Marco Antônio Pinheiro de Cristo

Orientador - Professor Berthier Ribeiro-Neto

Sobre Publicidade Direcionada baseada em Conteúdo

Tese de doutorado apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito para a obtenção do grau de doutor em Ciência da Computação.

Belo Horizonte
17 de Agosto de 2006

À Delle, Eric, Alexandre e André, o que de mais importante tenho na vida.
Aos meus pais e irmãos que tornaram este sonho realidade.
Aos meus amigos, por nunca o deixarem morrer.
...

Agradecimentos

Em primeiro lugar, agradeço de todo o meu coração à minha pequena e maravilhosa família, Danielle, Eric, Alexandre e André. Eles sempre foram a razão de tudo isto e, entretanto, foram os mais sacrificados pela minha ausência. Devo a eles tudo e quero lhes ser grato por todo seu amor, paciência e compreensão.

Este foi sempre um sonho para mim e realizá-lo só foi possível graças a visão dos meus pais, seu carinho e incentivo. Assim, lhes agradeço pois sei que sem eles, nada disto teria acontecido.

Agradecer a todas as pessoas que me ajudaram é uma tarefa penosa pois sempre corro o risco de omitir um ou outro nome. Ainda assim, não poderia deixar de dizer muitíssimo obrigado aos meus irmãos Ana, Maria, Mary e Adauto; à minha segunda família: Marília, Alberto, Daniel, Lucas e Mimi; e aos grandes amigos de sempre a quem devo muito mais do que poderia explicar: Adeíldo, Ademir, Adriano, Aldri, Altigran, Álvaro, Ana, André, André Soares, Andrea, Anísio, Baoping, Bárbara, Beatriz, Bruno, Brasílio, Bruno Possas, Charles, Claudine, Daniel, Daniela, Davi, Davi Reis, Denílson, Dilu, Edleno, Eduardo, Eric, Ernande, Estela, Fabiano, Fabíola, Guilherme, Hermes, Humberto, Ilmério, Jean, Jucimar, Kell, klessius, Leo, Luiz, Maurício, Márcio, Marcos, mateus, Michelle, Miranda, Míriam, Modesto, Monique, Paulinho, Paulo, Pável, Pinheiro, Pitel, Ramurti, Róbson, Rodrigo, Rosa, Thierson, Wesley, e Wilson, entre tantos outros.

À FUCAPI pela oportunidade.

Ao meu orientador, professor Berthier Ribeiro-Neto, por tudo que me ensinou.

A todos vocês, mais uma vez, muito obrigado.

Abstract

The current boom of online advertising is associated with the revenues originated from search advertising, which has become the driving force sustaining monetization of Web services. According to Forrester Research, search advertising revenues were projected to grow from US $3.6 billion in 2004 to US $11.6 billion by 2010. Actually, numbers might be quite larger. To illustrate, Yahoo reported search advertising revenues in the total amount of US $875 million for the second quarter of 2005 only, while Google reported revenues in the total amount of US $1.384 billion for the same period. Further, forecasts suggest that the influence of search advertising will increase in the upcoming years through diversification and the production of new types of search-related advertising. This rapidly consolidating market involves complex business networks and increasingly sophisticated technology. Thus, the exploitation of new forms of search advertising requires advances in the commercial and in the technology front.

In this work, we discuss the use of Information Retrieval (IR) techniques to improve the performance of ad placement methods in search advertising, with emphasis on content-targeted advertising. We investigate how the different sources of evidence already available to information gatekeepers (that operate keyword-targeted advertising systems) affect the matching of ads to the content of a Web page. As a result of this analysis, we propose new strategies for associating advertisements with Web pages. Experiments with a real ad collection show that the proper use of the available sources of evidence can lead to high quality matching algorithms.

We also exploit the combination of conceptual and syntactical evidence. To accomplish this, we first study how to improve Web document classification. We observe that methods based on link evidence can be successfully used to improve the classification based only on content. We then use the best classifiers obtained as source of conceptual information and conclude that matching algorithms in content-targeted advertising can be enhanced by combining the context-based ranking obtained through manual and automatic classification with the ranking provided by the syntactical matching methods.

Resumo

O grande sucesso da publicidade na Internet, observado atualmente, está diretamente relacionado ao investimento crescente em publicidade de busca que, por sua vez, tem sido essencial para o financiamento de conteúdos e serviços na Internet. De acordo com a Forrester Research, os ganhos em publicidade de busca devem subir de 3,6 bilhões de dólares em 2004 para 11,6 bilhões em 2010. De fato, estes números podem ser maiores ainda. Para se ter uma idéia, o Yahoo anunciou um ganho de 875 milhões de dólares referente apenas ao segundo trimestre de 2005, enquanto o Google reportou um ganho de 1,384 bilhões para o mesmo período. Além disso, previsões sugerem que a influência da publicidade de busca irá aumentar nos próximos anos através da diversificação e introdução de novos serviços. A exploração de tais possibilidades implica em avanços tanto nas áreas comercial quanto tecnológica.

Neste trabalho, nós discutimos o uso de técnicas de Recuperação de Informação para melhorar o desempenho de sistemas de atribuição de propagandas em publicidade de busca, com ênfase em publicidade baseada em conteúdo. Nós investigamos como as diferentes evidências já disponíveis para empresas (que operam sistemas de publicidade baseada em palavras-chave) afetam o casamento de propagandas com o conteúdo de páginas da *Web*. Como resultado deste estudo, nós propomos novas estratégias para atribuir propagandas a páginas da *Web*. Experimentos com uma coleção de propagandas real mostram que o uso adequado das evidências disponíveis possibilita um desempenho de alta qualidade no processo de atribuição de propagandas.

Nós também investigamos a combinação de evidências conceituais e sintáticas. Para isto, nós primeiro estudamos métodos para melhorar a classificação de páginas da *Web* e observamos que estratégias de classificação baseadas em apontadores são melhores que as tradicionais, baseadas unicamente na análise de texto. Ao utilizarmos os melhores classificadores obtidos como fonte de informação conceitual, concluimos que a combinação de métodos baseados em casamento sintático com os baseados em casamento conceitual apresenta melhor desempenho que aqueles baseados unicamente em casamento sintático.

Publications by the Candidate

Papers published or accepted for publication during the doctorate:

Book Chapters

1. Marco Cristo, Berthier Ribeiro-Neto, Paulo Golgher, and Edleno S. de Moura. Search Advertising. *Soft Computing in Web Information Retrieval: Models and Applications*, Series: Studies in Fuzziness and Soft Computing, v. 197, p.257–285. Herrera-Viedma, Enrique; Pasi, Gabriella; Crestani, Fabio (Eds.). Springer, 2006.

2. Baoping Zhang, Marcos André Gonçalves, Weiguo Fan, Yuxin Chen, Edward A. Fox, Marco Cristo, and Pável Calado. A Genetic Programming Approach for Combining Structural and Citation-Based Evidence for Text Classification in Web Digital Libraries. *Soft Computing in Web Information Retrieval: Models and Applications*, Series: Studies in Fuzziness and Soft Computing, v. 197, p.65–83. Herrera-Viedma, Enrique; Pasi, Gabriella; Crestani, Fabio (Eds.). Springer, 2006.

Journal Papers

1. Pável Calado, Marco Cristo, Marcos André Gonçalves, Edleno S. de Moura, Berthier Ribeiro-Neto, and Nivio Ziviani. Link-Based Similarity Measures for the Classification of Web Documents. *Journal of the American Society for Information Science and Technology*, v.57, n.2, p.208–221, John Wiley & Sons, Inc., New York, NY, USA. January 2006.

2. Marco Cristo, Pável Calado, Maria de Lourdes da Silveira, Ilmério Silva, Richard Muntz, Berthier Ribeiro-Neto. Bayesian Belief Networks for IR. *International Journal of Approximate Reasoning*, v.34, n.2–3, p.163–179., November 2003.

Conference Papers

1. Adriano Veloso, Wagner Meira Jr., Marco Cristo, Marcos André Gonçalves, and Mohammed Zaki. Multi-Evidence, Multi-Criteria, Lazy Associative Document Classification. *Proceedings of the 15th Conference on Information and Knowledge Management ACM CIKM 2006*, Arlington, VA, USA, November 6-11, 2006, to appear.

2. Anísio Mendes Lacerda, Marco Cristo, Marcos André Gonçalves, Weiguo Fan, Nivio Ziviani, and Berthier Ribeiro-Neto. Learning to advertise. *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, p.549–556, Seattle, WA, USA, August 6-11, 2006.

3. Thierson Couto, Marco Cristo, Marcos André Gonçalves, Pável Calado, Nivio Ziviani, Edleno S. de Moura, and Berthier Ribeiro-Neto. A Comparative Study of Citations and Links in Document Classification. *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, p.75–84, Chapel Hill, NC, USA, June 11-15, 2006.

4. Baoping Zhang, Yuxin Chen, Weiguo Fan, Edward A. Fox, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Intelligent GP Fusion from Multiple Sources for Text Classification. *Proceedings of the 14th Conference on Information and Knowledge Management ACM CIKM 2005*, p.477–484, Bremen, Germany, 31st October - 5th November, 2005.

5. Berthier Ribeiro-Neto, Marco Cristo, Paulo Golgher, and Edleno S. de Moura. Impedance Coupling in Content-targeted Advertising. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p.496–503, Salvador, Brazil, August 15-19, 2005.

6. Baoping Zhang, Yuxin Chen, Weiguo Fan, Edward A. Fox, Marcos André Gonçalves, Marco Cristo and Pável Calado. Intelligent Fusion of Structural and Citation-Based Evidence for Text Classification. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, p.667–668, Salvador, Brazil, August 15-19, 2005.

7. Baoping Zhang, Marcos André Gonçalves, Weiguo Fan, Yuxin Chen, Edward A. Fox, Pável Calado, and Marco Cristo. Combining Structural and Citation-Based Evidence for Text Classification. *Proceedings*

of the 13th Conference on Information and Knowledge Management ACM CIKM 2004, p.162–163, Washington D.C., U.S.A, November 8-13, 2004.

8. Pável Calado, Marco Cristo, Edleno Moura, Nivio Ziviani, Berthier Ribeiro-Neto, Marcos Gonçalves. Combining Link-Based and Content-Based Methods for Web Document Classification. *Proceedings of the 12th International Conference on Information and Knowledge Management ACM CIKM 2003*, p.394–401, New Orleans, Louisiana, USA, November 3-8, 2003.

9. Marco Cristo, Pável Calado, Edleno Moura, Nivio Ziviani, Berthier Ribeiro-Neto. Link Information as a Similarity Measure in Web Classification. *10th Symposium On String Processing and Information Retrieval SPIRE 2003*, p.43–55, Manaus, Brazil, October 8-10, 2003.

Resumo Estendido

Introdução

O grande sucesso da propaganda na Internet, observado atualmente, está diretamente relacionado ao investimento crescente em propaganda associada a máquinas de busca que, por sua vez, tem sido essencial para o financiamento de conteúdos e serviços na Internet. De acordo com a Forrester Research, os ganhos em publicidade de busca devem subir de 3.6 bilhões de dólares em 2004 para 11.6 bilhões em 2010. Além disso, previsões sugerem que a influência da propaganda associada a máquinas de busca, também denominada publicidade de busca, irá aumentar nos próximos anos através da diversificação e introdução de novos serviços. A exploração de tais possibilidades implica em avanços tanto nas áreas comercial quanto tecnológica.

Neste trabalho, nós discutimos o uso de técnicas de Recuperação de Informação para melhorar a relevância em publicidade de busca, com ênfase em publicidade baseada em conteúdo, isto é, propaganda associada ao conteúdo de páginas *Web*. Em particular, procuramos responder às seguintes questões: (1) dadas as evidências disponíveis em um sistema convencional de publicidade baseada em palavras-chave tais como as palavras-chave, o título, a descrição e o conteúdo das páginas dos anunciantes, é possível usar esta informação para melhorar o desempenho de sistemas de publicidade baseada em conteúdo? Além disso, qual o impacto destas evidências e como elas deveriam ser usadas? (2) Dada a diferença observada entre os vocabulários de páginas *Web* e propagandas de anunciantes, como poderíamos minimizar o impacto desta desigualdade? (3) Como melhorar a precisão de classificadores de páginas da *Web* para que estes forneçam informação conceitual de qualidade? (4) Finalmente, como usar tal informação conceitual para melhorar o desempenho de sistemas de publicidade baseada em conteúdo?

Para responder tais questões, nós estudamos o impacto de diferentes fontes de evidência em algoritmos de casamento de propaganda com o conteúdo de páginas *Web*. Propusemos novas estratégias, baseadas em casamento sintático, para seleção de propagandas para uma dada página *Web*,

assim como investigamos novas técnicas para melhorar a classificação de páginas da *Web* com o propósito de usar a informação conceitual fornecida por estes classificadores para melhorar sistemas de publicidade baseada em conteúdo.

Para testar nossos métodos e modelos, realizamos experimentos com coleções reais de propagandas e páginas da *Web*. Nossos resultados indicam que a consideração das evidências disponíveis é útil para melhorar a precisão no casamento de propagandas e páginas *Web* e que o impacto negativo do uso de diferentes vocabulários por anunciantes e provedores de conteúdo pode ser minimizado através de técnicas de expansão. Além disso, a informação de elos (*links*) entre páginas da *Web* se mostrou útil para melhorar a acuidade de classificadores. Finalmente, ao usarmos a informação conceitual fornecida por estes classificadores, fomos capazes de melhorar a precisão de algoritmos de ordenação de propagandas.

Casamento Sintático entre Propagandas e Páginas *Web*

Para avaliar o impacto de diferentes evidências em um sistema de publicidade direcionada baseada em conteúdo, propusemos dez estratégias de casamento sintático para a associação de propagandas a páginas *Web*. Cinco destas estratégias se baseiam na idéia de casar diretamente o texto da página com as evidências textuais associadas à propaganda, ou seja, o título, a descrição e as palavras-chave fornecidas pelos anunciantes. Em particular, a primeira estratégia, `AD`, consiste em casar o texto da página com o texto da propaganda em si, ou seja, seu título e descrição; a segunda estratégia, `KW`, consiste em casar a página com as palavras-chave associadas às propagandas, o que é comumente feito em sistemas de publicidade baseados em palavras-chave; em nossa terceira estratégia, `AD_KW`, são casados o conteúdo da página com o texto da propaganda e suas palavras-chave. Duas estratégias adicionais, `ANDKW` e `AAK`, são extensões das estratégias `KW` e `AD_KW`, respectivamente. Diferente de `KW` e `AD_KW`, em `ANDKW` e `AAK`, as propagandas são associadas às páginas apenas se, ao menos, uma das palavras-chave associadas às propagandas ocorre na página.

As outras cinco estratégias procuram flexibilizar o casamento entre as páginas e as propagandas através da expansão do texto associado às propagandas e às páginas *Web*. Desta forma, é possível reduzir o impacto negativo da diferença freqüentemente observada entre o vocabulário de uma página *Web* e o vocabulário das propagandas. Esta diferença é comum em várias

tarefas de recuperação de informação, uma vez que uma mesma entidade do mundo real pode ser referenciada de diferentes formas e por meio de diferentes palavras. Ela é particularmente grave em nosso caso porque muitos anunciantes tendem a associar poucas palavras-chave às propagandas. Como resultado, o vocabulário das páginas da *Web* tende a ser mais rico que o das propagandas.

Para expandir as propagandas nós usamos basicamente o texto extraído das páginas apontadas pelas mesmas. Para expandir as páginas onde serão colocadas as propagandas, nós usamos termos oriundos de páginas similares obtidas de uma coleção extraída da *Web*.

A idéia de expansão de páginas é motivada pela observação de que, ao descrever um tópico de forma resumida, o anunciante tende a escolher termos de caráter mais geral. Portanto, podemos flexibilizar o casamento da propaganda a uma página *Web* se associarmos termos mais gerais com a página. Assumindo que páginas similares à nossa página alvo p compartilham tópicos em comum, podemos encontrar termos que descrevem melhor o tópico de p inspecionando o vocabulário compartilhado por p com páginas similares a p.

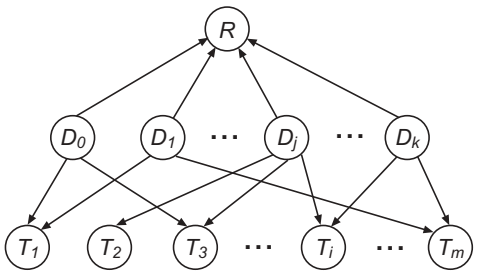


Figura 1: Modelo de rede Bayesiana para seleção de termos do vocabulário compartilhado por uma página p com páginas similares a p.

Para definir os termos deste vocabulário compartilhado, usamos um modelo baseado em rede de crenças, conforme proposto em [104] e ilustrado na Figura 1. Neste modelo, os nós representam páginas *Web* e termos encontrados nestas páginas. Com cada nó é associada uma variável aleatória binária. O nó R representa uma versão da página p estendida com termos do vocabulário compartilhado. O nó D_0 representa a página (ou documento) p. Os nós D_1 a D_k representam as k páginas mais similares a p em uma coleção *Web* $\mathcal{C}$ grande o suficiente. Os nós T_1 a T_m representam os termos no vocabulário de $\mathcal{C}$. Neste modelo, a probabilidade de um termo t_i ser um bom termo para descrever um tópico da página p pode ser calculada como a probabilidade de observar o termo t_i dada a representação estendida da página p, ou seja, $P(T_i = 1|R = 1)$. Isto se traduz na seguinte equação:

$$P(T_i|R) = \rho\ [(1-\alpha)\ w_{i0} + \alpha\ \sum_{j=1}^{k} w_{ij}\ sim(p, d_j)] \qquad (1)$$

onde ρ é uma constante de normalização e w_{ij} é o peso associado ao termo t_i na página d_j. A similaridade $sim(p, d_j)$ pode ser calculada como o valor do cosseno do ângulo entre os vetores correspondentes às páginas p e d_j. O peso w_{ij} pode ser calculado usando um esquema de pesos TF-IDF [108] e é zero se o termo t_i não ocorre em d_j. O peso w_{i0} representa a importância do termo t_i quando ocorrendo diretamente na página p. A constante α pode ser usada para determinar o grau de influência da página p em sua nova representação r. Se α é 0 então os termos na própria página p são considerados, o que é útil para quantificar os resultados sem expansão do vocabulário da página p.

Nós usamos a Equação (1) para determinar o conjunto de termos que irão compor r, a representação estendida da página p. Seja t_{top} o termo de maior probabilidade de acordo com a Equação (1). Então, o conjunto r é composto de termos t_i tal que $\frac{P(T_i|R)}{P(T_{top}|R)} \geq \beta$, onde β é um certo limiar. Em nossos experimentos, usamos $\beta = 0.05$. Note que r pode conter termos já presentes em p. Neste trabalho, nós tanto usamos r como uma versão expandida de p quanto como uma nova representação para p. Assim, nosso sexto método, `AAK_T`, é idêntico ao método `AAK` com a diferença de que o casamento é feito com a nova representação da página sem expansão ($\alpha = 1$). O sétimo método, `AAK_EXP`, é uma versão do método `AAK` onde o casamento envolve a página expandida ($0 < \alpha < 1$). O método `H` corresponde ao casamento da página-alvo p com o conteúdo da página indicada pelo anunciante. Os últimos dois métodos, `AAK_H` e `AAK_EXP_H`, correspondem aos métodos `AAK` e `AAK_EXP` usando propagandas expandidas com o conteúdo das páginas indicadas pelos anunciantes. A Tabela 1 sumariza as estratégias de casamento sintático estudadas neste trabalho.

Para avaliar os métodos descritos, uma série de experimentos foi realizada usando uma coleção de propagandas utilizadas para publicidade de busca no Brasil. Esta coleção possui 93.972 propagandas fornecidas por 1.744 anunciantes. Nesta coleção, apenas uma palavra-chave é associada com cada propaganda. A coleção de páginas-alvo corresponde a 100 páginas de diversos cadernos de um jornal brasileiro. As páginas *Web* utilizadas para expandir o vocabulário das páginas alvo correspondem a um conjunto de 5.939.061 milhões de páginas da *Web* brasileira.

Para cada uma das 100 páginas-alvo, as 3 propagandas do topo, de acordo com a Equação (1), foram inspecionadas por avaliadores humanos. Os avaliadores determinaram quais propagandas eram relevantes. Este processo foi realizado para cada um dos dez métodos da Tabela 1. De posse dos resultados

Estratégias de Casamento	Página Alvo p			Propaganda			
	termos originais	novos termos (r)	inclui palavra-chave	título	descrição	palavra-chave	página da propaganda
AD	×			×	×		
KW	×					×	
AD_KW	×			×	×	×	
ANDKW	×		×			×	
AD_ANDKW (AAK)	×		×	×	×	×	
AAK_T		×	×	×	×	×	
AAK_EXP	×	×	×	×	×	×	
H	×						×
AAK_H	×		×	×	×	×	×
AAK_EXP_H	×	×	×	×	×	×	×

Tabela 1: Sumário das estratégias de casamento sintático. Um "×" indica a evidência ou restrição considerada. Por exemplo, a estratégia ANDKW envolve o casamento da página-alvo p com uma ou mais propagandas, sujeito à restrição de que as palavras-chave associadas às propagandas ocorrem explicitamente em p.

da avaliação, determinamos a precisão média de cada método considerando todas as propagandas sugeridas (PAVG), as três do topo (PAVG@3) e para cada posição do topo (P@1, P@2 e P@3).

Os resultados obtidos são sumarizados na Tabela 2. Esta tabela mostra a precisão média obtida pelos vários métodos. Ela também apresenta os ganhos médios obtidos em relação ao método AAK. Observamos que o melhor desempenho nas primeiras posições (coluna PAVG@3) foi obtido com o método AAK_EXP. Contudo, o melhor resultado global em termos de precisão média (coluna PAVG) foi alcançado pelo método AAK_EXP_H, com um ganho de 42.9% sobre o método AAK.

