani, N. Uso de Programação Genética Para Propaganda Direcionada Baseada em Conteúdo. Dissertação de Mestrado Ciências da Computação. Universidade Federal de Minas Gerais. 2008.

Mei, T., Hua, X.-S., Yang, L. and Li, S., "VideoSense: Towards effective online video advertising," in Proceedings of ACM Multimedia, 2007, pp. 1075–1084.

Mei, T., Hua, X.-S., Li, S., "Contextual in-image advertising," in Proceedings of CM Multimedia, 2008, pp. 439–448.

Rappa, M. "The utility business model and the future of computing services." IBM Systems Journal 43(1): 32-43, 2004.

Ribeiro-Neto, B., Cristo, M., de Moura, E., and Golgher, P., Impedance coupling in content-target advertising. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 496_500, Salvador, Bahia, Brazil, July 2005.

Shields, M., Online publishers foresee dynamic ad spending. Adweek, February 2005. Available at http://www.adweek.com/aw/search/article_display.jsp?schema=&vnu_content_id=1000797161.

The Yahoo! Research Team, "Content, Metadata, and Behavioral Information: Directions for Yahoo! Research," *IEEE Data Engineering Bulletin*, December 2006.