

3

Classes Funcionais

Em todo trabalho de classificação funcional é necessário determinar quais serão as classes funcionais utilizadas. Esta divisão não se propõe a ser extensiva, ou seja, não pretende cobrir todo o conteúdo da Web. No entanto, as classes selecionadas neste trabalho são representativas de importantes canais de comunicação, e apresentam um contraste interessante entre tecnologias novas e tradicionais.

Ao mesmo tempo, embora as classes não sejam extensivas, a estratégia de coleta de informações, de extração de atributos e até mesmo de classificação adotadas podem facilmente ser adaptadas e expandidas para a inclusão de novas classes ou para utilização de classes completamente distintas. Em especial, os conjuntos de atributos selecionados e os métodos utilizados são de fácil reutilização.

A seguir são apresentadas cada uma das 4 classes selecionadas. Em seguida, é realizada uma breve discussão sobre o processo de coleta de informações, as dificuldades encontradas durante o mesmo e pontos de consideração para coletas futuras.

3.1.

Blogs

A primeira classe funcional é de **Blogs**. A palavra blog é uma contração do termo “*web log*”, que, traduzido ao pé da letra, significa diário na Web. Um blog é uma página que tem o seu conteúdo alterado freqüentemente através de pequenos textos, denominados *posts*. Embora alguns sejam focados em assuntos específicos, muitos funcionam como espaços de comentários sobre notícias e acontecimentos recentes, ou mesmo como diários de seus escritores. Portanto, pode não existir nenhuma correlação entre o conteúdo de diferentes posts num mesmo blog.

A maioria dos blogs utiliza texto como seu principal mecanismo de comunicação, mas existem também os que incorporam vídeos, fotos, música, e outros elementos de mídia. Existem ainda os que são exclusivamente compostos de outros elementos de mídia, em detrimento de texto, como os *fotoblogs* (apenas com fotos), ou *video blogs* (apenas por vídeos).

No geral, um blog é composto por uma seção principal, onde os posts mais recentes são apresentados em ordem cronológica reversa (dos mais novos para os mais antigos), uma seção denominada “*blogroll*”, onde o autor lista outros blogs que acha interessante, e uma seção de arquivos, onde posts antigos podem ser encontrados.