Os resultados obtidos nos permitem concluir que palavras-chave são essenciais para o casamento de propagandas e páginas. Em particular, métodos que garantem a presença de palavras-chave nas páginas são os melhores entre os métodos de casamento sintático sem expansão. Nós também observamos

Métodos	Precisão			PAVG			PAVG@3		
	@1	@2	@3	valor	ganho(%)	cl(%)	valor	ganho(%)	cl(%)
H	0.310	0.210	0.153	0.060	-65.7	99	0.134	-67.5	99
AD	0.410	0.365	0.287	0.110	-37.1	99	0.257	-37.6	99
AD_KW	0.510	0.395	0.320	0.124	-29.1	99	0.296	-28.2	99
KW	0.460	0.395	0.353	0.136	-22.3	99	0.323	-21.6	99
ANDKW	0.490	0.425	0.400	0.160	-8.6	90	0.364	-11.7	90
AAK	0.510	0.495	0.460	0.175	-	-	0.412	-	-
AAK_H	0.510	0.510	0.463	0.181	+3.4	-	0.421	+2.2	-
AAK_T	0.663	0.582	0.534	0.231	+32.0	95	0.498	+20.9	90
AAK_EXP	0.700	0.610	0.583	0.248	+41.7	99	0.554	+34.5	99
AAK_EXP_H	0.690	0.615	0.570	0.250	+42.9	99	0.533	+29.4	99

Tabela 2: Resultados obtidos por métodos de casamento sintático. Note que *cl* indica o nível de confiança, de acordo com o t-test, dos vários métodos relativo ao método AAK.

que apesar de ser de menor importância, o título e a descrição da propaganda são úteis para o casamento. Por outro lado, o conteúdo das páginas apontadas pelos anunciantes tem pouco impacto na precisão dos métodos de casamento sintático. Finalmente, os melhores resultados foram obtidos com métodos que expandem o conteúdo da página alvo com termos extraídos de páginas similares à página alvo.

Casamento Conceitual entre Propagandas e Página *Web*

Na primeira parte deste trabalho nos concentramos no casamento de termos sem nos preocuparmos com o contexto conceitual em que estes termos ocorrem. Como resultado, muitas propagandas foram colocadas em páginas devido à presença de termos ambíguos. Por exemplo, a associação de uma propaganda sobre o grupo musical Bush com uma página sobre o presidente dos Estados Unidos, Bush, poderia ser evitada se o sistema levasse em consideração a diferença entre os contextos da propaganda (música) e da página (política). Além disso, muitas associações entre páginas e propagandas conceitualmente relacionadas foram desconsideradas devido à não ocorrência de termos em comum. Isto pode ocorrer, em particular, em situações onde a propaganda e a página se relacionam em um contexto mais abrangente. Por exemplo, consumidores podem considerar uma propaganda sobre um CD de música apropriada para uma página sobre o show de um cantor, simplesmente porque ambos são sobre música.

Assim, motivados pela percepção de que associações conceituais indicam boas oportunidades para se colocar ou evitar a colocação de propagandas em páginas *Web*, na segunda parte deste trabalho nós estudamos como esta

informação pode ser explorada em sistemas de colocação de propagandas. Além disso, nós propomos uma estratégia de combinação através da qual a informação conceitual pode ser usada para melhorar a precisão de métodos baseados em casamento sintático. Para tanto, nós utilizamos categorias como o conjunto de conceitos relacionados às páginas e às propagandas. Em particular, nós consideramos o cenário normalmente observado em sistemas reais, em que as categorias das propagandas são préviamente definidas e as categorias das páginas podem ter de ser obtidas automaticamente.

Classificação de Páginas *Web*

Para tirar proveito de associações conceituais, nós estudamos como melhorar a classificação de páginas *Web* e como empregar esta informação *conceitual*, bem como aquela disponível para as propagandas, para melhorar algoritmos de casamento de propagandas a páginas *Web*. O primeiro passo é resolver o problema de classificação. Neste sentido, propomos um modelo de redes Bayesianas para combinar informação de elos (*links*) com informação textual de forma a obter classificadores mais precisos.

Para combinar as informações de elos entre páginas com a informação textual, nós propomos o modelo apresentado na Figura 2. Neste modelo, cada nó corresponde a uma variável aleatória binária. Os nós D_i representam os documentos utilizados para treinar o classificador. Os nós T_j representam a informação textual dos documentos a serem classificados, os nós L_j representam a informação de elos entre documentos (páginas) a serem classificados, o nó C representa a classe em consideração e os nós F_j representam a combinação final de evidências relativas aos documentos a classificar. Quanto maior o sinal F_j, mais provável é a classificação de d_j na classe C.

Dado o modelo, a probabilidade de um documento d_i pertencer à classe C, ou seja, $P(F_i = 1|C = 1)$ é dada por:

$$P(f_i|c) = \rho\Big[1-\big(1-W_t\times class(i,\mathcal{C},\overline{\mathcal{C}})\big)\big(1-W_\ell\times\alpha\sum_{j\in\mathcal{V}(i)\wedge j\in\mathcal{C}} link(i,j)\big)\Big] \quad (2)$$

onde $\rho = P(\mathbf{d})/P(c)$ é uma constante de normalização, $\mathcal{C}$ é o conjunto de documentos de treino declarados a priori como pertencentes à classe C e $\mathcal{V}(i)$ é o conjunto de documentos de treino relacionados ao documento d_i através de elos entre documentos. Finalmente, W_t e W_ℓ são pesos usados para ajustar as contribuições relativas das evidências textual e de elos entre documentos, respectivamente.

A função $class(i,\mathcal{C},\overline{\mathcal{C}})$ representa o valor retornado pelo classificador baseado em conteúdo para a probabilidade de d_i pertencer à classe C. Para

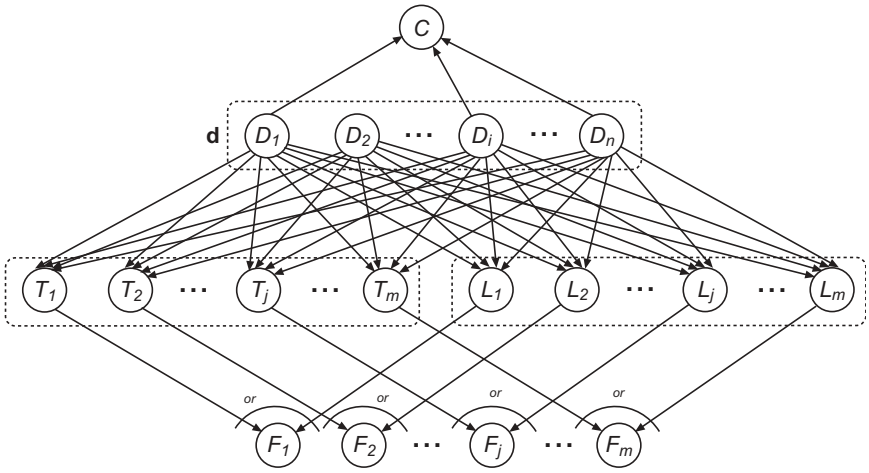


Figura 2: Rede Bayesiana para combinar os resultados fornecido pelos classificadores baseado em texto com os resultados dos classificadores baseados em informação de elos entre páginas.

calcular o valor de *class*, foram usados três classificadores: kNN [140], Naive Bayes [83] e Support Vector Machines [60]. A função $link(i, j)$ representa a similaridade entre o documento d_i e o documento d_j, com base nos elos entre eles. Neste trabalho, esta similaridade foi medida usando co-citação [118], acoplamento bibliográfico [63], a métrica de similaridade de Amsler [2] e o algoritmo Companion [30]. Note que o resultado fornecido por cada um dos classificadores consiste de um vetor de categorias $\vec{v_j} = \{v_{j1}, v_{j2}, ..., v_{jm}\}$ onde v_{ji} corresponde à probabilidade estimada de que o documento d_j seja classificado na classe c_i.

Para avaliar o desempenho dos novos classificadores, foram realizados experimentos com páginas pré-classificadas extraídas do diretório *Web* Cadê (`http://www.cade.com.br`). A partir deste diretório foram formadas duas coleções, o Cade12 e o Cade188. O Cade12 consiste de páginas classificadas nas 12 categorias do primeiro nível do Cadê, enquanto o Cade188 corresponde às páginas classificadas nas 188 categorias do segundo nível. As coleções foram então divididas em sub-coleções de treino e de teste de acordo com o método de validação cruzada de dez grupos [122]. Em outras palavras, cada coleção foi dividida aleatoriamente em dez grupos e, a cada rodada, um diferente grupo foi usado como coleção de teste enquanto os demais grupos foram usados como coleção de treino. Assim, os resultados finais obtidos representam a média das dez rodadas necessárias por experimento.

A Tabela 3 apresenta os resultados da combinação dos resultados dos classificadores baseados em evidências de texto e evidências de elos entre páginas.

Nesta tabela, os resultados são apresentados usando a métrica F_1 [141], ou seja, a média harmônica entre precisão e revocação. As colunas *Ganho/texto* e *Ganho/elos* mostram, respectivamente, o ganho obtido pelo classificador baseado em combinação das duas evidências sobre os resultados dos classificadores baseados em informação de texto e de elos entre documentos tomadas isoladamente. Para o cálculo das medidas de informação bibliométrica foram consideradas tanto as ligações internas (entre documentos do Cadê), quanto as externas (de, e para, páginas fora do Cadê). Finalmente, são apresentados resultados apenas para as medidas Amsler, Co-citação e Companion-autoridades, uma vez que a medida de acoplamento bibliográfico resultou em um grande número de classificações incorretas.

Coleção	Similaridade	W_ℓ/W_t	Class.	F_1	Ganho/texto	Ganho/elos
Cade12	Amsler	2	NB	76.63	44%	11%
	Co-citação	2	NB	76.77	45%	12%
	Companion-Autoridade	2	NB	74.70	41%	13%
Cade188	Amsler	4	*kNN*	67.26	80%	11%
	Co-citação	4	*kNN*	67.45	80%	11%
	Companion-Autoridade	2	*kNN*	64.16	76%	13%

Tabela 3: Resultados obtidos pela combinação de evidências para a classificação de páginas *Web* em classes de um diretório (NB indica o método *Naive Bayes* e *kNN* indica o método *K Nereast Neighbors*).

Destes primeiros resultados podemos concluir que a informação de elos entre páginas pode ser muito útil na classificação de páginas *Web*. De fato, classificadores baseados somente em informação de elos entre páginas tiveram melhor desempenho que os tradicionais baseados somente em texto. Os melhores resultados globais, entretanto, foram obtidos por meio de classificadores que usam a informação de elos entre páginas e a informação textual de forma combinada.

Utilizando informação conceitual (classe da página) para associar propagandas a uma página *Web*

Os melhores resultados finais de classificação de páginas em categorias de um diretório foram alcançados pela combinação das métricas co-citação com classificadores *kNN* e Naive Bayes. Dada a pequena diferença entre os métodos *kNN* e Naive Bayes e o fato de que o método *kNN* produz estimativas numéricas mais adequadas para uma combinação, nós usamos como fontes para evidência conceitual os vetores de categorias fornecidos pelos classificadores *kNN*, co-citação e *kNN* combinado com co-citação. Além destes, também consideramos a informação derivada de classificação manual, uma

Método	Descrição
$\texttt{AAK\_TXT}_{h,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{hard,txt}(a_j, d_i)$
$\texttt{AAK\_LNK}_{h,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{hard,lig}(a_j, d_i)$
$\texttt{AAK\_CMB}_{h,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{hard,comb}(a_j, d_i)$
$\texttt{AAK\_MAN}_{h,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{hard,man}(a_j, d_i)$
$\texttt{AAK\_TXT}_{s,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{soft,txt}(a_j, d_i)$
$\texttt{AAK\_LNK}_{s,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{soft,lig}(a_j, d_i)$
$\texttt{AAK\_CMB}_{s,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{soft,comb}(a_j, d_i)$
$\texttt{AAK\_TXT}_{h,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{hard,txt}(a_j, d_i))$
$\texttt{AAK\_LNK}_{h,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{hard,lig}(a_j, d_i))$
$\texttt{AAK\_CMB}_{h,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{hard,comb}(a_j, d_i))$
$\texttt{AAK\_MAN}_{h,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{hard,man}(a_j, d_i))$
$\texttt{AAK\_TXT}_{s,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{soft,txt}(a_j, d_i))$
$\texttt{AAK\_LNK}_{s,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{soft,lig}(a_j, d_i))$
$\texttt{AAK\_CMB}_{s,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{soft,comb}(a_j, d_i))$

Tabela 4: Métodos de combinação de evidências sintática e conceitual para associação de propagandas a uma página *Web*.

vez que, em alguns cenários, ela pode estar disponível. Com esta informação conceitual, é possível definir uma métrica de similaridade entre uma página d_i e uma propaganda a_j, dada por:

$$csim(a_j, d_i) = \frac{\sum_{t=1}^{m} s_{jt} \times p_{it}}{\sqrt{\sum_{t=1}^{m} s_{jt}^2} \times \sqrt{\sum_{t=1}^{m} p_{it}^2}} \quad (3)$$

onde $\vec{s_j} = \{s_{j1}, s_{j2}, ..., s_{jm}\}$ é o vetor de categorias associado à propaganda a_j e $\vec{p_i} = \{p_{i1}, p_{i2}, ..., p_{im}\}$ é o vetor de categorias associado à página d_i. Em outras palavras, a similaridade conceitual de a_j em relação a d_i é dada pelo cosseno do ângulo entre os vetores de categorias correspondentes.

Neste trabalho, os vetores de categorias $\vec{s_j}$ e $\vec{p_i}$ são usados de maneira a se considerar tanto todas as categorias não nulas (classificação *soft*) quanto apenas a categoria mais provável (classificação *hard*). Dados estes diferentes tipos de classificação (*hard* e *soft*), os vários classificadores a serem considerados pela função $csim(a_j, d_i)$ (kNN, co-citação, cocitação-combinado-comkNN e manual) e as formas de combinação estudadas (conjuntiva e disjuntiva), diversos tipos de combinação entre métodos conceitual e sintático são possíveis. Para refletir estas várias possibilidades, nos referimos à função $csim(a_j, d_i)$ como $csim_{cv,cl}(a_j, d_i)$, onde cv pode assumir os valores "hard" ou "soft" e cl pode assumir os valores "txt", "lig", "comb" ou "man". A tabela 4 descreve as várias combinações de evidências estudadas neste trabalho.

Métodos	Precisão			PAVG			PAVG@3		
	@1	@2	@3	valor	ganho(%)	cl(%)	valor	ganho(%)	cl(%)
linha de base									
AAK	0.640	0.620	0.600	0.173	-	-	0.552	-	-
combinação conjuntiva, classificação *hard*									
AAK_TXT$_{h,a}$	0.672	0.641	0.594	0.160	-7.5	99	0.559	+1.3	99
AAK_CMB$_{h,a}$	0.788	0.753	0.714	0.195	+12.7	-	0.684	+23.9	-
AAK_LNK$_{h,a}$	0.821	0.769	0.739	0.209	+20.8	-	0.710	+28.6	-
AAK_MAN$_{h,a}$	0.822	0.806	0.767	0.218	+26.0	-	0.733	+32.8	99
combinação conjuntiva, classificação *soft*									
AAK_TXT$_{s,a}$	0.688	0.667	0.639	0.183	+5.8	-	0.596	+8.0	-
AAK_LNK$_{s,a}$	0.753	0.706	0.691	0.200	+15.6	-	0.654	+18.5	95
AAK_CMB$_{s,a}$	0.750	0.745	0.720	0.207	+19.7	99	0.678	+22.8	99
combinação disjuntiva, classificação *hard*									
AAK_TXT$_{h,n}$	0.680	0.650	0.633	0.182	+5.2	-	0.583	+5.6	-
AAK_CMB$_{h,n}$	0.730	0.705	0.697	0.206	+19.1	98	0.652	+18.1	98
AAK_LNK$_{h,n}$	0.750	0.685	0.690	0.206	+19.1	99	0.641	+16.1	98
AAK_MAN$_{h,n}$	0.520	0.525	0.537	0.149	-13.9	-	0.457	-17.2	90
combinação disjuntiva, classificação *soft*									
AAK_TXT$_{s,n}$	0.670	0.665	0.640	0.179	+3.5	-	0.591	+7.1	-
AAK_LNK$_{s,n}$	0.750	0.695	0.680	0.206	+19.1	99	0.633	+14.7	95
AAK_CMB$_{s,n}$	0.710	0.710	0.707	0.206	+19.1	98	0.655	+18.7	98

Tabela 5: Precisão média das listas de propagandas obtidas por meio dos métodos de combinação conceitual (decisão dos classificadores) e informação sintática (método AAK). Note que *cl* representa o nível de confiança obtido pelo método t-test na comparação com o método AAK.

Para avaliar o desempenho dos métodos descritos na Tabela 4, um conjunto de experimentos foi realizado usando páginas da coleção Cade188 e uma coleção de propagandas reais pré-classificadas. A Tabela 5 apresenta a precisão das listas de propagandas obtidas por meio dos métodos descritos na Tabela 4 em comparação ao método AAK. Como podemos ver, vários métodos de combinação apresentam ganhos sobre o método AAK. Em particular, quando consideramos a métrica PAVG@3, o melhor método envolvendo classificação manual, AAK_MAN$_{h,a}$, apresenta um ganho de 32.8% sobre AAK enquanto o melhor método envolvendo classificação automática, AAK_CMB$_{s,a}$, apresenta um ganho de 22.8%.

Dos resultados obtidos, podemos concluir que a informação conceitual se mostrou útil para melhorar métodos de colocação de propagandas baseadas em casamento sintático. Além disso, em geral, quanto maior a precisão do classificador, maior o ganho obtido com o método de combinação de informação conceitual e sintática. Em nossos experimentos, a informação conceitual fornecida pelo classificador de menor precisão – o classificador baseado em texto que obteve uma precisão média de 40% na coleção de teste – não foi útil para melhorar a qualidade das listas de propagandas. Por outro lado, a informação fornecida pelos melhores classificadores – baseados em informação de elos entre páginas e combinação de informação de elos entre

páginas e texto – levaram a ganhos significativos sobre o método AAK. Finalmente, em geral, os melhores resultados de combinação foram obtidos com classificação manual. Nós também observamos que enquanto significativo para a tarefa de classificação, o ganho em precisão do método envolvendo combinação de informação textual e de elos entre páginas não foi grande o suficiente para fazer deste método a melhor fonte de informação conceitual. Finalmente, para classificadores mais precisos, tais como no cenário em que classificação manual é possível, a combinação conjuntiva das evidências textual e conceitual deveria ser preferida.

Conclusões

Neste trabalho, nós estudamos algoritmos para casamento de propagandas e páginas *Web* a serem utilizados em publicidade baseada em conteúdo. Em particular, investigamos o impacto de diferentes fontes de evidência bem como propusemos estratégias para lidar com a diferença entre os vocabulários usados nas páginas e nos anúncios e a tendência dos usuários de considerarem relevantes associações entre páginas e anúncios pertencentes a uma mesma categoria.

Experimentos efetuados considerando uma coleção real de propagandas e coleções de páginas extraídas da *Web* brasileira nos permitiram concluir que:

- A consideração cuidadosa das evidências disponíveis, tais como aquelas usadas comumente por empresas que operam sistemas de publicidade baseada em palavras-chave, pode levar a melhores sistemas de publicidade baseada em conteúdo. Em particular, observamos que as palavras-chave, o título e a descrição fornecida pelos anunciantes são úteis para melhorar a precisão obtida por algoritmos de casamento de páginas e propagandas. O mesmo não foi observado para o conteúdo das páginas apontadas pelos anúncios. Como esperado, as palavras-chave se mostraram a fonte de informação mais útil, especialmente quando consideramos relevantes apenas as páginas em que elas ocorrem;

- O impacto negativo do emprego de diferentes vocabulários para páginas e propagandas pode ser minimizado através da expansão do conteúdo destas páginas com termos extraídos de páginas de conteúdo similar;

- A informação de elos entre páginas se mostrou bastante útil para indicar se páginas da *Web* são similares, pelo menos, quando consideramos páginas muito referenciadas como aquelas extraídas de diretórios da *Web*. Assim, é possível melhorar a precisão de classificadores de

páginas *Web* através da combinação de informação baseada em elos entre páginas com aquela baseada no conteúdo textual das páginas;

- A informação conceitual obtida por meio de classificadores é útil para melhorar a precisão de algoritmos de ordenação de propagandas. Em particular, fomos capazes de melhorar o desempenho de algoritmos para a colocação de propagandas através da combinação de informação conceitual obtida por meio de classificadores automáticos e manuais com a informação derivada de métodos de casamento sintático.

Este trabalho não se encerra aqui, uma vez que tanto deixou questões em aberto como nos levou a novas idéias. Algumas destas questões e idéias são descritas a seguir. Inicialmente, novas estratégias de avaliação deveriam ser empregadas de forma a considerar elementos mais indicativos do retorno financeiro dos atores envolvidos. Além disso, dada a natureza universal da propaganda na Internet, outras coleções de páginas diferentes daquelas usadas aqui deveriam ser avaliadas. Outras evidências e técnicas para combiná-las também deveriam ser estudadas. Por exemplo, o valor que os anunciantes estão dispostos a pagar pela veiculação de suas propagandas deveria ser considerado no futuro. Entre outras técnicas de combinação, citamos aquelas baseadas em programação genética. Também no futuro, deverão ser estudadas estratégias para minimizar a colocação de anúncios irrelevantes ou inadequados. Entre novas idéias a serem exploradas, citamos o estudo de técnicas mais sofisticadas para o tratamento de frases e expressões, bem como para a expansão e agrupamento de propagandas.

On Content-targeted Advertising

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, we describe the motivation for our research and discuss our goals and contributions.

1.1 Motivation

The Internet's emergence represented a new marketing opportunity to any company – the possibility of global exposure to a large audience at a dramatically low cost. In fact, during the 90's many organizations were willing to spend great sums on advertising in the Internet with apparently no concerns about their investment return [136]. As a result, the Internet became the media of fastest growth in its first five years, according to the Interactive Advertising Bureau [57].

This situation radically changed in the following decade, when the failure of many Web companies led to a dropping in supply of cheap venture capital. This lead to wide concern over the value of these companies as reliable marketing partners and, as a result, to considerable reduction in on-line advertising investments [135, 136]. Such reduction caused consecutive declines of quarterly company revenues in the US market, beginning with the first quarter of 2001. This loss trend, however, has been reversed by the end of 2002 as seen in Figure 1.1. Further, it has been growing steadily since reaching peak values by the end of 2005 [57].

To better understand the reasons for this recover of the online industry, we have to analyze how different Web advertising formats have performed over time. Table 1.1 shows revenues generated by eight distinct forms of Internet advertising, as measured by IAB[1]: display ads, sponsorships, email,

[1]Display ads is the format in which advertisers pay on-line companies to display banners or logos on one or more of the company's pages. In Sponsorship advertising, an advertiser

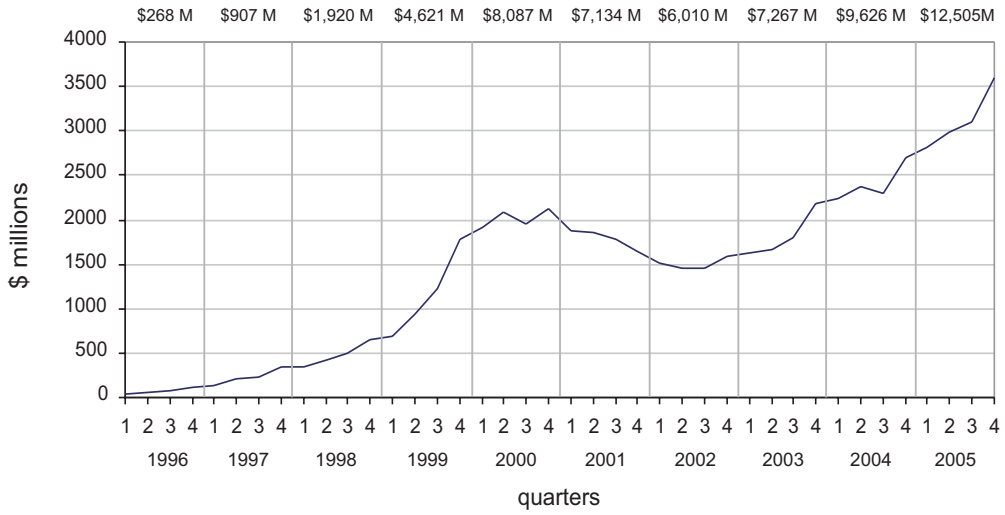


Figure 1.1: Quarterly Revenue Growth Comparisons 1996-2005. Source: IAB, 1996-2005.

classifieds/auctions, rich media, search, referrals, and slotting fees [57].

As we can see in Table 1.1, there were important changes in the popularity of the various forms of advertisements (ads). For example, display ads (which include banners) gradually declined from 56 percent in 1998 to 20 percent in 2005. Similar decrease in usage is observed for sponsorships. On the other hand, search advertising rose from 1 percent in 2000 to 40 percent in 2005, becoming the leading form of Internet advertising. Thus, the recovery of Web advertising coincided with the increasing adoption of search advertising. This growth has not been restricted to the USA, since similar gains have been reported in Europe [103]. It is not either a transitory phenomenon, since both advertisers and publishers have announced plans to increase their investments in search advertising [67,113]. In fact, according to Forrester Research projections, by 2010, search advertising alone

sponsors targeted Web site or email areas to build good-will more than traffic to its site. E-mail advertising accounts for ads associated with commercial e-mail communications. In Classifieds and Auctions, advertisers pay on-line companies to list specific products or services. Rich media is a generic term for a variety of interactive ads that integrate video and/or audio. In Referrals, advertisers pay on-line companies for references to qualified leads or purchase inquiries. In Slotting Fees, advertisers pay on-line companies for preference positioning of an ad on the company site. In search advertising, advertisers pay on-line companies to list and/or link the company site to a specific search keyword or page content, as well as to optimize their pages for search engines and ensure their insertion in search indexes.

Advertising Formats	1998	1999	2000	2001	2002	2003	2004	2005
Display ads	56	56	48	36	29	21	20	20
Sponsorships	33	27	28	26	18	10	9	5
Email	-	2	3	3	4	3	3	2
Classifieds/auctions	-	-	7	16	15	17	17	18
Rich media	5	4	6	5	10	10	8	8
Search	-	-	1	4	15	35	40	40
Referrals	-	-	4	2	1	1	2	6
Slotting fees	-	-	-	8	8	3	2	1
Other	6	11	3	-	-	-	-	-
Total	100	100	100	100	100	100	100	100

Table 1.1: Internet advertising revenues by type of advertisements, percentage figures - 1998-2005. Source: IAB, 1998-2005.

will represent a market of US$11.2 billion [80]. As a consequence, an entire new industry offering search advertising related services has emerged, in part by reverse engineering the search engine ranking algorithms [34]. Such services comprehend consultancy on keyword selection, performance analysis, site optimization, etc.

In search advertising methods, an advertiser company is given prominent positioning in ad lists in return for a placement fee. Because of this, such methods are called *paid placement strategies*. Amongst these methods, the most popular one is a non-intrusive technique called *keyword targeted advertising* [136]. In this technique, keywords extracted from the user's search query are matched against keywords associated with ads provided by advertisers. A ranking of the ads, which also takes into consideration the amount that each advertiser is willing to pay, is computed. The top ranked ads are displayed in the search result page together with the answers for the user query.

The success of keyword targeted advertising has motivated information gatekeepers to offer their ad services in different contexts. For example, relevant ads could be shown to users directly in the pages of information portals. The motivation is to take advantage of the users immediate information interests at browsing time. The problem of matching ads to a Web page that is browsed, which we also refer to as *content-targeted advertising* [74], is different from that of keyword targeted advertising. In this case, instead of dealing with users' keywords, we have to use the contents of a Web page to decide which ads to display.

It is important to notice that paid placement advertising strategies imply some risks to information gatekeepers. For instance, there is the possibility

of a negative impact on their credibility which, at long term, can demise their market share [6]. This makes investments in the quality of ad recommendation systems even more important to minimize the possibility of exhibiting ads unrelated to the user's interests. By investing in their ad systems, information gatekeepers are investing in the maintenance of their credibility and in the reinforcement of a positive user attitude towards the advertisers and their ads [133]. Further, that can translate into higher click-through rates which leads to an increase in revenues for information gatekeepers and advertisers, with gains to all parts [6].

1.2 Objectives and Contributions

In this work, we study how to improve the precision of the matching algorithms used in content-targeted advertising. In particular, this study was driven by some research questions which we intend to answer. These questions are described in the following paragraphs.

Content-targeted advertising is based on the idea that advertisers will bid on keywords that they believe are good indicators of the products and services to be advertised. However, ads are composed of more information than only keywords. In fact, if we consider only the evidence sources already available to information gatekeepers that operate keyword-targeted advertising systems, an ad can be viewed as a structured document composed of a title, a description, and a link to an external page whose contents are related to the ad. This leads us to our first research questions: could these fields provide useful information to enhance content-targeted advertising? What is the impact on ads selection of matching each of these fields? How should they be used?

Studying this problem, we observed a frequent mismatch between the vocabulary of a Web page and the vocabulary of an ad. This is aggravated by the fact that many advertisers bid on few keywords and select keywords of general nature. Consequently, specific terms present in targeting pages may have low impact on the ad selection. This leads us to other research question: how could we minimize the impact of this vocabulary mismatch?

We also observed that associations between ads and pages may be appropriate even when the ad and the page are related to each other in a broad conceptual scope. Further, common misplacements are caused by ambiguous terms that are matched without taking into consideration their meanings. This suggests that conceptual associations may indicate good opportunities to either avoid or place ads. However, obtaining reliable conceptual information is a hard task. This led us to the study of Web classification and to one

more question: how could we improve the accuracy of the classifiers used in Web classification to provide reliable conceptual information?

Finally, once the conceptual information is available, we are confronted with our last two questions: is the conceptual information obtained really useful to enhance content-targeted advertising? And how could we use it?

In our search for proper answers, we analyzed the impact on matching of ads to a Web page of different sources of evidence. Then, we proposed (a) new strategies based on syntactical matching for associating ads with Web pages, (b) new strategies for enhancing Web document classification, and (c) new strategies for combining syntactic information with conceptual information (on the classes of a Web page and of an ad) to improve the precision of matching algorithms in content-targeted advertising.

In particular, we proposed formal models based on Bayesian networks to expand Web pages and to combine different sources of evidence in Web classification. Based on our belief that the hypertextual nature of the Web can be used to indicate a document's topic and importance, we investigated how effective is link information to assist with document classification. We also studied how to combine rankings provided by methods based on conceptual and syntactical similarity metrics. To test these methods and models, we performed experiments using actual ads and Web collections. The major contributions of our work are:

- An empirical study on the impact of different sources of evidence to matching algorithms used in content-targeted advertising;
- A new method for expanding Web page contents to facilitate the match of ads and Web pages;
- New methods for Web document classification based on the combination of link-based and content-based information;
- A detailed empirical study on the effects of link information on classification of Web documents, including the analysis of several similarity metrics;
- A detailed empirical study on the effects of combining syntactical and conceptual similarity metrics in content-targeted advertising;
- A set of important guidelines on how links should be used to improve the effectiveness of document classification and how the conceptual information provided by these classifiers should be used to enhance matching algorithms in content-targeted advertising.

1.3 Organization of this Work

The first part of this work, composed of Chapters 1, 2, and 3 provides some background on topics related to our work. In particular, Chapter 2 introduces basic concepts related to search advertising, Information Retrieval (IR), link analysis, and Bayesian networks, essential to the understanding of this work. Chapter 3 discusses research on subjects related to our work.

In the second part, composed of Chapters 4 and 5, we present methods based on syntactical and conceptual information to improve the precision in content-targeted advertising. For this, two models are presented to expand Web pages and to combine link-based and content-based information. In particular, Chapter 4 presents new strategies for associating ads with Web pages according to their syntactical similarity. Five of these strategies are based on the idea of matching the text of the Web page directly to the text of the ads and its associated keywords. Five other strategies are based on the idea of expanding the Web page with new terms to facilitate the task of matching ads and Web pages. Experiments with a real ad collection indicate that, by reducing vocabulary mismatch between pages and ads, we are able to improve the precision of ad placement systems in content-targeted advertising.

Chapter 5 exploits the combination of conceptual and syntactical evidence. Initially, this chapter presents a Bayesian network model that combines content-based and link-based information to enhance Web document classification. Experiments with two collections of Web documents indicate that the model can be successfully used to improve classification methods based only on content. Following that, the best classifiers are used as source of conceptual information for matching ads to a Web page. Experiments are performed to investigate the impact of conceptual information of progressively better quality, how to treat the category decisions, and how to combine conceptual and syntactical information.

Finally, in Chapter 6, we present our final conclusions and some suggestions regarding future steps for this research.

Chapter 2

Basic Concepts

This chapter introduces basic concepts required for a better understanding of our proposed methods. We start by describing, in Section 2.1, the main types of search advertising, particularly, the keyword-targeted advertising and the content-targeted advertising. In this section, we also describe the dynamics of these systems. Since our framework is based on Bayesian networks, we review them in Section 2.2. After that, we present the vector space model in Section 2.3. This is the most common IR method for ranking documents and will be used as our comparison baseline in our discussion on content-targeted advertising matching methods. Further, this model is commonly used as part of *kNN* classifiers. This classifier, as well Naive Bayes and Support Vector Machine, will be presented in Section 2.4. In this work, they will be used as comparison baselines to our proposed classification methods. In Section 2.5 we discuss the linkage similarity measures to be used in our classification algorithms. Finally, Section 2.6 presents some measures commonly used to evaluate the retrieval performance of IR systems, useful to understand the results shown in the following chapters.

2.1 Search Advertising

Search advertising is in constant evolution and many different technical approaches have been proposed over the last few years. In this section, we discuss the main types of search advertising described by IAB: keyword-targeted advertising and content-targeted advertising. These were originally called paid listings and contextual search in IAB reports, respectively [57]. In spite of this work being focused on content-targeted advertising, we start by discussing keyword-targeted advertising because this approach introduced many of the basic concepts commonly used in the literature. We finalize this

section by describing search advertising networks and its actors.

2.1.1 Keyword-targeted Advertising

Keyword-targeted advertising was introduced by Overture[1] in 1998 [48]. It consists in showing a list of ads at the top or at the right hand side of a Web page search results. The ads displayed have to be "related" to the content of the user query (normally composed of one or more words). Figure 2.1 shows an example of keyword-targeted advertising.

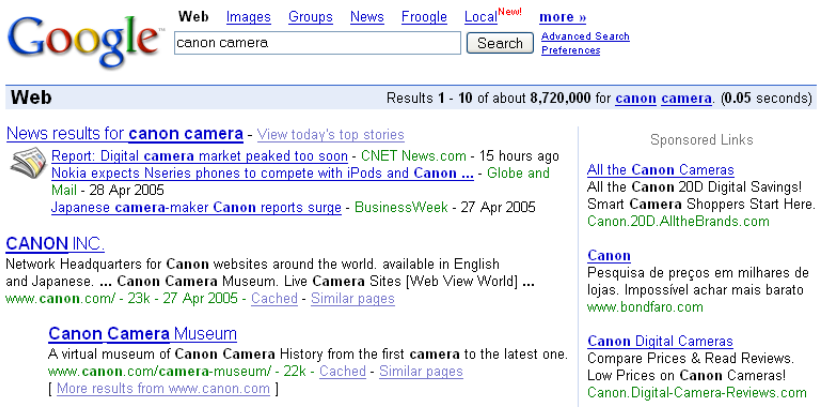


Figure 2.1: Example of keyword-targeted advertising. The ads are related to the user query "canon camera".

The list of ads is called a *paid list*. It is composed of a small number of ads, normally three or five, as shown in the right hand side of Figure 2.1. The ads can be presented to the users in different formats. However, they are generally showed as static text, to simplify ad creation and reduce impact on the page download time. This text normally comprises a title, a short description, and a URL address. The content of the title and the description is also called a *creative*. The creative consists of a concise action-oriented text designed to attract the user. Consequently, hot words such as "free" and call-to-action phrases like "Click here" or "Enter now to win a..." are very common. For the top ad exhibited in Figure 2.1, the title is "All the Canon Cameras"; the description is "All the Canon 20D Digital Savings! Smart Camera Shoppers Start Here." and the URL address is "Canon.20D.AlltheBrands.com".

Besides these visible parts, with each ad is associated a set of keywords $\mathcal{K} = \{k_1, k_2, ..., k_m\}$. The keywords comprise one or more words and are used

[1] `http://www.overture.com`

by the advertisers to describe the possible interests of the targeted audience. For instance, for the first ad shown in Figure 2.1, the ad keyword could be "Canon" or "digital camera". To associate a certain keyword k with one of its ads, the advertiser has to bid on k in an auction type system. The more the advertiser bids on k, greater are the chances that its ads will be shown in the ad list associated with that keyword. Notice that the advertisers will only pay for their bids when the users click on their ads. Because of this, keyword-targeted advertising is called a *pay-per-performance system*. Further, an advertiser can associate several ads with the same product or service. We refer to such group of ads as a *campaign*. Notice that only an ad per campaign should be placed in a Web page in order to ensure a fair use of the page advertising space and increase the likelihood that the user will find an interesting ad.

By creating a compelling ad, the advertiser expects to get users to click on it and jump to its *landing page*, that is, the page indicated by the URL address. In the landing page the user will find more information related to the ad or to the company, its products, and services, and possibly will start a transaction. If this is the case, we say that the click was converted into a transaction, an event called *conversion*. Here, a transaction is any action with which an advertiser associates value, such as the purchasing of a product or service, the downloading of a white paper, the requesting of a commercial proposal, etc.

2.1.2 Content-targeted Advertising

Content-targeted advertising was introduced by Google[2] in 2002 [67, 74]. It is analogous to keyword-targeted advertising in the form advertisers select and pay for keywords. The selection of the ads to be displayed, however, is based on the content of the page p being viewed, instead of on a user query. The page p is called *triggering page*.

Nowadays, content-targeted advertising is the dominant contextual approach in Web marketing. It has raised more interest than other approaches that, previously, were considered more promising, such as behavioral targeting [113]. Once the most relevant and profitable ads are known, they have to be shown to the users. As in the case of keyword-targeted advertising, the ads can be grouped in paid lists and positioned in the triggering page. Figure 2.2 shows an example of content-targeted advertising in which the ads are exhibited in the two slots of a paid list in a triggering page.

[2] `http://www.google.com`

Figure 2.2: Example of content-targeted advertising in the page of a newspaper. The middle slice of the page shows the beginning of an article about the launch of a movie on DVD. At the bottom slice, we can see ads picked for this page by Google's content-targeted advertising system, AdSense.

A Retrieval System for Content-targeted Advertising

In this work, we consider that the main task of a content-targeted advertising system consists of getting the first k ranked ads of a collection $\mathcal{A}$ according to their relatedness to the triggering page p. This task resembles that of a traditional IR system, i.e., to get the first k ranked documents satisfying a given query. In content-targeted advertising, the ads play the role of the documents and the triggering page plays the role of the query. Thus, given the triggering page $p \in \mathcal{D}$ and the set of ads $\mathcal{A}$, the ad retrieval function of a content-targeted advertising system can be defined by Equation (2.1):

$$adSearchTop(k, p, fsim(., p)) = \text{top } k \text{ ads in } sort(\mathcal{A}, fsim(., p)) \qquad (2.1)$$

where $fsim(., p)$ is a similarity function applied to p and $sort(\mathcal{A}, fsim(., p))$ sorts the ads in $\mathcal{A}$ according to $fsim(., p)$. In other words, function $fsim$ defines the ads to be placed in the triggering page. In this work, the content-targeted advertising methods to be proposed will basically redefine function $fsim$.

Note that function $adSearchTop(k, p, fsim(., p))$ does not ensure, as far as possible, that only one ad per campaign will be placed in the triggering page. In the following, we formally define this restriction. Let $\mathcal{C} =$

$\{C_1, C_2, ..., C_n\}$ be a partition of $\mathcal{A}$ that represents the set of campaigns $C_1, C_2, ..., C_n$. Let $\delta_{ijp}\colon \mathbb{N} \times \mathcal{C} \times \mathcal{D} \rightarrow \mathbb{R}$ be a function that represents the similarity value of the i-th top-ranked ad of campaign C_j according to the function $fsim$. For instance, if a_s is the second top-ranked ad of campaign C_5, $\delta_{25p} = 0.5$ indicates that $fsim(a_s, p) = 0.5$. Given these definitions, the ranking lists obtained by means of $sort(\mathcal{A}, fsim(., p))$ should satisfy the constraint:

$$\forall_{i,j,k|j\neq k}\ (\delta_{ijp} > 0 \wedge \delta_{(i+1)kp} > 0 \Rightarrow \delta_{ijp} > \delta_{(i+1)kp}) \tag{2.2}$$

A simple way to meet this requirement, without modifying function $fsim$, is to design the function *sort* according to a round robin strategy, as follows. For each campaign, a ranking is built according to the similarity function $fsim$. The *sort* function selects the top ranked ad of each ranking till that all the campaigns have been considered. The process is then repeated until that no ads remain to be selected. By doing this, we guarantee that the i-th top-ranked ad of a campaign will always be placed into a page above the $(i + 1)$-th top-ranked ad of any other campaign.

2.1.3 The Search Advertising Network

Many of the key elements of search advertising are not new. Both contextual placement and payment per performance have been used since the beginning of the Internet [55,111]. However, these first attempts were crude. Matching systems were very simple. As a result, irrelevant messages were often shown to users, annoying them. Also, questionable practices were very common such as analyzing user behavior without its consent and popping up ads in pages without the permission of their publishers, associating their image with improper companies, products, and services [55,71,134]. Such strategies were also not good for most of the advertisers since user attitude towards these approaches was very negative [79,134]. Consequently, this commonly implied in branding problems [55].

It is not surprising that the same elements that work so well in search advertising have failed consistently in the past. Contrary to these first efforts, search advertising is characterized by a relationship pattern in which all the participant actors are benefited[3]. In fact, the success of search advertising

[3]Such pattern is also observed in traditional advertising-financed mass media where content providers attempt to satisfy interests of recipients and advertisers. This pattern is characteristic of a contractual framework employed to emphasize cooperation and discourage participants' opportunistic behavior. The reader is referred to [129] for a detailed discussion about such relationship patterns from an economic theory perspective.

can be credited to the formation of such reliable networks [75]. In general terms, these networks are composed of four main actors: the broker, the advertisers, the publishers, and the users, as depicted in Figure 2.3.

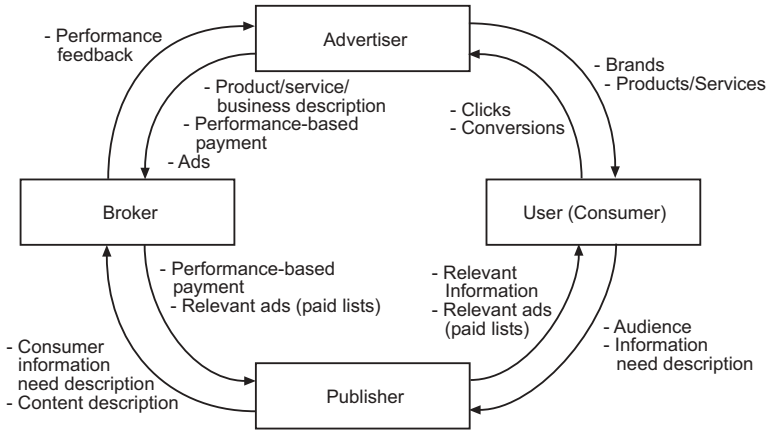


Figure 2.3: Search advertising network.

The broker is responsible for the maintenance of the network. It determines which advertisers and publishers can participate of the network as well as the publishing policies to be followed. For example, brokers cannot allow pornographic content, improper language, and copyright violation. They also want to avoid the participation of companies that promote or deal with illegal matters such as drugs or gambling games. Further, they are responsible for making the auction system work by offering tools (interfaces, databases, controlled vocabularies) that the advertisers use to bid on the keywords necessary to describe their products and services. The broker is also responsible for the technology that will be used to match keywords/content and ads [76] and for the measurement systems that will allow evaluating the performance of the publishers and the advertisers [48].