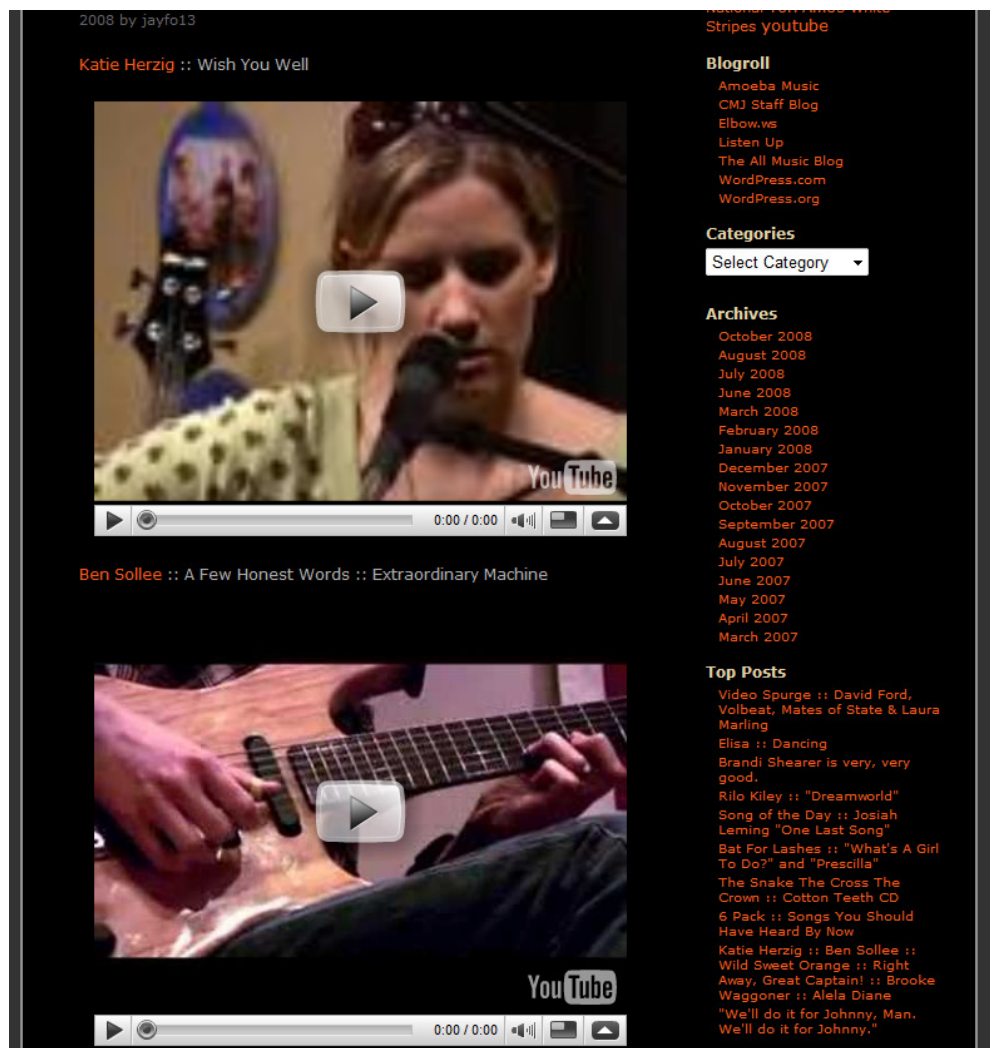


Figura 1 – Imagem retirada do blog portal “Mister Music Blog”

(<http://mistermusicblog.wordpress.com/>). Este é um exemplo de um blog com conteúdo

misto, consistindo principalmente de vídeos de música, composto e publicado através de uma ferramenta conhecida de blogs (wordpress). Ele possui também a maioria das seções tradicionais, como "Blogroll" e arquivos (posts antigos).

BLOGS. NYPOST.COM

RTS | ENTERTAINMENT | BUSINESS | OPINION | HEALTH | NY

November 2008
Su M T W Th F S
1
2 3 4 5 6 7 8
9 10 11 12 13 14 15
16 17 18 19 20 21 22
23 24 25 26 27 28 29
30

Tuesday, November 11, 2008
Last Update: 04:35 PM EST

NYC Weather
48° CLEAR

SEARCH THIS BLOG: **GO**

Advertisement

Rent Movies From Netflix

Only **\$4.99/mo.**
No late fees
Over 100,000 titles

Click Here

ARCHIVES LAST 10

- Oscar Watch: Mickey Rourke Benches Himself
- Oscar Watch: 'Australia' Gets Oprah's Vote
- DVD Extra: Big Day For Eddie Cantor Fans
- DVD Extra: 'This is the Army' Finally Gets Its Due
- Sundance Responds to Boycott Calls
- Oscar Watch: Storm Clouds Over 'Australia'
- Oscar Watch: Writing 'Changeling'
- Box Office: Bernie Mac's DOA as 'Madagascar' Scores and McLovin Outdraws Rogen
- DVD Extra: Chase Talks Sopranos Ending, Movie Version

LUMENICK ON FILM
CHIEF FILM CRITIC LOU LUMENICK TAKES YOU INSIDE THE MOVIE INDUSTRY, TELLING YOU WHAT NOT TO SEE AND WHAT NOT TO MISS. HE WATCHES THE GIANT CROCODILE MOVIES SO YOU DON'T HAVE TO.