The advertisers participate of the network with the expectation that they will be referred to quality users by the publishers. From the advertisers' point of view, quality users are those which are interested or could become interested in their products or services. This is the case of many users looking for information in directories and search engines or browsing editorial content in the Web. The advertisers compete among themselves for keywords by bidding in an auction system [34, 48, 59]. They pay to the broker according to the traffic provided to them by the publishers. Based on the performance reports they receive, it is possible to tune their campaigns dynamically, which allows them to maximize their revenues and, by extension, the quality of the

overall system [49, 111].

The publishers are interested in monetizing their pages through the loyalty of their audience. Thus, the publishers may provide the broker with a description of the content of their pages by using keywords and/or categories [87]. In general, however, an automatic system provided by the broker will infer *essential topics* related to a broker-provided content in an automatic or semi-automatic fashion. Notice that we also consider information gatekeepers, like search engines and directories, as publishers. They provide the broker with the user query or with the directory entries selected by the user. In this case, the publishers' payment is based on the traffic they provide.

The last actor in the network is the user or consumer. The users are interested in getting *relevant* information from the publishers. Consequently, they naturally segment themselves by describing their information needs by means of keywords or by surfing in Web pages whose content is of their interest [103]. Occasionally, they can click on the ads exhibited, jump to the advertisers' pages, and start commercial transactions.

2.2 Bayesian Networks

Bayesian networks provide a graphical formalism for explicitly representing independencies among the variables of a joint probability distribution. The probability distribution is represented through a directed acyclic graph, whose nodes represent the random variables of the distribution. Thus, two random variables, X and Y, are represented in a Bayesian network as two nodes in a directed graph, also referred to as X and Y. An edge directed from Y to X represents the influence of node Y, the *parent* node, on node X, the *child* node. Let x be a value taken by variable X and y a value taken by variable Y. The intensity of the influence of the variable Y on the variable X is quantified by the conditional probability $P(x|y)$, for every possible set of values (x, y).

In general, let $\mathbf{P}_X$ be the set of all parent nodes of a node X, $\mathbf{p}_X$ be a set of values for all the variables in $\mathbf{P}_X$, and x be a value of X. The influence of $\mathbf{P}_X$ on X can be modeled by any function $\mathcal{F}$ that satisfies the following conditions:

$$\sum_{x \in \mathbf{x}} \mathcal{F}(x, \mathbf{p}_X) = 1 \tag{2.3}$$

$$0 \leq \mathcal{F}(x, \mathbf{p}_X) \leq 1. \tag{2.4}$$

where $\mathbf{x}$ is the set of possible values for variable X. The function $\mathcal{F}(x, \mathbf{p}_X)$ provides a numerical quantification for the conditional probability $P(x|\mathbf{p}_X)$. Let $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ be the set of variables in a Bayesian network. The joint probability distribution over $\mathbf{X}$ is given by:

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|\mathbf{p}_{X_i}) \tag{2.5}$$

To illustrate, Figure 2.4 shows a Bayesian network for a joint probability distribution $P(x_1, x_2, x_3, x_4, x_5)$, where x_1, x_2, x_3, x_4, and x_5 refer to values of the random variables X_1, X_2, X_3, X_4, and X_5, respectively. Node X_1 is a node without parents and is called a *root node*. The probability $P(x_1)$ associated with a value x_1 of the root node X_1 is called a *prior probability* and can be used to represent previous knowledge of the modeled domain. By applying Equation (2.5), the joint probability distribution for the network shown in Figure 2.4 can be computed as:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_3)$$

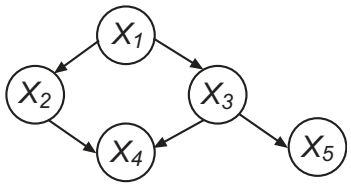


Figure 2.4: Example of a Bayesian network.

The most common task we wish to solve using Bayesian networks is probabilistic inference. The goal of the probabilistic inference is typically to find the conditional distribution of a subset of the variables, conditional on known values for some other subset (the evidence), and integrating over any other variables. In other words, given a piece of evidence, we can calculate the posterior probability of a possible explanation by applying the Bayes' rule:

$$P(r|e) = \frac{\sum_{\mathcal{U}-\{r\}} P(\mathcal{U}, e)}{P(e)} \tag{2.6}$$

where $P(r|e)$ denotes the probability that random variable R has value r given evidence e and $\mathcal{U}$ is a set representing the universe of variables in the

model. The denominator is just a normalizing constant that ensures the posterior probability adds up to 1. Notice that $P(\mathcal{U}, e)$ can be obtained through the application of Equation 2.5.

To illustrate this inference process, we now calculate the probability $P(w|x)$ for the Bayesian networks presented in Figure 2.5 and Figure 2.6. In these networks all the variables are binary, that is, they can assume only two possible values.

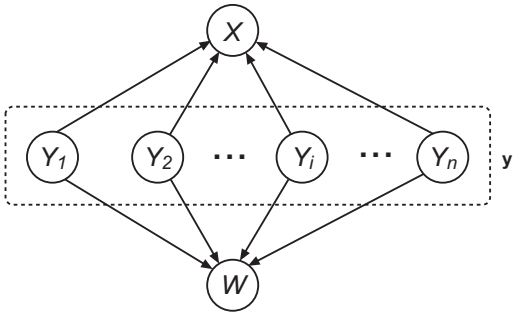


Figure 2.5: Example of a Bayesian network with several nodes represented as a set.

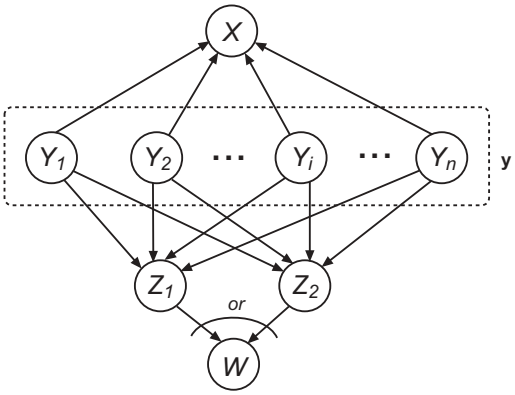


Figure 2.6: Example of a Bayesian network with a *noisy-OR* node.

By applying Equation (2.6) in the network of Figure 2.5, we obtain:

$$\begin{aligned} P(w|x) &= \frac{\sum_{\mathbf{y}} P(x, \mathbf{y}, w)}{P(x)} \\ &= \eta \sum_{\mathbf{y}} P(w|\mathbf{y})\ P(x|\mathbf{y})\ P(\mathbf{y}) \end{aligned} \tag{2.7}$$

where $\mathbf{y}$ is used to refer to any of the possible states of the root nodes Y_i and $\eta = \frac{1}{P(x)}$ is a normalizing constant. Equation (2.7) is particularly useful in this work. It will be used in the expansion technique to be described in Section 4.2.

Figure 2.6 provides another example for the inference process. However, differently from the network in Figure 2.5, the network in Figure 2.6 presents a new method for combining evidences, a *noisy-OR* node. In particular, the "or" mark above node W means that $P(W|Z_1, Z_2)$ is defined in such way that W is true if anyone of their parent nodes, Z_1 and Z_2, are true and W is false if both nodes Z_1 and Z_2 are false. In other words, $P(w|\overline{z}_1, \overline{z}_2) = 0$ and $P(w|z_1, z_2) = P(w|\overline{z}_1, z_2) = P(w|z_1, \overline{z}_2) = 1$, where z_i denotes that node

$Z_i = 1$ and $\overline{z}_i$ denotes that node $Z_i = 0$. We now calculate the probability $P(w|x)$ for this case.

$$\begin{aligned} P(w|x) &= \frac{\sum_{\mathbf{y},\mathbf{z}} P(x,\mathbf{y},\mathbf{z},w)}{P(x)} \\ &= \eta \sum_{\mathbf{y},\mathbf{z}} P(w|\mathbf{z})\ P(\mathbf{z}|\mathbf{y})\ P(x|\mathbf{y})\ P(\mathbf{y}) \\ &= \eta \sum_{\mathbf{y}} P(x|\mathbf{y})\ P(\mathbf{y})\ \sum_{\mathbf{z}} P(w|\mathbf{z})\ P(\mathbf{z}|\mathbf{y}) \\ &= \eta \sum_{\mathbf{y}} P(x|\mathbf{y})\ P(\mathbf{y})\ [P(z_1,\overline{z}_2|\mathbf{y}) + P(\overline{z}_1,z_2|\mathbf{y}) + P(z_1,z_2|\mathbf{y})] \\ &= \eta \sum_{\mathbf{y}} [1-(1-P(z_1|\mathbf{y}))(1-P(z_2|\mathbf{y}))]\ P(x|\mathbf{y})\ P(\mathbf{y}) \qquad (2.8) \end{aligned}$$

where $\mathbf{z}$ is used to refer to any of the possible states of nodes Z_1 and Z_2. Notice that Equation (2.8) will be used in the classification technique to be described in Section 5.2.2.

A key advantage of Bayesian networks is their synthesized representation of probabilistic relationships. In fact, it is necessary to consider only the known independencies among the variables in a domain, rather than specifying a complete joint probability distribution [24, 98]. The independencies declared at modeling time are then used to infer *beliefs* for all variables in the network. The inference mechanism, though exponential in the worst case, is efficient in many practical situations, particularly in those which arise in the IR arena.

2.3 The Vector Space Model

The vector space model is a simple and effective model for retrieving information from a document collection [109]. In it, documents and queries are represented as vectors in a space composed of index terms, i.e., words extracted from the text of the documents in the collection [137]. This vector representation allows us to use any vector algebra operation to compare queries and documents, or to compare a document to another one.

In the vector space model, with every term k_i in a document d_j is associated a weight w_{ij}. A document d_j is, thus, represented as a vector of term weights $\vec{d_j} = (w_{1j}, w_{2j}, ..., w_{tj})$, where t is the total number of distinct terms in the entire document collection. Each w_{ij} weight reflects the importance of term k_i in document d_j and is usually computed as:

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_i} \tag{2.9}$$

where tf_{ij} is the number of times the term k_i occurs in document d_j, n_i is the number of documents in which k_i occurs, and N is the total number of documents in the collection. The factor $\log(N/n_i)$ is called the *inverse document frequency* (IDF) and is used to stress the influence of terms that are more selective because they appear less frequently in the document collection. The expression for w_{ij} is usually referred to as *term frequency-inverse document frequency* (or TF-IDF) weight. Its foundations lie in the observation that a term is more important if it occurs many times in a document and less important if it occurs in many documents in the collection.

In the vector space model, users formulate their queries as sets of words. Thus, a query q also can be represented as a vector of term weights $\vec{q} = (w_{1q}, w_{2q}, ..., w_{tq})$. With this representation, we can use any vector related measure to compare a query with a document. The most commonly used measure is the so called *cosine similarity*, i.e., the cosine value of the angle between both vectors. Thus, we define the similarity between a document d_j and a query q as:

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \tag{2.10}$$

By computing the similarities between all the documents in a collection $\mathcal{D}$ and a given query q, we obtain an ordered set, where documents more likely to satisfy the query have higher similarities. To illustrate, Figure 2.7 shows the vectors corresponding to a document d_j and a query q, with terms k_a and k_b. The similarity between d_j and q is the cosine of angle θ.

In this work, we use the vector space model to obtain the information present in the contents of the documents and ads, both for ranking, in which documents are compared to ads to determine user's interests, as for document classification, where documents are compared to each other, to determine their similarity of topic.

2.4 Content-Based Classifiers

Document classification is the activity of assigning entries from a set of pre-specified categories to a document [110]. Given a set of categories $\mathcal{C} = \{c_1, c_2, ..., c_m\}$ and a set of documents $\mathcal{D} = \{d_1, d_2, ..., d_n\}$, the aim is to determine which elements of $\mathcal{C}$ can be assigned to each element of $\mathcal{D}$. Thus,

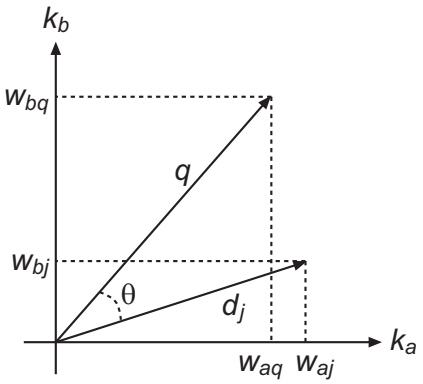


Figure 2.7: Representation of the similarity between document d_j and query q in the vector space model.

after classification, each document $d_i \in \mathcal{D}$ can be represented as a vector $\mathbf{p}_i = \{p_{i1}, p_{i2}, ..., p_{im}\}$, where p_{ij} indicates if category $c_j \in \mathcal{C}$ is assigned to document d_i. We call $\mathbf{p}_i$ the *category vector* (or *concept vector*) associated with d_i. The classification is called *hard* if p_{ij} is a boolean value, that is, the classifier output is an accept/reject decision. On the other hand, if p_{ij} can assume any value in the interval $[0, 1]$, the classification is called *soft*. This is the case of any classifier whose decisions can be interpreted as probabilities such that $\sum_{j=1}^{m} p_{ij} = 1$. To illustrate, let $\mathcal{C} = \{c_1, c_2, c_3\}$ and suppose a soft classification setting with $\mathbf{p}_i = \{0, 1, 1\}$. In this case, the classes of document d_i are c_2 and c_3. Now suppose a hard classification setting with $\mathbf{p}_i = \{0, 0.8, 0.2\}$. The likelihoods of d_i belonging to categories c_2 and c_3 are 0.8 and 0.2, respectively.

The main challenge in document classification comes from the fact that the rules to determine whether a document belongs or not to a category are not clearly defined. The most common solution consists of using machine learning algorithms, which are trained with a set of pre-classified documents, and use the knowledge thereby obtained to classify the whole collection.

Automatic classification algorithms are very useful due to the fact that there are many practical situations where the amount of data to be classified makes manual classification unfeasible. The most outstanding example is that of the World Wide Web, where we have millions of documents to be classified and where the data is commonly volatile, thus requiring frequent reclassifications. In the Web, the assignment of categories to documents is of great importance, since it can be used to improve tasks as ad placement by targeting users according to their topics of interest, to construct on-line

directories, to improve the precision of Web search engines, and even to help in the interactions between user and search systems [20, 127].

In this work, we use three well-known text classifiers: *kNN*, Naive Bayes, and Support Vector Machine (SVM). These methods have been extensively evaluated for text classification on reference collections and offer a strong baseline for comparison. We now briefly describe each of them.

2.4.1 The *kNN* Classifier

kNN, or *k nearest neighbors*, is a well known technique that has been widely studied in pattern recognition for over five decades [36]. It works by representing each data element to be classified as a point in an n-dimensional space. To assign a point to a category, its closest neighbors are examined and, in general terms, the object is classified under the category of the majority, as illustrated in Figure 2.8.

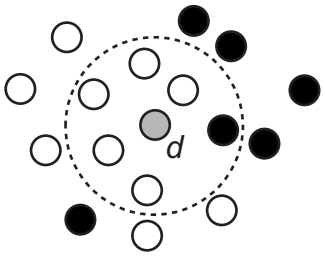


Figure 2.8: The *kNN* classifier. The class of point d is attributed according to the classes of its nearest neighbors. Points in black and white belong to distinct classes.

The most widely used *kNN* algorithm for IR was introduced by Yang in [140]. It assigns a category label to a test document based on the categories attributed to the k most similar documents in the training set. More specifically, to a given test document d is assigned a relevance score $s_{c_i,d}$ that associates d to the candidate category c_i. This score is defined as:

$$s_{c_i,d} = \sum_{d' \in \mathcal{N}_k(d)} sim(d, d') f(c_i, d') \tag{2.11}$$

where $\mathcal{N}_k(d)$ represents the k nearest neighbors of d, according to a given similarity function sim, and $f(c_i, d')$ is a function that returns 1 if document d' belongs to category c_i and 0 otherwise.

Traditionally, documents are represented by vectors of term weights and the similarity between two documents is measured by the cosine of the angle

between them, according to Equation 2.10. Term weights are computed using one of the conventional TF-IDF schemes [108], in which the weight of a term in a document is defined as in Equation (2.9). Based on the computed scores, we determine the top ranking category and assign it to the test document.

In this work, we used the *Bow* implementation of *kNN* [82], available at `http://www.cs.cmu.edu/~mccallum/bow`.

2.4.2 The Naive Bayes Classifier

The Naive Bayes method of classification uses Bayes theorem to determine the probability of a category, given the data to be classified. Its underlying assumption is that the attributes that characterize the data points are independent if their category is known, hence the name *naive* Bayes. In IR, Naive Bayes classification uses the probabilities of words and categories to estimate the probability of a category given a document [83]. It is assumed that the presence of each word in a document is independent of all words in the document, given its class.

Using this term independence hypothesis, the Naive Bayes classifier assumes that documents are generated from a distribution parameterized by θ. The likelihood of a document d_i being generated is defined as:

$$P(d_i|\theta) = \sum_{j=1}^{C} P(c_j|\theta)P(d_i|c_j,\theta) \tag{2.12}$$

where each c_j represents a class and C is the number of classes. Equation (2.12) states that a document is generated by selecting (1) one of the classes, with probability $P(c_j|\theta)$, and (2) a document from the class, with probability $P(d_i|c_j,\theta)$. The probability of selecting a class is defined by the proportion of documents belonging to the class, i.e.:

$$P(c_j|\theta) = \frac{\sum_{i=1}^{N} P(c_j|d_i)}{N} \tag{2.13}$$

where N is the total number of documents in the training set and $P(c_j|d_i)$ is defined as 1 if document d_i belongs to class c_j and 0 if otherwise.

In the model used in this work, each term in a document is seen as an event. Each term is represented by a random variable that takes a value from 1 to V, where V is the number of terms in the collection. A document of length n can be represented as a sequence of n term events, i.e., a sequence of n multinomial trials, where each trial is independent of the previous (due to the naive Bayes assumption). The probability of generating a document

d_i given a class c_j is, therefore, defined as the multinomial distribution:

$$P(d_i|c_j,\theta) = n_i! \sum_{k=1}^{V} \frac{P(t_k|c_j,\theta)^{f_{ki}}}{f_{ki}!} \tag{2.14}$$

where n_i is the length of document d_i and f_{ki} is the number of times term t_k occurs in document d_i. The probability of a term t_k in a class c_j is defined as:

$$P(t_k|c_j,\theta) = \frac{1 + \sum_{i=1} N f_{ki} P(c_j|d_i)}{V + \sum_{l=1} V \sum_{i=1} N f_{si} P(c_j|d_i)} \tag{2.15}$$

which is the proportion of occurrences of term t_k in the documents of class c_j, normalized to avoid probabilities of 0 or 1.

Finally, once all the parameters are learned, a document can be classified by computing the probability of a class given a document. This is accomplished by applying Bayes' rule, i.e.:

$$P(c_j|d_i,\theta) = \frac{P(c_j,d_i,\theta)}{P(d_i,\theta)} = \frac{P(c_j|\theta)P(d_i|c_j,\theta)}{P(d_i|\theta)} \tag{2.16}$$

To classify document d_i, the class that maximizes Equation (2.16) is chosen.

In this work, we used the *Bow* implementation of Naive Bayes [82], available at `http://www.cs.cmu.edu/~mccallum/bow`.

2.4.3 The SVM Classifier

SVM is a relatively new method of classification introduced by Vapnik in [132] and first used in text classification by Joachims in [60]. The method is defined over a vector space where the problem is to find a hyperplane with the maximal margin of separation between two classes. Classifying a document corresponds to determining its position relative to this hyperplane.

Figure 2.9 illustrates a space where points of different classes are linearly separable. The dashed line represents a possible hyperplane separating both classes. This hyperplane can be described by:

$$(\vec{w} \cdot \vec{x}) + b = 0, \tag{2.17}$$

where $\vec{x}$ is an arbitrary data point that represents the document to be classified and the vector $\vec{w}$ and the constant b are learned from a training set of linearly separable data. Classifying a vector is achieved by applying the decision function

$$f(\vec{x}) = \text{sign}((\vec{w} \cdot \vec{x}) + b) \tag{2.18}$$

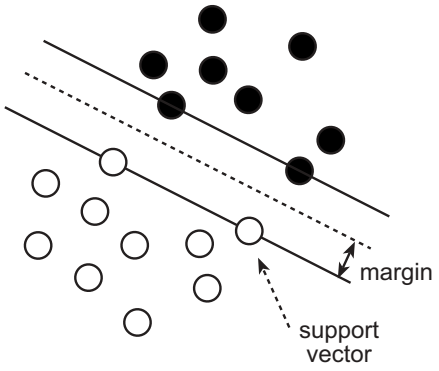


Figure 2.9: The SVM classifier. An separating hyperplane is found by maximizing the margin between the candidate hyperplane and the classes.

which determines the position of $\vec{x}$ relative to the hyperplane.

In Figure 2.9, the solid lines represent how much the hyperplane can be moved while still separating the classes. The SVM classifier tries to maximize the margin between the hyperplane and the points in the boundaries of each class. This is achieved by solving a constrained quadratic optimization problem. The solution can be found in terms of a subset of training patterns that lie in the marginal planes of the classes, the *support vectors*, and is of the form:

$$\vec{w} = \sum_i v_i \vec{x}_i \tag{2.19}$$

where each v_i is a learned parameter and each x_i is a support vector. The decision function can be, thus, written as:

$$f(\vec{x}) = \text{sign}(\sum_i v_i (\vec{x} \cdot \vec{x}_i) + b) \tag{2.20}$$

In the original data space, also called the *input space*, classes may not be separable by a hyperplane. However, the original data vectors can be mapped to a higher dimensional space, called the *feature space*, where classes are linearly separable. This is achieved through the use of *kernel functions*. Using kernel functions the optimization problem is solved in the feature space, instead of the input space, and the final decision function thus becomes:

$$f(\vec{x}) = \text{sign}(\sum_i v_i \kappa(\vec{x} \cdot \vec{x}_i) + b) \tag{2.21}$$

where κ is the kernel function.

Support vector machines only take binary decisions: a document belongs or not to a given class. In a multiple class setting, such as the one of this

work, a different classifier needs to be learned for each class. To make the final decision, each classifier can be compared to all the others and a voting scheme can be used in which the class of the classifier with the more votes is chosen [56].

In this work, we used the *LIBSVM* implementation of support vector machines [17], available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

2.5 Linkage Similarity Measures

In our classification experiments, we used five different similarity measures, derived from link structure, to determine a degree of similarity among Web pages: co-citation, bibliographic coupling, Amsler, authority degrees provided by the Companion algorithm, and hub degrees provided by the Companion algorithm. The first three were introduced in bibliometric science, to quantify the relationship between two scientific papers [2, 63, 118]. In this work, we evaluate how they perform when applied to the Web environment, where we assume that links between Web pages have the same role as citations between scientific papers. However, it is important to note that several distinctions must be made between the Web and the domain of scientific publications. Citations between scientific papers are commonly used to provide background information, give credit to the authors of an idea, discuss or criticize existing work, among others. Web links, on the other hand, can be seen as a generalized form of citation. Besides the same functionality, they are also used for advertising, in-site navigation, providing access to databases, among others. Moreover, authors of a Web page may refuse to reference other pages even if they are authorities on the same subject (for instance, pages from rival companies, like Sun and Microsoft). Also, links can be used to artificially increase the likelihood that a page be retrieved by a given search engine (spamming). These extra roles can make links a less reliable source of evidence, when used as an indicator of similarity between Web pages. However, several works have shown that there is enough functionality in common between links and citations to allow that links can be treated as citations. Authors in [120] present a detailed discussion on the role of Web links versus citations.

The Companion algorithm was proposed by Dean and Henzinger [30], as a method to find Web pages related to each other. Here, we use it to provide a value of similarity between documents. We now describe in detail each of the proposed linkage similarity measures.

2.5.1 Co-Citation

Co-citation was first proposed in [118] as a similarity measure between scientific papers. Two papers are co-cited if a third paper has citations to both of them. This reflects the assumption that the author of a scientific paper will cite only papers related to his own work. Although Web links have many differences from citations, we can assume that many of them have the same meaning, i.e., a Web page author will insert links to pages *related* to his own page. In this case, we can apply co-citation to Web documents by treating links as citations. As illustrated in Figure 2.10, we say that two pages are co-cited if a third page has links to both of them.

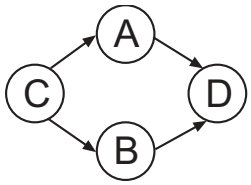


Figure 2.10: Pages A and B share one unit of co-citation, since they are both linked to by page C. They also share one unit of bibliographic coupling, since they both link to page D. By extent, they share two units of Amsler similarity, since they both link to page D and are both linked to by page C.

To further refine this idea, let d be a Web page and let P_d be the set of pages that link to d, called the *parents* of d. The co-citation similarity between two pages d_1 and d_2 is defined as:

$$cocitation(d_1, d_2) = \frac{|P_{d_1} \cap P_{d_2}|}{|P_{d_1} \cup P_{d_2}|} \tag{2.22}$$

Equation (2.22) tells us that, the more parents d_1 and d_2 have in common, the more related they are. This value is normalized by the total set of parents, so that the co-citation similarity varies between 0 and 1. If both P_{d_1} and P_{d_2} are empty, we define the co-citation similarity as zero.

2.5.2 Bibliographic Coupling

Also with the goal of determining the similarity between papers, Kessler [63] introduced the measure of bibliographic coupling. Two documents share one unit of bibliographic coupling if both cite a same paper. The idea is based on the notion that paper authors who work on the same subject tend to cite the same papers. As for co-citation, we can apply this principle to the Web. We assume that two authors of Web pages on the same subject tend to

insert links to the same pages. Thus, we say that two pages have one unit of bibliographic coupling between them if they link to the same page, as shown in Figure 2.10.

More formally, let d be a Web page. We define C_d as the set of pages that d links to, also called the *children* of d. Bibliographic coupling between two pages d_1 and d_2 is defined as:

$$bibcoupling(d_1, d_2) = \frac{|C_{d_1} \cap C_{d_2}|}{|C_{d_1} \cup C_{d_2}|} \tag{2.23}$$

According to (2.23), the more children in common page d_1 has with page d_2, the more related they are. This value is normalized by the total set of children, to fit between 0 and 1. If both C_{d_1} and C_{d_2} are empty, we define the bibliographic coupling similarity as zero.

2.5.3 Amsler

In an attempt to take the most advantage of the information available in citations between papers, Amsler [2] proposed a measure of similarity that combines both co-citation and bibliographic coupling. According to Amsler, two papers A and B are related if (1) A and B are cited by the same paper, (2) A and B cite the same paper, or (3) A cites a third paper C that cites B. As for the previous measures, we can apply the Amsler similarity measure to Web pages, replacing citations by links, as illustrated by Figure 2.10.

Let d be a Web page, let P_d be the set of parents of d, and let C_d be the set of children of d. The Amsler similarity between two pages d_1 and d_2 is defined as:

$$amsler(d_1, d_2) = \frac{|(P_{d_1} \cup C_{d_1}) \cap (P_{d_2} \cup C_{d_2})|}{|(P_{d_1} \cup C_{d_1}) \cup (P_{d_2} \cup C_{d_2})|} \tag{2.24}$$

Equation (2.24) tell us that, the more links (either parents or children) d_1 and d_2 have in common, the more they are related. The measure is normalized by the total number of links. If neither d_1 nor d_2 have any children or parents, the similarity is defined as zero.

2.5.4 Companion

On a different approach, the Companion algorithm was proposed by Dean and Henzinger in [30]. Given a Web page D, the algorithm finds a set of related pages by examining its link structure, and returns a degree of how related each page is to D. In this work, we proposed that this degree can be used as a similarity measure between D and the remaining pages.

To find a set of pages related to a page D, the Companion algorithm has two main steps:

1. build a Vicinity Graph of D, and
2. compute the degrees of similarity.

In step 1, pages that are linked to D are retrieved. We build the set $\mathcal{V}$, the vicinity of D, which contains the parents of D, the children of the parents of D, the children of D, and the parents of the children of D. This is the set of pages related to D. Let $\mathcal{E}$ be the set of links among the pages in $\mathcal{V}$. The pages in $\mathcal{V}$ and the links in $\mathcal{E}$ constitute a graph $G = (\mathcal{V}, \mathcal{E})$ called the vicinity graph of D, as illustrated in Figure 2.11.

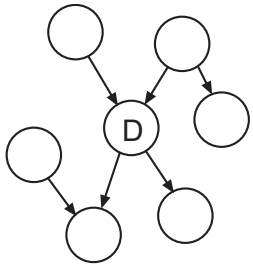


Figure 2.11: Vicinity graph of page D.

In step 2 we compute the degree to which the pages in $\mathcal{V}$ are related to D. To do this, we consider the pages in G and calculate their authority and hub values. In particular, the authority value of a document p is defined as:

$$a(p) = \sum_{d|(d,p)\in\mathcal{E}} h(d) \tag{2.25}$$

Recursively, the hub value of a document p is defined as:

$$h(p) = \sum_{d|(p,d)\in\mathcal{E}} a(d) \tag{2.26}$$

Thus, a document is considered a good authority if it is linked by many good hubs. Conversely, a document is considered a good hub if it links to many good authorities. These definitions for authority and hub were first presented by Kleinberg in an algorithm called HITS [40, 64]. Notice that we can use the degree of authority or hub (or a combination of both) as a measure of similarity between D and each page in $\mathcal{V}$. In this work we experimented with the Companion algorithm using either the authority or the hub degree in isolation as a similarity measure. We define the similarity between D and any page that is not in $\mathcal{V}$ as zero.

2.6 Evaluation Measures

In this section, we introduce the metrics which we will use to evaluate the models and results shown in the following chapters.

2.6.1 Precision and Recall

To evaluate the ad ranking results we use the classical IR evaluation metrics *precision* and *recall* [5, 78]. We use these metrics considering that, for each triggering page, a set of relevant ads has been defined. These relevance judgments are accomplished through human evaluations, made by a group of system users.

Given a triggering page T and a set R of relevant ads for T, precision and recall figures can be used to evaluate the quality of a retrieval method. Using the retrieval method, we obtain a set A of ads to place in page T. This set is then compared to the set R of relevant ads. The higher the overlap between them, the better is considered the result. Precision and recall are defined as a means to characterize this overlap, as follows.

Precision, p, is the fraction of all answers in A that are correct, i.e.:

$$p = \frac{|A \cap R|}{|A|}$$

Precision is defined 1 if no ads were retrieved, i.e., if $|A| = 0$.

Recall, r, is the fraction of correct answers that were properly retrieved in A, i.e.:

$$r = \frac{|A \cap R|}{|R|}$$

Recall is defined as 1 if there are no relevant ads, i.e., if $|R| = 0$.

Frequently, we want to evaluate average precision at given recall levels. The standard 10-point average precision measure returns precision at 10%, 20%, ..., 100% of recall. For instance, precision at 10% recall is the precision when 10% of the relevant ads in the set R have been seen in the ranking, starting from the top. Average precision at 10% recall is the average precision for all test triggering pages, taken at 10% recall. Plotting the precision at the 10 standard recall points allows us to easily evaluate and compare the quality of ranking algorithms.

Finally, as single value summaries of the performance of the ranking algorithms we use the precision at the first k top ranked ads [5], P@k, and two versions of the non-interpolated average precision [51], PAVG and PAVG@k.

Measure P@k is defined as:

$$P@k = \frac{\sum_{j=1}^{k} rs(a_j)}{k} \tag{2.27}$$

where k is the number of ads to be displayed in a page, a_j is the j-th top ranked ad and $rs(a_j) \in \{0, 1\}$ is the relevance score assigned to a_j, being 1 if a_j is relevant and 0 otherwise. Notice that this measure takes into account the number of relevant ads in the top positions but not the order in which they appear. As this is an important element for the ranking quality assessment we also evaluate our methods using PAVG, a measure commonly used in TREC evaluations [51]. It is defined as:

$$PAVG = \frac{1}{|R|} \sum_{i=1}^{|D|} \left(rs(a_i) \times \left(\frac{\sum_{j=1}^{i} rs(a_j)}{i} \right) \right) \tag{2.28}$$

where $|R|$ is the number of relevant ads in set R and $|D|$ is the number of retrieved ads. Finally, given the importance of the k top positions in the content-targeted advertising problem, we also report our results in PAVG@k, a slightly modified version of PAVG, defined as:

$$PAVG@k = \frac{1}{k} \sum_{i=1}^{k} \left(rs(a_i) \times \left(\frac{\sum_{j=1}^{i} rs(a_j)}{i} \right) \right) \tag{2.29}$$

In this metric, $|R|$ and $|D|$ are substituted by k. Thus, a ranking function which places relevant ads in all the k top ad slots of the triggering page will receive the maximum PAVG@k value equal to 1.

2.6.2 The F-measure

In classification tasks precision and recall are taken for every class. This yields a great number of values, making the tasks of comparing and evaluating algorithms more difficult. In these cases, it is often convenient to combine precision and recall into a single quality measure. One of the most commonly used such measures is the *F-measure* [141].

The F-measure combines precision and recall values and allows the assignment of different weights to each of these measures. It is defined as:

$$F_\alpha = \frac{(\alpha^2 + 1)pr}{\alpha^2 p + r} \tag{2.30}$$

where α defines the relative importance of precision and recall. When $\alpha = 0$, only precision is considered. When $\alpha = \infty$, only recall is considered. When $\alpha = 0.5$, recall is half as important as precision, and so on.

In our classification experiments, we assign equal weights to precision and recall by defining $\alpha = 1$. This yields the so called F_1 measure, defined as:

$$F_1 = \frac{2rp}{p + r} \tag{2.31}$$

The F_1 measure allows us to conveniently analyze the performance of the classification algorithms used in our experiments on each of the used classes, and can also be averaged over the classes to quantify overall retrieval performance.

Chapter 3

Related Work

In Section 3.1, we review previous research done in search advertising. In Section 3.2, we also review previous work on the use of cross-referencing information in bibliometric science and in IR. Since our proposed models use Bayesian networks as a formal basis, important works related to the use of Bayesian networks in IR are described in Section 3.3. As some of our methods are based on the technique of query expansion, we describe some works on query expansion in Section 3.4. Finally, in Section 3.5, we present some works which have used automatic classification as a means of improving Web search effectiveness.

3.1 Search Advertising

In this section, we present a broad view of the research in search advertising. In particular, we review works on relevance matching, ranking, fraud detection, feedback tools, and implementation design. Notice that we do not cover several other topics related to search advertising such as privacy issues [19] and ethical/social issues [9, 58, 90]. Since there is little academic research on these subjects [79, 136], our study will rely largely on research conduced by advertising agencies and consulting companies.

3.1.1 Relevance Matching

Many works in advertising research have stressed the importance of relevant associations for consumers [79,84,129]. For example, the studies in [97] point out that the user perception of Web content can be more positive whenever there is a strong relationship between that content and the advertised products. According to the authors in [79], even forced exposure is considered

less intrusive if it is editorially congruent. Studies in [133] show that by ensuring the relevance of the ads, information gatekeepers are investing in the reinforcement of a positive user attitude towards the advertisers and their ads. The authors in [34] also notice that irrelevant or offensive ads can turn off users and relevant ads are more likely to be clicked on, which leads to better performance since click-through is a key element in evaluation. The results in [91] enforce this conclusion by pointing out that the more targeted the advertising, the more effective it is.

Therefore it is not surprising that other works have addressed the relevance issue. For instance, in [70] it is proposed a system called ADWIZ that is able to adapt online advertising to the short-term interests of a user's in a non-intrusive way. Contrary to our work, ADWIZ does not directly use the content of the page viewed by the user. It relies on search keywords supplied by the user to search engines and on the URL of the page requested by the user. On the other hand, in [66] the authors presented an intrusive approach in which an agent sits between advertisers and the user's browser allowing a banner to be placed into the currently viewed page. In spite of having the opportunity to use the page's content, the agent infers relevance based on category information and user's private information collected along the time.

In fact, sophisticated matching strategies have been developed to ensure that relevant ads will be shown to the users. This implies considering exact and approximate matching methods. In the exact method, an ad keyword and a user query are matched if they are identical. This excludes, for instance, the case in which the keyword is *part* of the query. In the approximate method, the keywords are matched to the user query regardless of the order and of the distance of their component words. Further the words can be automatically expanded to include synonyms, related terms and plural forms. In many systems, the advertisers can still define terms that should not occur in the user query or even define sub-phrases that should occur in it [49, 81].

Studies on keyword matching have also shown that the nature and size of the keywords have impact on the likelihood of an ad to be clicked [62, 112]. These studies were motivated by the following intuition. A user in a self-education process could start a search by employing short and generic keywords. In a later phase of his buying cycle, however, he tends to be more focused. Consequently, his query evolves to include more words and these words are more specific. In such phase, it is common the use of brand names. To illustrate, a query that starts as "camera" can evolve to "canon powershot a95". Thus, conversion rates probably tend to increase as the number of words in a search query increases. In fact, in order to test this intuition, the authors in [93] tracked conversion rates per query in 2004. They found that, discounted the single word keywords, conversion rates peak at four-

word keywords, dropping with larger keywords. In the case of single word keywords (keyword size = 1), they found very high conversion rates when brand names were considered. However, after removing these brand names, the conversion rates for single word keywords dropped to values lower than those obtained by keywords with size 2 to 4 words. This indicates that people searching for a particular company's name are more predisposed to make a conversion. In general, results can be explained by the fact that large keywords and corporate names tend to be less ambiguous, which leads to a more precise matching. Thus, these keywords could indicate a more qualified user, that is, a user in buying mode [112].

On a different approach, the authors in [126] proposed a system that learns how to extract keywords from triggering pages. Their system is initially trained with a set of example pages that have been hand-labeled with relevant keywords. Based on this training, it can extract new keywords from previous unseen pages. They considered different approaches to deal with phrases and a number of features on the learning. Examples of such features are the term frequency, IDF, frequency in meta-data and in search query logs. They found large improvements over a simple extraction approach based on the TFIDF weighting scheme and observed that the several features tested were helpful. Surprisingly, the best feature was the frequency in the query log. They also tested their strategy to extract keywords from emails with similar results [45].

Other evidence sources have been used as a means of improving precision in search advertising matching systems, such as, for example, geographic information [72, 73]. With this information, it is possible to advertise products or services in user neighborhood. Further, by means of geographic databases, they are able to recognize that some terms in the keywords represent specific locations. In some cases, the advertisers can explicitly target specific countries and languages [72]. This is particularly interesting for products and services whose scope is geographically restricted as, for example, movie rent services and sales promotions.

Finally, it is important to notice that there are situations in which even relevant association between ads and page contents could lead to improper advertising [55, 123]. For example, it is possible that an ad could be relevant to a page's content and damaging to the brand. This is the case of pages about catastrophes. Such pages hardly offer good opportunities to the announcement of products and services. As a consequence, matching systems normally have to be able to recognize and filter these contents. Such systems will also try to recognize unethical and illegal advertising to avoid placing ads of competitors, violating trademarks, and supporting pornography, drugs, and gambling games. For that, brokers normally use filtering

lists [87, 96, 121]. In many cases, however, manual editorial control is employed [34].

3.1.2 Ranking

After the matching system finds the key topics of a Web page, a ranking system will try to maximize the utility of the search network to all its participants. For this, it has to satisfy different interests [71]: (a) the users want to receive relevant information, (b) the advertisers want to receive quality traffic at a minimum cost and with a minimum risk of negative user attitude towards them, and (c) the brokers and publishers want to maximize their revenues at the minimum risk of negative user attitude towards their brands, contents, and services.

A good ranking system has to meet these interests in a fair way in order to be useful to all participants. For example, as pointed by [6], higher is the rank position in a paid list, higher are the click-through rates received by the advertiser and, by extension, higher are the revenues received by the brokers. This could lead to a very simple ranking strategy where user interests are not taken into account. In this ranking, paid list positions would be defined only by how much advertisers pay. However, one such ranking can lead the users to perceive the whole service as unreliable. That is, in the long run, it can diminish users' loyalty towards the publisher. Thus, such biased ranking can lead to less traffic. A similar result has been observed in the case of information gatekeepers, such as search engines, that favor pages in their ranking that carry their ads. They risk their credibility with severe consequences in the long term [124]. These examples clearly show the necessity of fair and clever ranking algorithms.

A detailed study on such ranking algorithms is provided in [34]. In that work, the authors analyze paid placement strategies for keyword-targeted advertising. In particular, they study stylized versions of the ranking strategies employed by Google and Overture. Their comparison ignores, however, many factors present in real systems, including editorial control, inexact query matching, new pricing models like payment per conversion, marketing efforts, brand awareness, legal controls, and strategic alliances.

From the strategies considered, Google's ranking performed best in almost all cases. This corroborates the intuition that click-through rates, used in Google technique, can be taken as a useful relevance judgement about the ads. To evaluate the impact of the editorial control over the ranking strategies, a variation of Overture's strategy was developed to simulate the filtering of irrelevant or objectionable ads before the ranking process. This new strategy was significantly better than the original. This suggests that a careful review

of the mappings made by the ranking algorithms could be worthy if it ensures that ads are properly targeted [116, 123]. As pointed out by [34], brokers can therefore choose a suitable level of investment in the editorial process by trading the costs with the consequent increases in revenues. The authors also studied the dynamic nature of the strategies. As a result, they suggest that ranking algorithms should employ a revision mechanism, in which the reward for a click is larger if it is received in a lower rank. Such clicks indicate situations in which it is more probable that the ranking algorithm has failed.

Finally, they conducted experiments to evaluate how many ads should be placed in a paid list. They observed that the optimal revenue values were obtained for lists with sizes varying from 3 to 7 ads. This is due to the tradeoff between direct revenue increases and indirect revenue losses due to consumer defection. This tradeoff can be explained as follows. Intuitively, the more ads a publisher shows, the more revenue it will receive from advertisers. However, large paid lists are more likely to enroll irrelevant ads. This has a negative impact on the overall quality of the publisher. Consequently, total traffic of the publisher and click-through rates will reduce lowering revenue from paid placement.

The authors in [85] have studied this same ranking problem considering an additional restriction, the fact that advertisers have different daily budgets. The authors showed that this problem is clearly a generalization of the online bipartite matching problem and proposed new algorithms to solve it.

3.1.3 Fraud Detection

The revenues in search advertising are directly associated with the user traffic in the network. Thus, the more publishers have users clicking in the ads shown in their pages, the more advertisers will pay to them. Clearly, there is a potential for frauds since publishers could simulate that traffic [33, 79, 88]. This is in fact a problem in the industry. For example, CompUSA[1] spent more than U$10 millions in 2004 due to fake traffic [79]. This is serious because advertisers could lose the confidence in the network, which could profoundly harm all the business [62].

To deal with this, research has concentrated on trying to characterize the fake traffic. For that, common strategies consist in analyzing the distribution of the clicks along the time. If an unusual number of clicks from a same user or group of users is detected in a certain window of time, or a pattern of invalid clicks is found, these clicks are considered illegal and are ignored. A more sophisticated treatment for the problem was suggested by the author

[1] `http://www.compusa.com`

in [33]. It considers this a classification problem in which real and fake traffic have to be distinguished with maximum precision. This is necessary to protect advertisers from paying for excess clicks and, at same time, to avoid penalizing the broker by discharging valid clicks. The author suggests the use of unlabeled data in the training phase because a huge amount of clicks is generated continuously and it would be impossible to label all of them. However, as suggested by [3], probably the only effective way to avoid fake traffic is by moving from pay-per-click to pay-per-sale or pay-per-lead programs. To show this, the authors presented a technique to yield fake traffic that seems to be virtually undetectable since that, given today's Web infrastructure, it is not possible to achieve sufficient auditability to address the problem.

3.1.4 Feedback Information

In search advertising, advertisers have the possibility of getting detailed feedback about their performance. This can be used to determine how much they have to pay for the received traffic and what can be done to improve their campaigns. To help advertisers in such tasks, research, metrics, and tools have been developed. In fact, studies have shown that a careful analysis of feedback information is rewarding for advertisers [59,62]. In general, companies that excel on biding, selecting keywords, selecting ad text, and preparing landing pages obtain higher revenues with search advertising.