November 11, 2008

Oscar Watch: Mickey Rourke Benches Himself

Mickey Rourke has been considered a lock for a Best Actor nomination for his comeback performance in "The Wrestler," but that didn't stop the famously self-destructive actor from effectively taking himself out of contention. In a case of spectacularly bad timing, Rourke used an anti-gay hate word to refer to a writer – just days after the passage of Prop 8. Worse, this [article](#) about his subsequent apology contains a link to a video of the offending incident that is no doubt finding an interested audience among academy members. Tip O'Hat: [Lt. Smith \(Ret\)](#)

Posted by Lou Lumenick at 3:30 PM | [Comments \(0\)](#) [Buzz up!](#)

Oscar Watch: 'Australia' Gets Oprah's Vote

Even before Nicole Kidman and Hugh Jackman started playing her with Australian cookies and candy, Oprah Winfrey was hyperventilating about "Australia," which could use some good buzz after [word filtered out](#) from Down Under that director Baz Luhrmann has bowed to studio pressure for a "happy" ending, not generally a good omen for the Oscars. "This is the movie to see over Thanksgiving," [proclaimed](#) Oprah on her show yesterday when she wasn't nibbling on goodies or asking the frozen-faced Oscar winner

Figura 2 – Imagem retirada do blog portal de filmes do jornal NY Post

(<http://blogs.nypost.com/movies/>). Este é um exemplo de um blog corporativo, que segue a estrutura visual do jornal ao qual é afiliado. Ele apresenta também algumas das seções tradicionais de blogs, como arquivos e melhores posts, mas não todas (ausência de blogroll, por exemplo).

A natureza temporal da informação contida dentro dos blogs os diferencia de páginas tradicionais da web. Sua importância, em termos de ranqueamento, está relacionada com o quão recentes são suas informações. Se o autor é sempre um dos primeiros a trazer novos temas, e esses se espalham rapidamente, então o blog é importante, mesmo que só possua um ou dois links externos [17].

3.1.1. Classificação de Blogs

A classificação temática é fundamental para os serviços de busca de blogs, uma vez que a maioria dos usuários realizam consultas buscando assuntos, e não nomes de blogs ou autores específicos. Ao mesmo tempo, a classificação funcional é necessária para a indexação desses serviços, já que apenas blogs devem ser indexados e rankeados, e demais páginas devem ser ignoradas.

De um ponto de vista estrutural, blogs também apresentam um desafio complexo. Com ferramentas como Wordpress e Blogger, qualquer usuário pode publicar seu conteúdo na rede sem nenhum conhecimento técnico sobre HTML ou tecnologias relacionadas. Assim, as únicas informações presentes dentro de tags HTML especiais são as colocadas lá pelas ferramentas de publicação de conteúdo.

3.1.2. Classificação Funcional e Temática de Blogs

A grande variedade de blogs existentes afeta diretamente o trabalho de classificação, seja ela funcional ou temática. No caso de classificação funcional, a decisão a ser tomada é se todos os tipos diferentes de blogs devem ou não ser agrupados em uma única classe funcional. No caso de classificação temática, a variação de estilos de escrita dificulta a classificação do conteúdo abordado.

Já o estudo apresentado em [7] define uma série de classes funcionais para os blogs, o que é interessante para tarefas de classificação funcionais. Temos os agregadores e filtros de notícias, nos quais os autores passam o dia lendo notícias e repassam as que consideram mais interessantes; existem os *blogs* relacionados com “novidades,” onde os autores estão sempre buscando assuntos novos e interessantes para repassar aos leitores; existem os que funcionam como diários *on-line*, onde os autores simplesmente expõe seus pensamentos e seu estado de espírito; os que funcionam como vitrine para escritores e poetas, onde eles expõem seus textos para uma ampla audiência; e dezenas de outros tipos mais.

Pode-se traçar ainda classes funcionais com base nas distinções estruturais e de formato dos diferentes blogs que surgiram ao longo do tempo. Temos blogs com fotos (os *fotoblogs*), blogs publicados através de dispositivos móveis

(*moblogs*), blogs escritos colaborativamente por diversos autores, e, com o surgimento de sites como YouTube, blogs inteiramente videofilmados.