Advertisers know that clicks will not necessarily be converted. Thus, they are also concerned with aspects such as brand awareness and perception, and are always pressuring for more detailed information that can provide them with a reliable estimate of their return of investments (ROI). To meet this demand, brokers and third-party companies have made available several tools to help advertisers analyze their conversion rates and improve their campaigns accordingly [48]. These tools normally include tracking mechanisms that work in the advertiser's site. With these tools, advertisers can track users during their buying cycles, which makes possible to estimate impact of brand awareness, to determine which actions the user takes, how many leads, orders or sales transactions are generated, etc [49].

Using such tools, advertisers can also infer the more profitable keywords and auction strategies. For that, brokers can provide the advertisers with detailed information on their bids that includes, in some cases, demographic information about the users. For example, the MSN[2] adCenter system is able to inform the gender, age, lifestyle, and income of the users who search

[2] `http://www.msn.com`

for a certain keyword [77]. In fact, a whole industry has flourished by selling consultancy services related to tasks as keyword selection and definition of auction strategies. For an example of such tools, consider a keyword suggestion tool that, for any given keyword, provides a sorted list of correlated keywords and suggests them to the advertiser.

Notice that by identifying a cluster to which a certain keyword belongs, it is possible to take the other keywords in that cluster as suggestions. The authors in [14] evaluated two clustering methods for determining (a) groups of keywords belonging to the same marketplace and (b) submarkets of advertisers showing common bidding behavior. The first clustering method was based on the idea that advertisers with common interests will bid on the same subset of keywords forming a submarket. Thus, the problem of finding these submarkets can be approached through partitioning a graph to find strongly connected subgraphs. For this, they employed a flow-based graph partitioning method. Their second clustering method was based on the idea that related keywords present similar bidding patterns. Thus, these keywords can be represented as graph nodes connected by edges whose weights are proportional to the amount of overlapping between their set of bidders. Their method consisted in clustering these nodes using an agglomerative technique. After comparing the proposed methods against other commonly employed in the literature, they found that their first approach is better for providing a small number of larger clusters while their second approach is better for providing a large number of small clusters.

The authors in [42] also presented a keyword suggestion tool. As they were interested in controlling the level of generality of the keywords suggested, they used a vector space model with a singular value decomposition (SVD) approach. Differently from the methods previously described, their strategy allows each keyword to potentially form soft clustering of related keywords. In their method, keywords are represented as vectors of advertisers. A nonzero entry in these vectors corresponds to an advertiser bidding on the keyword. The similarity between keywords is calculated as the cosine of the angle between their corresponding vectors. By projecting these vectors in an SVD subspace, they can perform a conceptual match instead of a simple exact match. In other words, they can match keywords globally or conceptually without the need of explicit bidding associations.

3.1.5 Implementation Design

Unlike the traditional IR search model, a search advertising model is characterized by dynamic ad collections and ranking functions based on parameters that vary dynamically and must be updated frequently. This makes it hard

the employment of query optimization techniques such as those based on index pruning [100]. To cope with these search advertising characteristics, authors in [4] proposed an architecture in which the employed optimization techniques aim to achieve efficient query search, incremental ad updates, and dynamic ranking. Efficient query search is essential particularly in keyword-targeted advertising where the ads have to be shown along with the results of the user query. The possibility of incremental ad updates makes it possible on-line modification of the ad collection. Finally, a dynamic ranking is necessary to model the competition between advertisers continuously updating the ranking parameters that they control. Figure 3.1 illustrates their architecture.

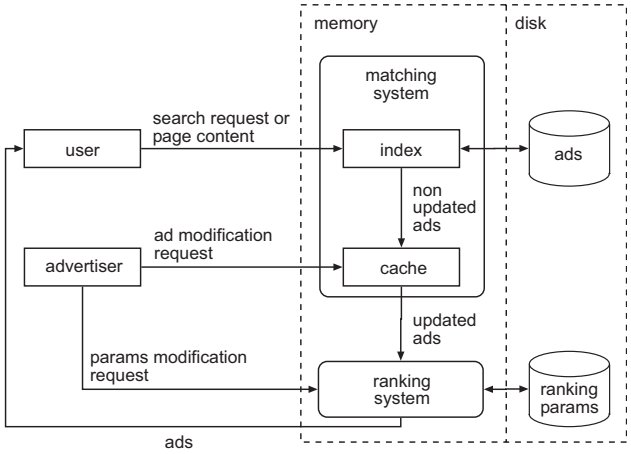


Figure 3.1: Search advertising system

As we can see, ads and their associated selection criteria (e.g., keywords) are stored in a collection as well as the parameters related to each advertiser. For each user request all ads that give a positive match are selected and ordered according to a ranking measure. A positive matching indicates that the ad is relevant to the user request. The ranking measure indicates how useful is the match for the actors in the search advertising network. Advertisers can modify their ads and the ranking parameters that they control such as, for example, the amount that they are willing to pay by a user click given a certain query. A caching system is used to support real time updates of queries and ranking parameters.

3.2 Cross-referencing Information

Cross-referencing information was first used in bibliometric science. In particular, citations among scientific papers were used both to find papers on related topics and to measure the importance of a publication. Similarly, since in the Web, links among the pages can take the role of cross-references, they can be used to find pages on related topics and to evaluate their importance. In this section, we review some works in which such cross-referencing information have been studied and applied.

In 1963, the authors in [63] introduced the notion of bibliographic coupling (see Section 2.5.2). This measure can be used to determine documents with similar topics. For this reason, bibliographic coupling was also used to cluster scientific journals [119]. Later, the measure of co-citation was introduced in [118] (see Section 2.5.1). Co-citation and bibliographic coupling have been used as complementary sources of information for document retrieval and classification [2,7]. Citations also were suggested as a means to evaluate the importance of scientific journals [39], where the importance of a journal was assumed proportional to the number of citations to its papers.

In [107], Salton introduced the idea of using citations for automatic document retrieval. Later works have presented hypertext retrieval systems capable of using citations, or any other type of link, as a complement to full-text searching [8, 23, 37]. Taking advantage of the network formed by documents and the links between them, the authors in [26] used spreading activation techniques for hypertext retrieval. The authors in [27] have also suggested the use of Bayesian networks as a modeling framework for hypertext. More recently, citations have been used to index and retrieve scientific papers published in the Web [41].

The ideas used for citations among documents can be transposed to the Web environment. The application of bibliometric techniques to the Web, and the necessary adaptation of these techniques to the new context, has given rise to algorithms for improving retrieval performance in the Web, such as PageRank and HITS. PageRank was proposed by the authors in [10] based in the idea that a page is considered popular if it is linked by many other pages. In fact, a page is recursively defined as popular when it is linked to by many other popular pages. HITS (Hyperlink-Induced Topic Search) was introduced by the authors in [40,64]. In HITS, pages assume two distinct roles: hubs and authorities. An authority is a page that contains important information on a given subject. A hub is a page that may not have relevant information, but links to many authority pages. Thus, a good hub page links to many good authority pages and, recursively, a good authority page is linked to by many good hub pages.

3.2.1 Link Information for Web Classification

Link information has been previously proposed as a way of finding Web documents related to a same topic. The Companion algorithm [30], for instance, uses links to determine pages related to a given initial page. Its functionality is briefly described in Section 2.5.4. Similarly, the authors in [54] propose a clustering algorithm that groups Web pages by operating on the graph defined by their link structure. Co-citation and text similarity measures are used to assign weights to the edges of the graph and partitioning algorithms are used to split the set of pages into clusters. In [128], three measures of linkage similarity are compared to a human evaluation of similarity between Web pages. However, the authors come to conclusions quite different from our own, mainly due to the collection used—a set of academic sites from the U.K. This collection has a very different link structure where, for instance, many of the pages link to each other, a phenomena that we cannot expect in in the Web in general [68].

Differently from simply finding related documents, several other works in the literature have reported the successful use of links as a means to improve classification performance. Using the taxonomy presented in [125], we can summarize these efforts in three main approaches: hypertext, link analysis, and neighborhood.

In the hypertext approach, Web pages are represented by context features, such as terms extracted from linked pages, anchor text describing the links, paragraphs surrounding the links, and the headlines that structurally precede them. In [38], [44] and [125], it was achieved good results by using anchor text together with the paragraphs and headlines that surround the links, whereas the authors in [142] show that the use of terms from linked documents works better when neighboring documents are all in the same class.

In the link analysis approach, learning algorithms are applied to handle both the text components in Web pages and the linkage between them. The authors in [117] exploit the hyperlink topology using a HITS based algorithm [64] to discover test set regularities. In [61], it was studied the combination of support vector machine kernel functions representing co-citation and content information. By using a combination of link-based and text-based probabilistic methods, the authors in [22] improved classification performance over a text-based baseline. Authors in [35] extended this work by showing that link information is useful when the document collection has a high link density and most links are of high quality.

Finally, in the neighborhood approach, the document category is estimated based on category assignments of already classified neighboring pages. The algorithm proposed in [15] uses the known classes of training documents

to estimate the class of the neighboring test documents. Their work shows that co-citation based strategies are better than those using immediate neighbors. The authors in [92] improved on this work by using a filtering process to further refine the set of linked documents to be used.

The classification method presented in this work mixes the link analysis and neighborhood approaches, differing from previous works in two main issues. First, we analyze several distinct link-based similarity measures and determine which ones provide the best results in predicting the category of a document. Second, we evaluate how effective these measures are in improving the results of a text-based classifier. This is achieved by combining link-based and text-based information through a Bayesian network model.

3.3 Bayesian Networks in IR

Bayesian networks, introduced by Pearl in [99], provide a graphical formalism for explicitly representing independencies among the variables of a joint probability distribution. They were first used in IR problems in the *Inference Network Model* [130, 131]. In that model, index terms, documents and user queries are seen as events and are represented as nodes in a Bayesian network. The model takes the viewpoint that the observation of a document induces belief on its set of index terms, and that specification of such terms induces belief in a user query or information need. This model was shown to perform better than traditional probabilistic models and used to effectively combine different sources of information for the task of document ranking. Later, a second model was proposed in [104], where the elements of an IR system are formally defined as concepts in a sample space. Their work not only provides a probabilistic justification for the model, but also demonstrates that the combination of evidence from past queries with evidence from the vector space model yields better results than the use of a vector ranking alone. More recently, the authors in [1] presented a third model whose network topology is defined in such way that an exact propagation algorithm, also proposed in their work, can be used to efficiently compute the relevance probabilities of the documents. When compared to the Inference Network Model for the task of document ranking, it shows better performance in four out of five reference collections.

Bayesian networks have also been applied to other IR problems besides ranking as, for instance, relevance feedback [50], automatic construction of hypertext [114], query expansion [28], information filtering [13], assigning structure to database queries [12], document clustering and classification [32], retrieval of juridical information [29], and retrieval of images from

the Web [21].

In this work, we adopt the Bayesian framework proposed by Ribeiro-Neto and Muntz in [104] for modeling a new representation for a triggering page and for combining content-based and link-based information.

3.4 Query Expansion

As it can be assumed that query terms are useful to identify relevant documents, expansion terms that are closely related to the query terms should be useful for ranking. This idea has motivated the research of several approaches on query expansion. Such approaches can be classified into Global Analysis and Local Analysis [5]. Global analysis consists in obtaining corpus-wide statistics, such as co-occurrence between pairs of terms in the whole collection, and using them to expand the original query. A common problem with this approach is the necessity of considerable computation resources to obtain corpus-wide statistics. Examples of this approach are Latent Semantic Analysis [31], the use of similarity thesauri [101], and term clustering, where queries are expanded by terms that appear in the same clusters of the query terms. Local Analysis, by other hand, uses only some initially retrieved documents for further query expansion. A well-known local analysis technique is Relevance Feedback [106]. In this technique, the query is modified based on relevance judgements provided by the users. Unfortunately, in real search context, users are reluctant to provide such feedback information. As a result, other methods were proposed, such as Pseudo-relevance Feedback [11]. In this method, it is assumed that top ranked documents are relevant. Thus, the terms in these documents are used to modify the original query.

An example of a pseudo-relevance feedback technique is presented in [16], which focuses on improving search engine results in a TREC collection by means of a method based on *kNN* [140]. Such method resembles our expansion approach to be described in section 4.2. Our method, however, is different from that presented by [16]. They expand user queries applied to a document collection with terms extracted from the top k documents returned as answer to the query in the *same* collection. In our case, we use two collections: an ad and a Web collection. We expand the triggering pages with terms extracted from the Web collection using statistics of the Web collection. We then match these expanded pages to the ads from the ad collection using statistics of the ad collection. By doing this, we emphasize the main topics of the triggering pages, increasing the possibility of associating relevant ads with them.

3.5 Search using Automatic Classification

Previous work in literature have exploited automatic classification to improve search effectiveness. For the Web, in particular, classification has been used mainly to contextualize queries, enhance link analysis, and reduce the search space.

Contextualization methods deal with the high degree of redundancy found in Web search queries. The main idea of these methods is to modify the query by using terms related to the query category. For instance, Inquirus2 [43] uses entropy loss to determine new terms for the query whereas Keyword spices [94] and TAX-PQ [95] use decisions trees to find new terms in documents classified according to hierarchical and flat taxonomies. Our conceptual matching method presented in Chapter 5 differs from these works in that we neither modify the triggering page nor the ads.

Classification has also been used to improve link analysis methods, in particular, PageRank. In PageRank, the relative importance of Web pages is calculated as a unique ranking vector independent of any particular search query. As a consequence, heavily linked pages can be highly ranked for queries for which they have no particular authority. To avoid this problem, the authors in [52] proposed a method called *topic sensitive PageRank*, in which a set of category-specific ranking vectors is computed instead of just a generic one. The category-sensitive scores for pages satisfying the query are computed using the category of the query keywords. Our conceptual method differs from this in that we use category information to reorder the ranking directly whereas topic sensitive PageRank uses the information to select the most appropriate link-based ranking.

Finally, category-based methods used for reducing the search space attempt to disambiguate queries (or query terms) by providing information about their conceptual frame. In [18] it was proposed a system which initially retrieves documents that match the query. After that, the documents are filtered out to ensure that only those that were classified under one of the query categories will be shown. Similarly, the authors in [102] proposed a search engine that retrieves documents based on a combination of keyword and conceptual matching. Notice that this idea is closely related to ours and can be viewed as a generalization of the previous approach. The authors in both works used automatic classification and found rather modest precision increases. In particular, the authors in [102] attributed this result to a very restrictive definition of relevance. They also reported that gains were better for short queries and that keyword matching is far more important than concept matching. As we will see in Chapter 5, we came to different conclusions and studied different scenarios due to intrinsic characteristics of

the content targeted advertising such as a broader definition of relevance and the possibility of dealing with manual classification.

Chapter 4

Syntactical Matching Strategies

In this chapter, we propose new strategies for associating ads with Web pages in content-targeted advertising. Five of these strategies are referred to as *simple matching strategies*. They are based on the idea of matching the text of the Web page directly to the text of the ads and its associated keywords. Five other strategies, which we here introduce, are referred to as *impedance coupling strategies*. They are based on the idea of expanding the Web page with new terms to facilitate the task of matching ads and Web pages. This is motivated by the observation that there is frequently a mismatch of the vocabulary of a Web page with the vocabulary of an ad. We say that there is a *vocabulary impedance problem* and that our technique provides a positive effect of *impedance coupling* by reducing the vocabulary impedance. Further, all our strategies rely on information that is already available to brokers that operate keyword targeted advertising systems. Thus, no other data from the advertiser is required.

The chapter is organized as follows. In Section 4.1, we introduce five simple matching strategies to solve the content-targeted advertising problem. In Section 4.2, we present our impedance coupling strategies. In Section 4.3, we describe our experimental methodology and datasets and discuss our results. Finally, in Section 4.4 we present our conclusions.

4.1 Simple Matching Strategies

In this work we are particularly interested in the relevance aspect of the content-targeted advertising. Thus, given a Web collection $\mathcal{D}$ and a set of ads $\mathcal{A}$, our task is to select ads $a_i \in \mathcal{A}$ related to the contents of a Web page $p \in \mathcal{D}$ and rank them according to how relevant they are. The ad list is then built in such way that more relevant ads are placed in top positions.

We consider that an ad a_i is composed of a title, a textual description, a set of keywords, and a hyperlink, as described in Section 2.1.1. Formally, an ad $a_i \in \mathcal{A}$ can be defined as a tuple as given by Equation 4.1:

$$a_i = \langle \vec{c_i}, \mathcal{K}_i, \vec{h_i} \rangle \tag{4.1}$$

where $\vec{c_i}$ represents the ad creative, that is, its title and description; $\mathcal{K}_i = \{\vec{k_1}, \vec{k_2}, ..., \vec{k_n}\}$ represents the set of keywords associated with a_i and $\vec{h_i}$ represents the contents of the landing page pointed to by a_i.

A simple way of ranking $a_i \in \mathcal{A}$ with regard to p is by matching the contents of p to the contents of a_i, i.e., its creative $\vec{c_i}$. For this, we use the vector space model (see Section 2.3). By considering p as the query and a_i as the document, we can rank the ads with regard to the Web page p. This is our first matching strategy. It is represented by the function AD given by:

$$\texttt{AD}(a_i, p) = sim(c_i, p)$$

where AD stands for "direct match of the ad, composed by title and description" and $sim(c_i, p)$ is computed according to Equation (2.10). In this case, the k first ranked ads are obtained by means of $adSearchTop(k, p, \texttt{AD})$ (cf. Section 2.1.2).

In our second method, we use other source of evidence provided by the advertisers: the keywords. With each ad a_i an advertiser associates a set of keywords $\mathcal{K}_i = \{\vec{k_1}, \vec{k_2}, ..., \vec{k_n}\}$. Each keyword $\vec{k_s} \in \mathcal{K}_i$ may be composed of one or more terms. In here, we use the keywords to match ads to the triggering page p. This provides our second method for ad matching given by:

$$\texttt{KW}(a_i, p) = sim(k_1 \cup k_2 \cup ... \cup k_n, p)$$

where $\vec{k_1}, \vec{k_2}, ..., \vec{k_n} \in \mathcal{K}_i$ and KW stands for "match the ad keywords". In this case, the k first ranked ads are obtained by means of $adSearchTop(k, p, \texttt{KW})$. Note this corresponds to the matching approach used in keyword targeted advertising with the triggering page taking the place of the user query.

We notice that most of the keywords selected by advertisers are also present in the ads associated with those keywords. For instance, in our ad test collection, this is true for 90% of the ads. Thus, instead of using the keywords as matching devices, we can use them to emphasize the main concepts in an ad, in an attempt to improve our AD strategy. This leads to our third method of ad matching given by:

$$\texttt{AD\_KW}(a_i, p) = sim(c_i \cup k_1 \cup k_2 \cup ... \cup k_n, p)$$

where $\vec{k_1}, \vec{k_2}, ..., \vec{k_n} \in \mathcal{K}_i$ and AD_KW stands for "match the ad and its keywords". In this case, we use $adSearchTop(k, p, \texttt{AD\_KW})$ to get the k first ranked ads.

Finally, it is important to notice that a keyword $\vec{k_s} \in \mathcal{K}_i$ associated with a_i could not appear at all in the triggering page p, even when a_i is highly ranked. However, if we assume that the keywords summarize the main topic of a_i according to an advertiser viewpoint, it can be interesting to ensure the presence of, at least, one of them in p. This reasoning suggests that requiring the occurrence of the keyword $\vec{k_s} \in \mathcal{K}_i$ in the triggering page p as a condition to associate a_i with p might lead to improved results. This leads to two extra matching strategies as follows:

$$\texttt{ANDKW}(a_i, p) = \begin{cases} sim(k_1 \cup ... \cup k_n, p) & \text{if } \exists_{\vec{k_s} \in \mathcal{K}_i} k_s \subseteqq p \\ 0 & \text{if otherwise} \end{cases}$$

$$\texttt{AD\_ANDKW}(a_i, p) = \texttt{AAK}(a_i, p) = \begin{cases} sim(c_i \cup k_1 \cup ... \cup k_n, p) & \text{if } \exists_{\vec{k_s} \in \mathcal{K}_i} k_s \subseteqq p \\ 0 & \text{if otherwise} \end{cases}$$

where $\vec{k_1}, ..., \vec{k_n} \in \mathcal{K}_i$, ANDKW stands for "match the ad keywords and force their appearance", and AD_ANDKW (or AAK for "ads and keywords") stands for "match the ad, its keywords, and force their appearance". In these cases, the k first ranked ads are obtained by means of $adSearchTop(k, p, \texttt{ANDKW})$ and $adSearchTop(k, p, \texttt{AAK})$, respectively.

As we will see in our results, the best among these simple methods is AAK. Thus, it will be used as baseline for our impedance coupling strategies which we now discuss.

4.2 Impedance Coupling Strategies

One key issue become clear as one plays with the content-targeted advertising problem. The association between a good ad and the triggering page might depend on a topic that is not mentioned explicitly in the triggering page. This is a common issue in many IR tasks based on syntactic match and is, in general, due to the fact that a same entity or subject of the world might be referred to in many distinct ways and through several different labels.

However, this situation is aggravated in content-targeted advertising because many advertisers bid on few ad keywords and Web pages present a rich vocabulary whereas ads are concise. To illustrate theses points, Figure 4.1 shows the number of ad keywords the advertisers associated with their campaigns in our test collection and the first line of Table 4.1 shows an example of a sports-related triggering page.

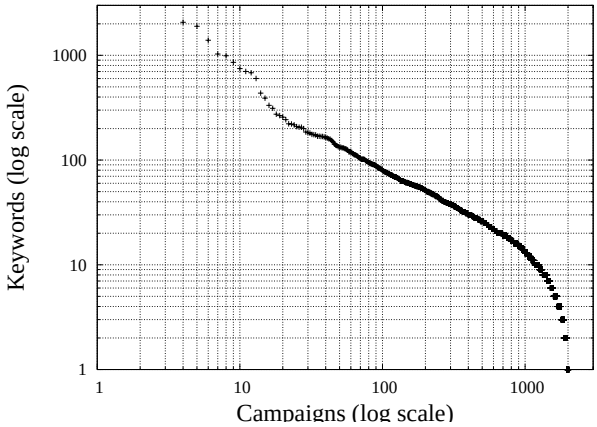


Figure 4.1: Number of keywords per campaign.