A classificação temática de blogs é especialmente complexa devido à diversidade de conteúdo presente em cada um. De fato, tentar classificar um blog como um todo de acordo com um determinado tema é praticamente impossível, uma vez que cada post pode tratar de um assunto completamente diferente. Por outro lado, a classificação de posts individuais é mais simples, uma vez que geralmente cada post trata de um único assunto. Assim, a única maneira de se realizar esse tipo de classificação em um blog é se classificar os posts individuais e depois realizar uma seleção sobre os temas mais frequentes (como tecnologia, esportes, etc.).

Outra dificuldade inerente ao trabalho de classificação de conteúdo dentro de blogs são as características temporais das informações contidas dentro de um blog [6]. Quando um usuário realiza uma busca dentro de um blog, ele não está apenas preocupado em conseguir informação sobre algum assunto, mas também em conseguir informação *dentro de certo intervalo de tempo*. O usuário pode estar interessado nas informações mais recentes sobre um determinado assunto, ou sobre informações históricas de outro assunto. De uma maneira ou de outra, existe uma componente temporal fundamental nas buscas dentro dos blogs, que dificilmente existe em buscas na web.

3.1.3. Splogs

Um problema relacionado com a classificação de blogs é a detecção de *splogs* – blogs que correspondem a *spam*, páginas que não interessam a nenhum leitor e alteram os rankings de máquinas de busca artificialmente. Segundo [8], 75% dos *pings* recebidos por serviços que monitoram atualizações de blogs e até 20% dos resultados retornados por máquinas de busca especializadas correspondem a splogs.

Devido ao tamanho do problema, a detecção de splogs foi explorada a fundo em [4], [5] e [8]. Nestes estudos, o problema é formalizado, e técnicas de detecção são elaboradas e testadas em ambientes reais, apresentando ótimos resultados. Em [5], especificamente, os autores aplicam técnicas similares ao problema de

classificação binária de páginas em “blogs” e “não-blogs”. Utilizando atributos de conteúdo das páginas, os autores construíram modelos baseados em SVMs para a classificação de páginas web, chegando a 98% de acurácia sobre alguns conjuntos de treino.

A classificação de splogs é essencialmente uma classificação funcional. Nos estudos mencionados acima, os autores utilizam abordagens híbridas no que diz respeito tanto aos atributos utilizados quanto a classificação por contexto/conteúdo.

3.1.4. Classificação Binária de Blogs

A classificação binária de blogs consiste em classificar uma página qualquer da web em duas classes funcionais distintas: “blog” ou “não-blog.” Qualquer tipo de blog deve pertencer à classe “blog,” seja um fotolog, um splog, ou qualquer outro. Ao mesmo tempo, todas as outras páginas devem estar na classe “não-blog.”

Para esse trabalho, já foram exploradas até hoje duas abordagens. A primeira, e mais simples, é uma abordagem baseada em heurísticas simples. Embora elas apresentem bons resultados quando são criadas, tendem a perder sua eficácia rapidamente. Em [5], os autores mostram que a utilização dos tags HTML que indicam a presença de *feeds* RSS ou ATOM resultava em acurácia de 70% em 2006, mas essa acurácia vem decaindo rapidamente, conforme a tecnologia passa a ser adotada amplamente.

A outra abordagem, adotada em [5] e também em [9], consiste na utilização do conteúdo dos documentos, através de pré-processamento, para a classificação do documento em “blog” e “não-blog”. Embora ambos os métodos apresentem excelentes resultados, uma abordagem baseada em características estruturais é desejável, pois evita uma série de problemas que estão presentes na classificação por conteúdo, como a dependência lingüística e a necessidade de recomposição do conjunto de atributos a cada vez que o conjunto de treino se altera.

3.2. Blog Posts

A segunda classe considerada são os ***Blog Posts***. Um blog, conforme descrito acima pode conter o texto de diversos blog posts em seu corpo, mas estes também são disponibilizados na forma de páginas separadas, que estão sendo denominadas neste trabalho de blog posts.

Estas páginas consistem de pequenos textos acompanhados de comentários. Alguns podem conter grandes quantidades de texto e um número pequeno de comentários, enquanto outros podem possuir vários comentários mas um texto curto.