As we can see in Figure 4.1, although some advertisers have bid on a large number of keywords, only half of them selected more than ten keywords to their campaigns. The remaining used, in average, only 6.44 keywords. By betting on so few keywords, the advertisers tend to select terms of a more general nature to describe their interest areas. As a consequence, many specific terms that appear in the triggering page find no match in the ads. For instance, this might be the case of the triggering page p shown in Table 4.1. In spite of it being about a football match between Barcelona and Milan, it does not contain words such as "football", "barcelona", or "milan". On the other hand, it has a rich vocabulary full of specific terms related to football, such as "derby", "rossoneri", and the name of some players. Unfortunately, a sports clothing shop probably would prefer to use terms like "sports" or "football" instead of "eto'o" to describe its core business topic, possibly missing the opportunity to place ads in page p.

In summary, the vocabularies of pages and ads may have low intersection even when the ads are related to the pages. We cite this problem from now on as the *vocabulary impedance problem*. In our experiments, we realized that this problem limits the final quality of syntactical matching strategies. Therefore, we studied alternatives to reduce the referred vocabulary impedance.

For this, we propose to expand the triggering pages with new terms. Figure 4.2 illustrates our intuition. We already know that the addition of keywords (selected by the advertiser) to the ads leads to improved results. We say that a keyword reduces the vocabulary impedance by providing an alternative matching path. Our idea is to add new terms (words) to the Web page p to also reduce the vocabulary impedance by providing a second

Pages	Content
p	*Easter Treat in the **Champions League***: the derby is over and now all thoughts turn to the **Champions League**. The **Rossoneri** is proud to be once again in the last four of Europe's premier competition and to come up against the Catalan side who have the likes of **Ronaldinho**, **Eto'o** and Puyol. It will be a tough encounter against the Spanish **league** leaders but after we have all digested our Easter eggs there is another mouth-watering treat in store.
d_1	***Football** News*: **Barcelona** have Spain **football league** title in their sights – without **Ronaldinho** and Messi, **Eto'o** shines again; AC **Milan** leave it late to progress for UEFA **Champions League Football** – the **Rossoneri** waits for either FC **Barcelona** or SL Benfica;
d_2	*Singapore Soccer Community > **MILAN** vs **BARCELONA***: Watching **Milan** against Cagliari on Sunday I was struck by how little creativity they produced. Even accounting for the resting of **Kaka**, Inzaghi and Gatusso for most of the match, the team should have dealt with their opponents in a far more decisive manner than the eventual 1-0 scoreline. Against the catalans, I see them playing a tight **football** as usual, and trying to win that midfield battle of **Kaka**/Gatusso vs **Ronaldinho** /Deco as this is where the **game** will be won or lost. Like **Milan**, Barca are not renowned for their high goal average but I feel that they have enough talent up front with **Eto'o**, Larsson, and the 'magician' **Ronaldinho**, that I think it will be too much of a job for **Milan** at this time.
d_3	*CL **Barcelona** - **Milan** (Matchday 4)*: **Barcelona** has the home advantage and the Nou Camp will be packed to the rafters to create one of the most awesome atmospheres that is seen in **football**. **Barcelona** will try to utilize their full pitch to try and spread out the **Rossoneri** outfit and thereby create space for their adventurous cavalier play. The **Rossoneri** has been steadily improving and started to keep clean sheets again which is an ominous sign for **Barcelona**. However I don't think that **Milan** will keep a clean sheet in this **game**. **Milan** will have to play out of their skins to keep a clean sheet in the Nou Camp and shut out shop against **Barcelona** for two **games** in a row is highly unlikely. **BARCELONA**: Valdes; Gio, Puyol, Marques, Oleguer; Van Bommel, Edmilson, Van Bommel, Iniesta; Giuly, **Ronaldinho**, **Eto'o**. **MILAN**: Stam; Cafu, Nesta, Kaladze, Serginho; Pirlo, Gatusso, Seedorf; **Kaka**, Inzaghi, Shevchenko.

Table 4.1: Example of the contents of a triggering page (p) and other Web pages similar to p (d_1, d_2, and d_3).

alternative matching path. We refer to our expansion technique as impedance coupling. For this, we proceed as follows.

An advertiser trying to describe a certain topic in a concise way probably will choose general terms to characterize that topic. Thus, to facilitate the matching of this ad with triggering page p, we need to associate new general terms with p. For this, we assume that Web documents similar to the trigger-

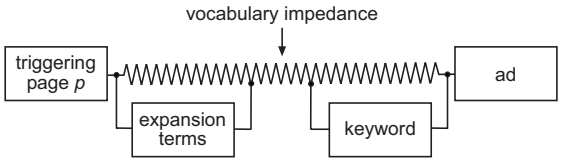


Figure 4.2: Addition of new terms to a Web page to reduce the vocabulary impedance.

ing page p share common topics. Therefore, by inspecting the vocabulary of these similar documents we might find good terms for better characterizing the main topics in the page p. To illustrate, Table 4.1 shows the content of page p and other three pages similar to p, obtained through a search engine. In this table, the most frequent terms in the shared vocabulary of the pages are highlighted in bold. Among the terms observed, some emphasize specific topics of p ("milan", "barcelona", "eto'o" and "ronaldinho") whereas others clearly characterize the contents of p in a more general fashion ("football"). Notice that many of these terms were not present in p ("milan", "barcelona", and "football"). We now describe this idea using a Bayesian network model depicted in Figure 4.3.

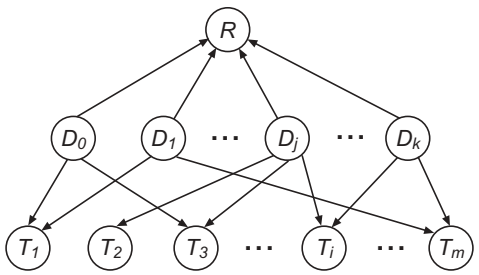


Figure 4.3: Bayesian network model for our impedance coupling technique.

In our model, which is based on the belief network in [104], the nodes represent pieces of information in the domain. With each node is associated a binary random variable, which takes the value 1 to mean that the corresponding entity (a page or terms) is *observed* and, thus, relevant in our computations. In this case, we say that the information was *observed*. Node R represents the page r, a new representation for the triggering page p. Let $\mathcal{N}$ be the set of the k most similar documents to the triggering page, including the triggering page p itself, in a large enough Web collection $\mathcal{C}$. Root nodes D_0 through D_k represent the documents in $\mathcal{N}$, that is, the triggering

page p (node D_0) and its k nearest neighbors, d_1 through d_k, among all pages in $\mathcal{C}$[1]. There is an edge from node D_j to node R if document d_j is in $\mathcal{N}$. Nodes T_1 through T_m represent the terms in the vocabulary of $\mathcal{C}$. There is an edge from node D_j to a node T_i if term t_i occurs in document d_j. In our model, the observation of the pages in $\mathcal{N}$ leads to the observation of r, the new representation of the triggering page p, and to a set of terms describing the main topics associated with p and its neighbors.

Given these definitions, we can now use the network to determine the probability that a term t_i is a good term for representing a topic of the triggering page p. In other words, we are interested in the probability of observing the final evidence regarding a term t_i, given that the new representation of the page p has been observed, $P(T_i = 1|R = 1)$. This translates into the following equation[2]:

$$P(T_i|R) = \frac{1}{P(R)} \sum_{\mathbf{d}} P(T_i|\mathbf{d}) P(R|\mathbf{d}) P(\mathbf{d}) \tag{4.2}$$

where $\mathbf{d}$ represents the set of states of the document nodes. As we are interested just in the states in which *only* a single document d_j is observed and $P(\mathbf{d})$ can be regarded as a constant, we can rewrite Equation (4.2) as:

$$P(T_i|R) = \frac{\nu}{P(R)} \sum_{j=0}^{k} P(T_i|\mathbf{d_j}) P(R|\mathbf{d_j}) \tag{4.3}$$

where $\mathbf{d_j}$ represents the state of the document nodes in which *only* document d_j is observed and ν is a constant associated with $P(\mathbf{d_j})$. Equation (4.3) is the general equation to compute the probability that a term t_i is related to the triggering page. We now define the probabilities $P(T_i|\mathbf{d_j})$ and $P(R|\mathbf{d_j})$ as follows:

$$P(T_i|\mathbf{d_j}) = \eta \; w_{ij} \tag{4.4}$$

$$P(R|\mathbf{d_j}) \;\; = \;\; \begin{cases} (1-\alpha) & j = 0 \\ \alpha \; sim(p, d_j) & 1 \leq j \leq k \end{cases} \tag{4.5}$$

where η is a normalizing constant, w_{ij} is the weight associated with term t_i in the document d_j, and $sim(p, d_j)$ is given by Equation (2.10), i.e., is the cosine similarity between p and d_j. The weight w_{ij} is computed using

[1]The neighbors documents in $\mathcal{N}$ can be obtained by means of the clustering technique discussed in Sections 2.4.1 and 5.2.1.

[2]To simplify our notation we represent the probabilities $P(X = 1)$ as $P(X)$ and $P(X = 0)$ as $P(\overline{X})$.

Equation 2.9 and is zero if term t_i does not occur in document d_j. Notice that $P(\overline{T}_i|\mathbf{d_j}) = 1 - P(T_i|\mathbf{d_j})$ and $P(\overline{R}|\mathbf{d_j}) = 1 - P(R|\mathbf{d_j})$. By defining the constant α, it is possible to determine how important should be the influence of the triggering page p to its new representation r. By substituting Equation (4.4) and Equation (4.5) into Equation (4.3), we obtain:

$$P(T_i|R) = \rho \; ((1-\alpha) \; w_{i0} + \alpha \; \sum_{j=1}^{k} w_{ij} \; sim(p, d_j)) \tag{4.6}$$

where $\rho = \eta \; \nu$ is a normalizing constant.

We use Equation (4.6) to determine the set of terms that will compose r, as illustrated in Figure 4.2. Let t_{top} be the top ranked term according to Equation (4.6). The set r is composed of the terms t_i such that $\frac{P(T_i|R)}{P(T_{top}|R)} \geq \beta$, where β is a given threshold. In our experiments, we have used $\beta = 0.05$. Notice that the set r might contain terms that already occur in p. That is, while we will refer to the set r as expansion terms, it should be clear that $p \cap r \neq \varnothing$.

By using $\alpha = 0$, we simply consider the terms originally in page p. By increasing α, we relax the context of the page p, adding terms from neighbor pages, turning page p into its new representation r. This is important because, sometimes, a topic apparently not important in the triggering page offers a good opportunity for advertising. For example, consider a triggering page that describes a congress in London about digital photography. Although London is probably not an important topic in this page, ads about hotels in London would be appropriate. Thus, adding "hotels" to page p is important. This suggests using $\alpha > 0$, that is, preserving the contents of p and using the terms in r to expand p.

In this work, we examine both approaches. Thus, in our sixth method we match r ($\alpha = 1$), the set of new expansion terms, directly to the ads, as follows:

$$\texttt{AAK\_T}(a_i, p) = \texttt{AAK}(a_i, r)$$

where `AAK_T` stands for "match the ad and keywords to the set r of expansion terms". In this case, we use $adSearchTop(k, p, \texttt{AAK\_T})$ to get the k first ranked ads.

In our seventh method, we match an expanded page p ($\alpha < 1$) to the ads as follows:

$$\texttt{AAK\_EXP}(a_i, p) = \texttt{AAK}(a_i, p \cup r)$$

where AAK_EXP stands for "match the ad and keywords to the expanded triggering page". In this case, we use $adSearchTop(k, p, \texttt{AAK\_EXP})$ to get the k first ranked ads.

To improve our ad placement methods, other external source that we can use is the content of the page h_i pointed to by the ad's hyperlink, that is, its *landing* page. After all, this page comprises the real target of the ad and perhaps could present a more detailed description of the product or service being advertised. Our eighth method consists of matching the triggering page p to the landing pages pointed to by the ads, as follows:

$$\texttt{H}(p, a_i) = sim(p, h_i)$$

H stands for "match the hyperlink pointed to by the ad". In this case, we use $adSearchTop(k, p, \texttt{H})$ to get the k first ranked ads.

We can also combine this information with the more promising methods previously described, AAK and AAK_EXP as follows. Thus, our two last methods are given by:

$$\texttt{AAK\_H}(a_i, p) = \begin{cases} sim(c_i \cup h_i \cup k_1 \cup ... \cup k_n, p) & \text{if } \exists_{\vec{k_s} \in \mathcal{K}_i} k_s \subseteqq p \\ 0 & \text{if otherwise} \end{cases}$$

$$\texttt{AAK\_EXP\_H}(a_i, p) = \begin{cases} sim(c_i \cup h_i \cup k_1 \cup ... \cup k_n, p \cup r) & \text{if } \exists_{\vec{k_s} \in \mathcal{K}_i} k_s \subseteqq (p \cup r) \\ 0 & \text{if otherwise} \end{cases}$$

where $\vec{k_1}, ..., \vec{k_n} \in \mathcal{K}_i$, AAK_H stands for "match ads and keywords also considering the page pointed by the ad", and AAK_EXP_H stands for "match ads and keywords with expanded triggering page, also considering the page pointed by the ad". In these cases, the k first ranked ads are obtained by means of $adSearchTop(k, p, \texttt{AAK\_H})$ and $adSearchTop(k, p, \texttt{AAK\_EXP\_H})$, respectively.

Notice that other combinations were considered in this study. However, they led to poor results in our experimentation and for this reason were discarded. We also have tested stemming in trying to enhance our methods. Nonetheless, this technique did not yield consistent improvements. Table 4.2 summarizes all the matching strategies presented in this work.

4.3 Experiments

4.3.1 Test Collections

To evaluate our ad placement strategies, we performed a series of experiments using a sample of a real case ad collection with 93,972 ads, 1,744 advertisers,

Matching Strategies	Triggering Page			Advertisement			
	original terms	new terms (set r)	keyword required	title	description	keyword	landing page
AD	×			×	×		
KW	×					×	
AD_KW	×			×	×	×	
ANDKW	×		×			×	
AD_ANDKW (AAK)	×		×	×	×	×	
AAK_T		×	×	×	×	×	
AAK_EXP	×	×	×	×	×	×	
H	×						×
AAK_H	×		×	×	×	×	×
AAK_EXP_H	×	×	×	×	×	×	×

Table 4.2: Summary of the matching strategies. An "×" indicates the evidence or restriction considered. For example, strategy ANDKW is a matching of the triggering page with the ad keyword restricted to the appearance of, at least, one keyword in the triggering page.

and 68,238 keywords[3]. The ads are grouped in 2,029 campaigns with an average of 1.16 campaigns per advertiser. In this collection, only one keyword is associated with each ad. This makes campaigns very important since they are used by the advertisers to associate several keywords with a product or service.

As we are initially interested in the placement of ads in the pages of information portals, our triggering page collection was composed of 100 pages extracted from a Brazilian newspaper. They were crawled in such a way that only the contents of their articles were preserved. As we have no preference for particular topics, these pages cover subjects as diverse as culture, local news, international news, economy, sports, politics, agriculture, cars, children, real estate, computers and internet, TV, travels, and economy.

For the strategies AAK_T and AAK_EXP, we had to generate a set of expansion terms. For that, we used a database of Web pages crawled by the

[3]Data in portuguese provided by an on-line advertisement company that operates in Brazil.

TodoBR search engine [115] (`http://www.todobr.com.br/`). This database is composed of 5,939,061 pages of the Brazilian Web, under the domain ".br". For the strategies `H`, `AAK_H`, and `AAK_EXP_H`, we also crawled the landing pages pointed to by the advertisers. No other filtering method was applied to these pages besides the removal of HTML tags.

4.3.2 Methodology and Evaluation

As previously mentioned in Section 3.1.1, several works have stressed the importance of relevance in search advertising for consumers, brokers, information gatekeepers, and advertisers. Thus, we evaluate our methods essentially by considering their performance in terms of relevance. More specifically, we adopt the same pooling method used to evaluate the TREC Web-based collection [53], as described as follows. For each of our 100 triggering pages, we selected the top three ranked ads provided by each of our 10 ad placement strategies. Thus, for each triggering page we selected no more than 30 ads. As a result, a total of 1,860 distinct ads were selected. They were then inserted into pools corresponding to its triggering pages. These top ads were then inserted in a pool for that triggering page. Each pool contained an average of 15.81 ads. All ads in each pool were submitted to a manual evaluation by a group of 15 users. Each user was asked to evaluate the ads selected to each page according to its relevance to the pages[4]. The average number of relevant ads per page pool was 5.15.

We present the results of our experiments considering that a triggering page offers three ad slots and the studied methods always attempt to fill all the slots. To quantify the precision of our methods, we used PAVG and PAVG@3 metrics (cf. Section 2.6.1). For all the methods, we also report the average precision at the first, second, and third top-ranked ad. Since we are not able to evaluate the entire ad collection, recall values are relative to the set of the evaluated ads. All the results to be presented were found statistically significant, when tested with the two-tailed paired t-test, at least at the 90% level.

[4]Since our methods were not designed to avoid inappropriate, unethical or illegal situations, the users were asked to pay special attention to these aspects of the ad placement problem. In particular, they were asked not to consider irrelevant sex-related ads, ads from competitors, and ads about for-sale term papers only because they should be inappropriate for the publisher or considered unethical.

4.3.3 Tuning IDF factors

We start by analyzing the impact of different IDF factors in our ad collection. IDF factors are important because they quantify how discriminative is a term in the collection. In our ad collection, IDF factors can be computed by taking ads, advertisers or campaigns as documents. Hence, we can compute ad, advertiser or campaign IDF factors. As we observe in Figure 4.4, for the `AD` strategy, the best ranking is obtained by the use of campaign IDF, that is, by calculating our IDF factor so that it discriminates campaigns. Similar results were obtained for all the other methods.

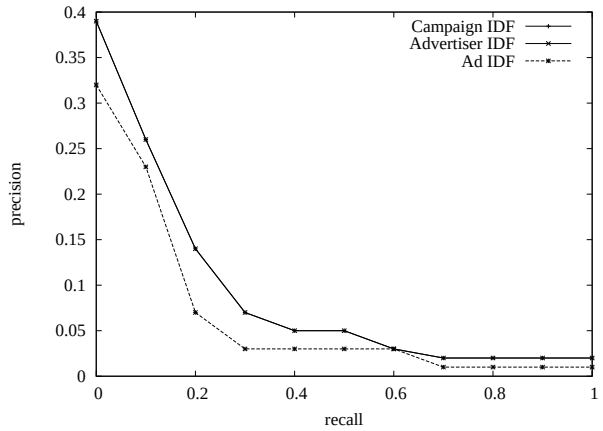


Figure 4.4: Precision-recall curves obtained for the `AD` strategy using ad, advertiser, and campaign IDF factors.

This reflects the fact that terms might be better discriminators for a business topic than for a specific ad. This effect can be accomplished by calculating the factor relative to IDF advertisers or campaigns instead of ads. In fact, campaign IDF factors yielded the best results. Thus, they will be used in all the experiments reported from now on.

4.3.4 Results

Simple Matching Strategies

Figure 4.5 displays the results for the simple matching strategies presented in Section 4.1. As shown, directly matching of the contents of the ad to the triggering page (`AD` strategy) is not so effective. The reason is that the ad contents are very noisy. It may contain messages that do not properly describe the ad topics such as requisitions for user actions (e.g, "visit our

site") and general sentences that could be applied to any product or service (e.g, "we delivery for the whole country"). On the other hand, an advertiser provided keyword summarizes well the topic of the ad. As a consequence, the `KW` strategy is superior to the `AD` and `AD_KW` strategies. This situation changes when we require the keywords to appear in the target Web page. By filtering out ads whose keywords do not occur in the triggering page, much noise is discarded. This makes `ANDKW` a better alternative than `KW`. Further, in this new situation, the contents of the ad becomes useful to rank the most relevant ads making `AD_ANDKW` (or `AAK` for "ads and keywords") the best among all described methods. For this reason, we adopt `AAK` as our baseline in the next set of experiments.

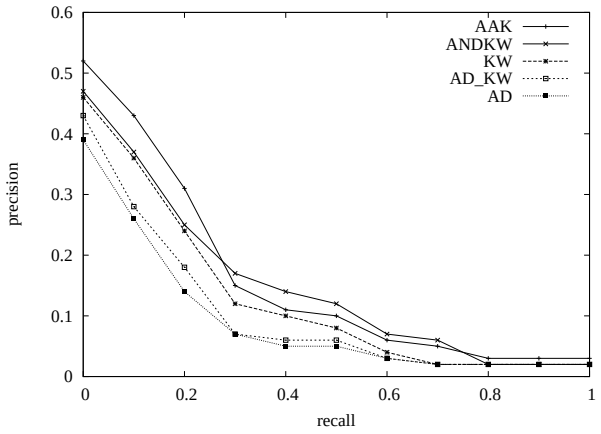


Figure 4.5: Comparison among our five simple matching strategies. `AAK` ("ads and keywords") is superior.

Table 4.3 illustrates average precision figures for Figure 4.5. We notice that our `AAK` strategy provides a gain in average precision of about 60% relative to the trivial `AD` strategy. Both gains are significant at the 99% confidence level. This shows that careful consideration of the evidence related to the problem does pay off.

Impedance Coupling Strategies

Table 4.4 shows top ranked terms that occur in a page covering Argentinean wines produced using grapes derived from the Bordeaux region of France. The p column includes the top terms for this page ranked according to the TF-IDF weighting scheme. The r column includes the top ranked expansion terms generated according to Equation (4.6). Notice that the expansion

Methods	Precision			PAVG			PAVG@3		
	@1	@2	@3	score	gain(%)	cl(%)	score	gain(%)	cl(%)
AD	0.410	0.365	0.287	0.110	-	-	0.257	-	-
AD_KW	0.510	0.395	0.320	0.124	+12.7	90	0.296	+15.2	95
KW	0.460	0.395	0.353	0.136	+23.6	95	0.323	+25.7	95
ANDKW	0.490	0.425	0.400	0.160	+45.5	99	0.364	+41.6	99
AD_ANDKW (AAK)	0.510	0.495	0.460	0.175	+59.1	99	0.412	+60.3	99

Table 4.3: Average precision figures, corresponding to Figure 4.5, for our five simple matching strategies. The AAK strategy provides improvements of about 60% relative to the AD strategy in both PAVG and PAVG@3 metrics. Note that *cl* stands for t-test confidence level.

terms not only emphasize important terms of the target page (by increasing their weights) such as "wines" and "whites", but also reveal new terms related to the main topic of the page such as "aroma" and "red". Further, they avoid some uninteresting terms such as "obtained" and "country".