Como este trabalho trata de páginas individuais, e não de sites como um todo, blog posts podem ser diferenciados dos blogs aos quais pertencem. Além de existirem como páginas separadas do blog principal, os blog posts na maioria dos casos possuem ainda um *permalink*, um link permanente para um determinado blog post que garante a sua acessibilidade mesmo que ele tenha desaparecido do blog que o originou.

Assim como os blogs dos quais fazem parte, blog posts possuem características estruturais próprias, e seguem normalmente o padrão Título, Data, Texto, Comentários, Permalink e Metadados. O título é o título do blog post em questão, e é geralmente exibido nos textos de âncora de links que apontam para o mesmo. A data é a data de publicação daquele blog post. Como muitos blogs são atualizados múltiplas vezes por dia, a data geralmente contém também o horário de atualização. O texto é o conteúdo do blog post em si, escrito pelo autor. Os comentários são justamente isso, comentários sobre o texto, escritos por leitores em resposta ou complementando as informações do texto. A funcionalidade dos permalinks foi descrita acima. Os metadados têm duas informações fundamentais: o autor do blog post, especialmente importante para blogs com múltiplos autores, e temas associados com o blog post.

Blog posts podem ser alterados ao longo do tempo, com correções às informações escritas ou com respostas a alguns comentários. A maneira como estas alterações são realizadas diferem de autor para autor. Alguns anexam novos textos ao final do blog post, outros alteram diretamente o texto, marcando o que

foi alterado com ~~striketrough~~, enquanto outros ainda simplesmente escrevem novos blog posts.

A classificação temática de blog posts é complexa, pois cada um deles pode ter um número pequeno de palavras, e os comentários contidos no mesmo podem utilizar linguagem completamente diferente do blog post em si, ou até mesmo tratar de assuntos diferentes.

A dificuldade da classificação funcional de um blog post depende diretamente das classes sendo consideradas. Visto como um documento HTML, um blog post apresenta poucas características estruturais marcantes, compartilhando sua estruturação visual e segmentação em seções com o blog do qual se originou, mas apresentando menor quantidade de texto. Para blog posts mais longos ou com um número maior de comentários, esta distinção fica consideravelmente reduzida.

Um trabalho que realiza a classificação temática de blog posts é apresentado em [3], onde estas páginas são classificadas de acordo com seu “humor,” ou seja, com a emoção associada ao texto principal presente nelas (ignorando-se os comentários). Nesse estudo, os autores mostram que, embora qualquer tipo de classificação que requer a utilização de blog posts individuais seja difícil, a utilização do conjunto de todos os blog posts de um blog ou da blogoesfera, dentro de certo intervalo de tempo, para a análise de características do conjunto de todos os blogs é eficaz.

3.3. Portais de Notícias

A terceira classe utilizada neste trabalho foram os *Portais de Notícias*. Portais de Notícias são sites que apresentam compilações de notícias dos mais diversos tipos. No geral, estão vinculados a algum veículo de informações fora da Internet, como jornais, revistas ou redes de televisão, embora isso não seja necessário. Para este estudo, como o foco é dado a páginas individuais, considera-se um portal de notícias a primeira página de um veículo de informações deste tipo (www.cnn.com, por exemplo).

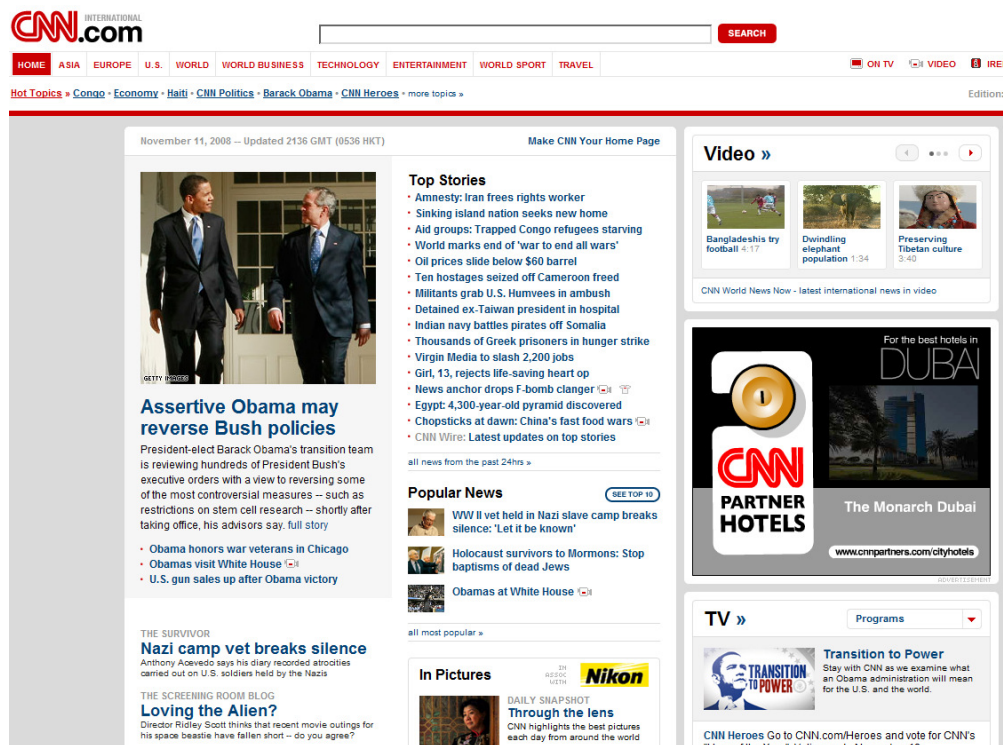


Figura 3 – Imagem do site da CNN (www.cnn.com). Enquanto um portal de notícias, o site apresenta chamadas para diversas reportagens, tanto através de pequenos textos quanto através apenas de manchetes com links. Este portal combina ainda uma série de mídias, como fotos, vídeos e TV via web para fornecer mais informações aos seus usuários.

Um portal de notícias apresenta uma série de complexidades particulares para os métodos de classificação. Primeiro, geralmente combinam uma série de mídias em sua página principal, como reportagens audiovisuais para complementar as tradicionais. Em seguida, alguns destes portais apresentam o título das principais matérias, outros o título e pequenas chamadas, e assim por diante.

Assim como no caso dos blogs, a diversidade de estruturas e temas implica na complexidade das tarefas de classificação temática e funcional. No caso da classificação temática, o principal complicador é a diversidade de temas abordados. Mesmo portais de notícias especializados em ciências podem ter reportagens sobre natureza, clima, tecnologia, e outros temas mais em sua página principal. Já para o caso da classificação funcional, temos dois complicadores. O primeiro é a similaridade de estruturas dos news portals maiores com sites comuns. Muitos jornais apresentam hoje seções de entretenimento, esportes,

lazer, e assim por diante, e se confundem estruturalmente com páginas como MSN (www.msn.com), que possuem elementos de informação parecidos. O segundo é a utilização de plataformas de construção de site. Como temos muitos news portals construídos sobre as mesmas plataformas, qualquer classificador estrutural precisa ser construído de forma a evitar o *overfitting* às características relativas a estas plataformas.

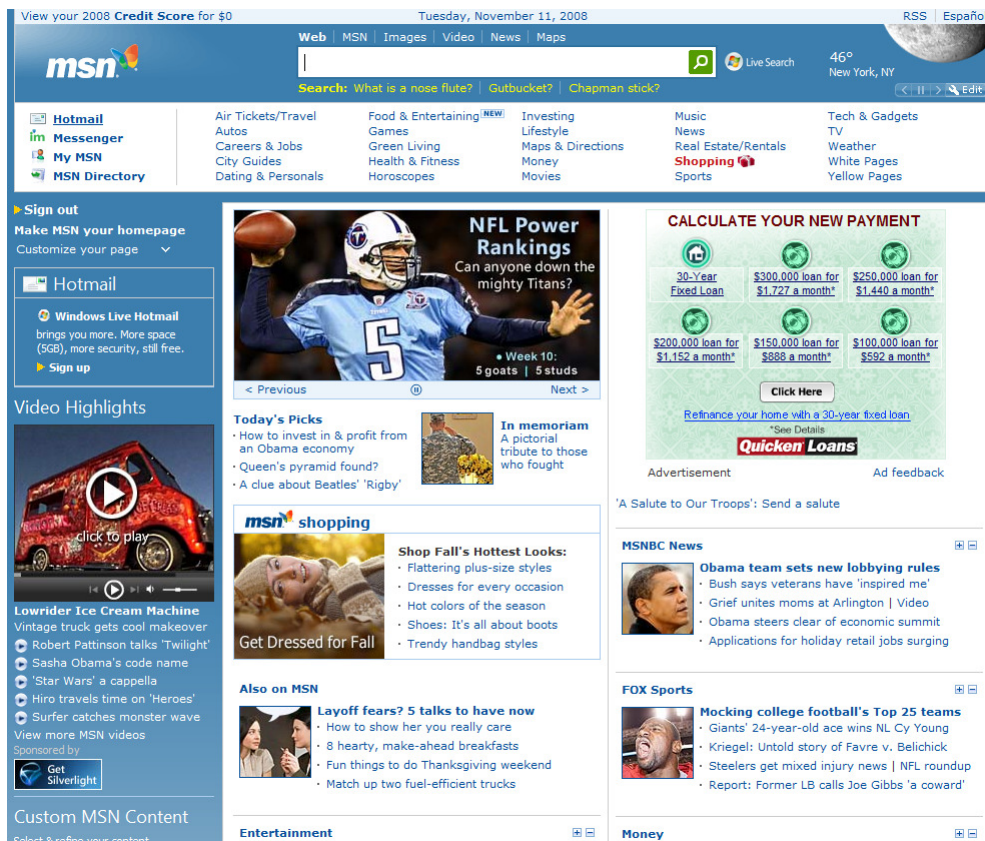


Figura 4 – Imagem do site MSN (www.msn.com). Apesar de apresentar notícias, o site é um portal para os mais diversos tipos de conteúdo, como ferramentas de busca, mapas, vídeos, outros canais de comunicação, ambientes de compras on-line, jogos, acesso a e-mail, dentre outras funcionalidades.

3.4. Notícias

A quarta e última classe selecionada é a de *Notícias*. Notícias estão para portais de notícias assim como blog posts estão para blogs. São páginas referentes a notícias individuais, que estão contidas dentro de news portals. Nem todas as

páginas presentes dentro de um portal de notícias, no entanto, podem ser consideradas notícias. Portais modernos contêm, além de elementos multimídia, conforme mencionado acima, outros veículos de comunicação, como blogs, e até mesmo jogos.

Assim, notícias são páginas que contêm um texto sobre uma determinada matéria, seguindo, no geral, uma estrutura simples, de título e texto. Alguns sites jornalísticos hoje realizam a publicação de matérias on-line em um formato similar a um blog, ou seja, cada notícia fica contida dentro de um blog post realizado por um jornalista. As diferenças fundamentais entre notícias deste tipo e blog posts são a linguagem utilizada (no geral, a linguagem utilizada em posts é muito mais informal do que a utilizada em serviços de notícias) e a ausência de elementos estruturais como blogrolls e comentários.

Notícias se assemelham bastante a blog posts em termos estruturais, uma vez que possuem blocos de texto relativamente curtos, e uma quantidade menor de links externos e características estruturais marcantes, que as diferenciem dos portais das quais se originaram.

SWEDEN'S NEWS IN ENGLISH

News Travel Lifestyle Galleries Dating Opinion Blog Discuss Noticeboard Weather Jobs Property

National Money Politics Sci & Tech Society Education Sport Archive

Swede arrested in US teen sex sting

Published: 8 Aug 08 10:08 CET

A 47-year-old Swede has been arrested in the United States for attempting to arrange sex with a 13-year-old girl. The man from Stockholm could now face up to 30 years in prison.

The 47-year-old father of three had said to his wife and kids that he was heading to the US to visit relatives.

In fact he was travelling to Utah to have sex with a 13-year-old girl. The date did not however work out as planned and he is now detained in a police cell in Layton, just north of Salt Lake City.

The Stockholm man began chatting with a person on the internet in June that he thought was a 13-year-old girl, according to Aftonbladet. In fact he had been in contact with an undercover police officer.

The 47-year-old Swede sent pictures of himself in sexual situations.

"He has on two occasions used a web camera and sent pictures. Once of his penis, and once of himself engaged in sexual activity," said Richard Larsson at the prosecutor's office in Davis county to Aftonbladet.

"The Swede was looking for sexual services," Larsson said to Aftonbladet.

The Local reported on a similar case in June. 40-year-old Jesper Lundberg from Stockholm was imprisoned for almost six years in Virginia after having been lured to the US by police in another internet sting operation.

The 47-year-old man was also unable to resist temptation and flew to the US and to Layton in Utah. On Tuesday he was arrested by police who had set up the meeting with the fictitious teenager.

The man has been charged with trying to engage in illicit sexual conduct with a minor and for distributing dangerous material to a minor. The two offences carry a maximum penalty of up to 30 years.

TT/The Local (news@thelocal.se/08 656 6518)

Today's most read

1. Young woman raped on Malmö street »
2. Brand Sweden enters global top ten »
3. Owner-occupied flats coming to Sweden »
4. Swedish inventions - from the zipper to the pacemaker »
5. Swedish bestiality ring exposed »

Today's most emailed

Police Exam prep books
Since 1987 books and prep materials for entrance and promotion exams
www.fire-police-exams.com

Law Enforcement Jobs
New job listings daily. Browse 1000s of openings.
worldjobster.com

Shop at WebSpider.com
Save time and money by comparing all the online merchants.
www.WebSpider.com

Ads by Google

Poll

Do you think the introduction of owner-occupied apartments will help alleviate the shortage of apartments for rent in Sweden's big cities?

☐ Yes, more people will consider subletting their apartments if they can do so without restrictions

☐ No, what's really needed is a system which allows rents to be set by the market rather than collective bargaining

☐ Maybe, but it will take years before we can know for sure

View Results **Vote »**

Highlights

TICKET ONLINE
arlandexpress.com

Take the train to the meeting from: **Stockholm to Göteborg**
fr. 95:-
Buy now!

Noticeboard

- 11/11 Need a translator? Need to learn English or Swedish?
- 10/11 AutoCad part time work required
- 10/11 Coach or Partner
- 10/11 POSTERS - Free shipping!
- 10/11 Racquetball
- 10/11 Navy Blue Sofa Bed

SHOW ALL NOTICES »

Discuss

- 22:57 History of the Swedish man / 2 posts »
- 22:42 Hunting in Sweden / 25 posts »
- 22:16 changing family ties / 4 posts »
- 22:05 HOW CAN I FIND A FLAT TO RENT IN ... / 5 posts »
- 21:57 Is Sweden less 'couple... / 3 posts »

FULL INDEX »
JUST THE NEWS »
LATEST TOPICS »

Figura 5 – Imagem de uma notícia retirada do site “The Local” (www.thelocal.se), um portal de notícias sueco. A notícia em questão compartilha diversas características com o

portal da qual se originou, com links para outras notícias e outras seções dentro do portal, além de pesquisas e seções para participação.



Figura 6 – Imagem de uma notícia retirada do site do International Herald Tribune (<http://www.ihb.com>), um portal de notícias vinculado a um veículo físico de informações. Apesar de compartilhar características visuais com o portal a qual pertence, esta notícia apresenta muito menos interconexão com outros elementos do portal e um texto mais longo.