Figure 4.6 illustrates our results when the set r of expansion terms is used. They show that matching the ads to the terms in set r instead of to the triggering page p (AAK_T strategy) leads to a considerable improvement over our baseline, AAK. The gain is even larger when we use the terms in r to expand the triggering page (method AAK_EXP). This confirms our hypothesis that the triggering page could have some interesting terms that should not be completely discarded.

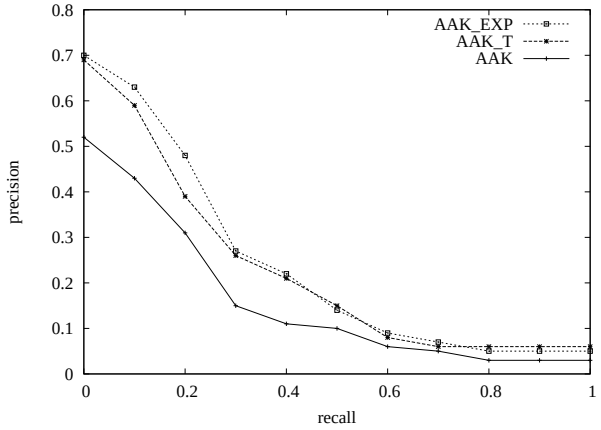


Figure 4.6: Impact of using a new representation for the triggering page, one that includes expansion terms.

Finally, we analyze the impact on the ranking of using the contents of pages pointed to by the ads. Figure 4.7 displays our results. It is clear that

Rank	p		r	
	term	score	term	score
1	argentina	0.090	wines	0.251
2	obtained*	0.047	wine*	0.140
3	class*	0.036	whites	0.091
4	whites	0.035	red*	0.057
5	french*	0.031	grape	0.051
6	origin*	0.029	bordeaux	0.045
7	france*	0.029	acideness*	0.038
8	grape	0.017	argentina	0.037
9	sweet*	0.016	aroma*	0.037
10	country*	0.013	*blanc**	0.036
...				
35	wines	0.010	-	-
...				

Table 4.4: Top ranked terms for the triggering page p according to our TF-IDF weighting scheme and top ranked terms for r, the expansion terms for p, generated according to Equation (4.6). Ranking scores were normalized in order to sum up to 1. Terms marked with '*' are not shared by the sets p and r.

using only the contents of the pages pointed by the ads (`H` strategy) yields very poor results. We also did not achieve statistically significant gains by combining this evidence with our baseline. However, combining our best strategy so far (`AAK_EXP`) with pages pointed by ads (`AAK_EXP_H` strategy) leads to superior results. This happens because the two additional sources of evidence, expansion terms and pages pointed by the ads, are distinct and complementary, providing extra and valuable information for matching ads to a Web page.

Figure 4.8 and Table 4.5 summarize all results described in this section. In Figure 4.8 we show precision-recall curves and in Table 4.5 we show average precision figures. We also present gains in average precision relative to our baseline, `AAK`. We notice that the best performance in the first positions (PAVG@3) was achieved by method `AAK_EXP`. However, the method with best overall retrieval performance was `AAK_EXP_H`, yielding a gain in PAVG figures of roughly 43% over the baseline (`AAK`).

All the experiments here described were repeated for an alternative setting where ads in a same campaign were grouped. By doing this, we simulated the situation in which several keywords are associated with only one ad. The results obtained were similar to those described in this section. General performance was, however, slightly worse. This occurred because some

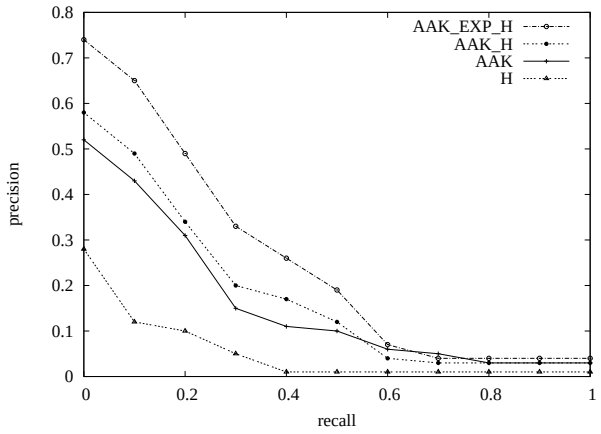


Figure 4.7: Impact of using the contents of the page pointed by the ad (the hyperlink).

companies bid on large number of keywords hurting the performance of the matching algorithms.

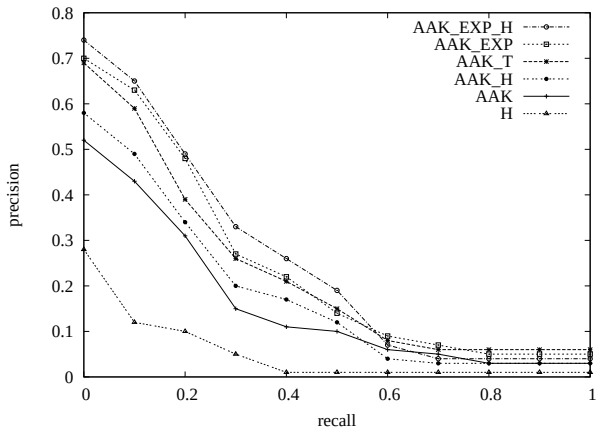


Figure 4.8: Comparison among our ad placement strategies.

4.3.5 Analysis of Misplacements

Since method `AAK_EXP` presented the best overall results, we carefully evaluated its ad misplacements. In particular, we studied a random sample composed by 87 ad suggestions provided by `AAK_EXP` which were considered

Methods	Precision			PAVG			PAVG@3		
	@1	@2	@3	score	gain(%)	cl(%)	score	gain(%)	cl(%)
H	0.310	0.210	0.153	0.060	-65.7	99	0.134	-67.5	99
AAK	0.510	0.495	0.460	0.175	-	-	0.412	-	-
AAK_H	0.510	0.510	0.463	0.181	+3.4	-	0.421	+2.2	-
AAK_T	0.663	0.582	0.534	0.231	+32.0	95	0.498	+20.9	90
AAK_EXP	0.700	0.610	0.583	0.248	+41.7	99	0.554	+34.5	99
AAK_EXP_H	0.690	0.615	0.570	0.250	+42.9	99	0.533	+29.4	99

Table 4.5: Results for our impedance coupling strategies. Note that *cl* stands for t-test confidence level.

not correct by the evaluators. The aim of this analysis was to classify the observed misplacements, pointing out possible reasons for the selection of improper ads.

As a result of this analysis, we observed the misplacements can be classified into the following cases (note that a same misplacement can be classified under several cases):

Term ambiguity Ads are selected due to the occurrence of ambiguous terms. This case was observed in 52% of the misplacements in our sample. An example of such issue is the placement of ads about exchange programs for learning Spanish (language) in Barcelona (city) into an article about the European soccer transfer season involving Spanish teams such as Barcelona. Other examples of ambiguous matching involved brands, proper names, nouns, and verbs.

Wrong focus Ads are selected due to the occurrence of very specific terms not related to the main focus of the page. This case was observed in about 34% of the misplacements. For instance, ads about bank services and a tourist resort (Ilhabela) were associated with a page describing the final race of the Brazilian sailing championship that took place in Ilhabela and whose champion team was sponsored by a bank. In this case, the article was clearly focused on the sports event and not on the location where it took place nor the service provided by the sponsor company. In particular, we noticed that contextual terms indicating geographical location tend to attract ads about regional topics, tourism, exchange programs, and language courses, among others. In many of these cases, the evaluator considers the placement not suitable.

Hard topics No ads in the database are directly related to the content of the page. This is a problem because AAK_EXP always suggests ads. This case was observed in 10% of the misplacements. Examples of this case are pages about politics, war, weapons, and historic events, among others.

Other cases The remaining 4% of the misplacements comprise different situations. For instance, an association between an ad and a document was considered improper because the page presented a negative viewpoint about the advertised product. In another case, the expansion technique emphasized the politic context of a page (instead of its economic context) making it hard to find proper advertisements.

As we can observe, ambiguity and focusing on the wrong topic are the main reasons for misplacements. The techniques discussed in this chapter attempted to deal with these issues by disambiguating and emphasizing topics by providing additional terms for matching. However, new techniques should be explored for better characterizing the topics of the ads and the triggering pages.

4.3.6 Performance Issues

In a keyword targeted advertising system, ads are assigned at query time, thus the performance of the system is a very important issue. In content-targeted advertising systems, we can associate ads with a page at publishing (or updating) time. Also, if a new ad comes in we might consider assigning this ad to already published pages in offline mode. That is, we might design the system such that its performance depends fundamentally on the rate that new pages are published and the rate that ads are added or modified. Further, the data needed by our strategies (page crawling, page expansion, and ad link crawling) can be gathered and processed offline, not affecting the user experience. Thus, from this point of view, the performance is not critical and will not be addressed in this work.

4.4 Conclusions

In this chapter we investigated ten distinct strategies for associating ads with a Web page that is browsed (content-targeted advertising). Five of our strategies attempt to match the ads directly to the Web page. Because of that, they are called *simple matching strategies*. The other five strategies recognize that there is a vocabulary impedance problem among ads and Web pages and attempt to solve the problem by expanding the Web pages and the ads with new terms. Because of that they are called *impedance coupling strategies*.

Using a sample of a real case database with over 93 thousand ads, we evaluated our strategies. For the five simple matching strategies, our results

indicated that planned consideration of additional evidence (such as the keywords provided by the advertisers) yielded gains in average precision figures (for our test collection) of about 60%. This was obtained by a strategy called `AAK` (for "ads and keywords"), which is taken as the baseline for evaluating our more advanced impedance coupling strategies.

For our five impedance coupling strategies, the results indicate that additional gains in average precision of about 42% (now relative to the `AAK` strategy) are possible. These were generated by expanding the Web page with new terms (obtained using a sample Web collection containing over five million pages) and the ads with the contents of the page they point to (a hyperlink provided by the advertisers).

Chapter 5

Conceptual Matching Strategies

In this chapter, we analyze how conceptual information could be used to improve the precision of ad placement methods based on syntactical matching. This study is motivated by the perception that conceptual associations can indicate good opportunities to either avoid or place ads. In Section 5.2, we first propose how to use link information to enhance traditional content-based classifiers since we believe that the hypertextual nature of the Web is a useful resource to indicate page's topic and importance. In Section 5.3, we study how to combine the syntactical matching approach discussed in the previous chapter with the category information provided by our best classifiers. In Section 5.4, we present the results obtained through experiments performed on a real ad collection, using a manually-built ad taxonomy and a Web directory as source of triggering pages. Finally, in Section 5.5, we present our conclusions.

5.1 Classification in Content-targeted Advertising

The document classification technology is not new for search advertising. Over the years, it has been employed as a filtering tool as well as a means of targeting audiences. As a filtering tool, it is useful to avoid the ad misplacement common in situations where, in spite of being related to the same subject, the association between ad and page is not appropriate [55, 123]. An example of such a case is the assignment of an ad to a page about a catastrophe. Moreover, matching systems have to be able to recognize unethical and illegal advertising to avoid placing ads supporting specific subjects such as pornography, drugs, and gambling games. As a means of targeting audiences, document classification has been exploited for many years in Web

directory advertising. Directories provide to the advertisers the possibility to reach users interested in specific and well defined topics.

In content-targeted advertising, in particular, document classification was introduced by FindWhat[1] [87]. In FindWhat's system, publishers (advertisers) classify their pages (ads) according to a pre-defined taxonomy. Thus, an ad is associated with a page if both are classified under the same concept[2], that is, they match conceptually. The motivation for the employment of such classification filter is to avoid the matching challenges and ensure the ads on a page are at least generally related to its content [89]. Notice that FindWhat uses document classification in a similar way to that of Web directories.

Differently from FindWhat, we are interested in applying document classification to enhance the performance of syntactical matching systems. As seen in Chapter 4, these systems are based on algorithms that return a match whenever an exact occurrence of an ad term is found in the triggering page, regardless of its meaning. By considering the context in which the match occurs, we believe that we can avoid many cases of ad misplacement. For instance, the association of an ad on the rock band Bush with a page about the US president Bush could be avoided if the system took into consideration the difference between the contexts of the ad (music) and the page (politics). Further, from our previous experiments in content-targeted advertising, we learned that associations between ads and pages may be appropriate even when the ad and the page are related to each other in a broad conceptual scope. For instance, users may consider an ad about a music CD appropriate for a page about a music show just because they are about music. The same was observed for topics related to tourism, movies, books, and real estate, among others.

To take advantage of this conceptual matching approach we need to build taxonomies for the ad and page collections. As in FindWhat scenario, we can rely on the advertisers and publishers to provide this information manually. However, manual classification poses some challenges since it can be unfeasible, in particular, for large sites and very complex taxonomies. Thus, providing the concepts in a completely automatic fashion would be beneficial. Further, for our purposes, a classification method that provides an ordered list of concepts ranked according to likelihood may be useful even if the top concept or the order of the concepts is not correct. Motivated by these observations, in this chapter, we also consider the scenario where manual classification is not available and, therefore, automatic classifiers have to be

[1] `http://www.findwhat.com/`

[2] In this work, we will use the terms concept, class, and category in a largely interchangeable way.

used. Nevertheless, as it is known that traditional classifiers tend to perform poorly in the Web [15, 47], we propose methods to improve their effectiveness. In summary, in the following subsections, we present new methods for automatic classification of Web documents and combination of conceptual and syntactical evidence.

5.2 Classifying Web Documents

Automatic classification is particularly difficult in the Web because pages are usually noisy and with little text, containing images, scripts and other types of data unusable by traditional text classifiers. Furthermore, they can be created by many different authors, with no coherence in style, language or structure. By other hand, the Web provides ground to explore a new set of possibilities. Multimedia documents, semi-structured data, user behavior logs, and many other sources of information allow a whole new range of IR algorithms to be tested. In this section, we focus on one such source of information, widely available in the Web: its link structure. As the hypertextual nature of the Web can be used to indicate a document's topic and importance, we intend to use this resource to improve text-based classification algorithms.

We now describe how the link-based similarity measures of Section 2.5 can be used on a Web classification task. We start by showing how they can be applied in a well-known classification algorithm: the *kNN* classifier. This application will allow us to determine the effectiveness of each link-based similarity measure in determining how related two Web pages are. Following that, we use a Bayesian network model to combine the results achieved by the similarity measures with the results of a traditional text-based classifier. This will determine if the results of the text-based classifier can be improved by the use of link-based information. To end this section, we provide a brief description on how text-based classifier decisions will be used in this work.

5.2.1 Link-based Similarity Measures in *kNN* Classifier

To evaluate the link-based similarity measures, we used a strategy based on a nearest neighbor classifier, described in Section 2.4.1. The *kNN* algorithm was chosen because it is simple, efficient, and makes a direct use of similarity information. To test the link-based similarity measures, Equations 2.22, 2.23, 2.24 and the values returned by the Companion algorithm were used in place of the cosine similarity, in Equation 2.11. This allowed us to test all measures under the same set of conditions, and evaluate how accurate they

are in predicting the subject of Web pages.

5.2.2 Combining Link-Based and Text-Based Classification

To evaluate the effects of combining the link-based classification, described in Section 5.2.1, with the results of a traditional text-based classifier, we propose the use of the Bayesian network model shown in Figure 5.1.

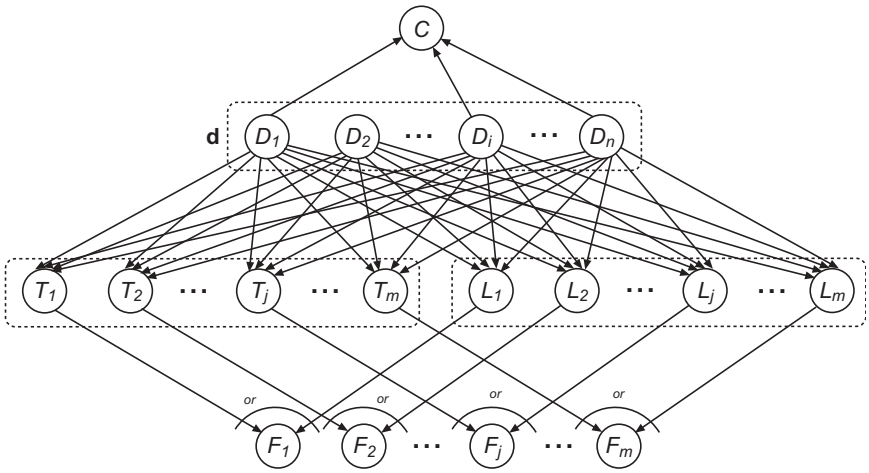


Figure 5.1: Bayesian network model to combine a text-based classifier with evidence from link structure.

In the network of Figure 5.1, the root nodes, labeled D_1 through D_n, represent our prior knowledge about the problem, i.e., a set of classified documents (the training set), node C represents category c, each node T_j represents evidence from the text-based classifier indicating that test document j belongs to category c, each node L_j represents evidence given from the link-based classifier indicating that document j belongs to category c, and nodes F_1 through F_m represent the final evidence that each test document belongs to category c.

Given these definitions, we can use the network to determine the probability that a test document j belongs to category c. This translates to the following equation:

$$P(f_j|c) = \eta \sum_{\mathbf{d}} \Big[1 - \big(1 - W_t \; P(t_j|\mathbf{d})\big)\big(1 - W_\ell \; P(\ell_j|\mathbf{d})\big)\Big] \; P(c|\mathbf{d}) \qquad (5.1)$$

where $\eta = P(\mathbf{d})/P(c)$ is a normalizing constant and $\mathbf{d}$ is a possible state of all the variables D_j. The probability $P(c|\mathbf{d})$ is now used to select only

the training documents that belong to the category we want to process. We define $P(c|\mathbf{d})$ as:

$$P(c|\mathbf{d}) = \begin{cases} 1 & \text{if } \forall i,\, d_i = 1 \text{ iff } i \in \mathcal{T}_c \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

where $\mathcal{T}_c$ is the set of training documents that belong to category c. By applying Equation 5.2 to Equation 5.1, we obtain:

$$P(f_j|c) = \eta \sum_{\mathbf{d}_c} \Big[1 - \big(1 - W_t\ P(t_j|\mathbf{d}_c)\big)\big(1 - W_\ell\ P(\ell_j|\mathbf{d}_c)\big) \Big] \tag{5.3}$$

where $\mathbf{d}_c$ is the state of variables D_j where only the variables corresponding to the training documents of class c are active. Constants W_t and W_ℓ are the weights given to the text-based classifier and to the link-based similarity measures, respectively. They can be used to regulate the importance of each source of evidence on the final result. The introduction of weights in the model is accomplished by the use of a *noisy-OR* combination [99].

To compute the final probability, we simply define $P(t_j|\mathbf{d}_c)$ as the value given by a text-based classifier of document j belonging to class c, and $P(\ell_j|\mathbf{d}_c)$ as the value given by the link-based classification algorithm described in Section 5.2.1. Both values are assumed to be normalized to fit between 0 and 1.

We note that the Bayesian network model here presented takes an epistemological view (as opposed to a frequentist view) of the information retrieval problem and interprets probabilities as degrees of belief devoid of experimentation, as also done in [131,138]. The value returned by each of the classifiers is, therefore, interpreted as the belief that a given document belongs to a given category and applied to Equation (5.3).

5.2.3 Text-Based Classifier Decision as a Belief

In our experiments, three well known classification algorithms were used as the text-based sources of evidence: *kNN*, Naive Bayes, and Support Vector Machine (SVM).

The *kNN* classifier used is described in Section 2.4.1 and was applied here with the conventional TF-IDF weights defined in Equation (2.9). To be applied in the Bayesian network, the scores $s_{c_i,d}$, yielded by Equation (2.11), were normalized such that $\sum_{\forall c_i} s_{c_i,d} = 1$. This normalized value can be interpreted as the degree of belief that document d belongs to category c_i and

applied to Equation (5.3). If the sum of all scores is zero, the prior probability of the category, i.e., the fraction of training documents that belong to the category, is used.

The Naive Bayes classifier used is described in Section 2.4.2. In this classifier, a document can be classified by computing the probability of a class given a document $P(c_j|d_i, \theta)$. This probability can also be used in the Bayesian network model of Section 5.2.2 as the degree of belief for the text-based classifier.

The SVM classifier used is described in Section 2.4.3. As the SVM decision is based on a voting scheme, to apply this classification to the Bayesian network model of Section 5.2.2, we interpret the number of votes given to a class c_i as the degree of belief that document d belongs to c_i. This value, normalized by the total number of votes, can then be applied to Equation (5.3).

5.3 Applying Classification to Content-targeted Advertising

We now describe how the category information can be combined with syntactical matching strategies, such as `AAK` (cf. Section 4.1), to improve the ranking precision. We start by defining a concept-based similarity function that associates ads with triggering pages if they are classified under the same concept. Following that, we discuss how to incorporate these concept-based similarities into our syntactical matching strategies.

5.3.1 Concept-based Similarity

Since the output of all previously described classifiers corresponds to lists of categories ordered according to likelihood, we can use them for hard and soft classification. As for soft classification, no additional processing is necessary because the classifiers already provide soft decisions. Soft classification is interesting for content-targeted advertising because a triggering page can be about several topics. For instance, a page about fishing can belong to categories such as *Sports* and *Leisure*. Further, taxonomies can be ambiguously defined making it hard to distinguish categories such as, for example, *Internet* and *Computers*. However, soft classification can introduce some noise since each page may be associated with a large number of categories which increases the chance of misclassification. To determine the impact of such noise, we will also test hard classification. For this, we convert soft decisions to hard decisions as follows.

Let $\mathbf{p}_i = \{p_{i1}, p_{i2}, ..., p_{im}\}$ be the category vector associated with document d_i, where p_{ij} is the probability that page d_i belongs to category c_j. As previously mentioned, for soft classification, p_{ij} is a real value in the interval $[0, 1]$. For hard classification p_{ij} is given by Equation (5.4).

$$p_{ij} = \begin{cases} 1 & \text{if } (\nexists_{k \neq i}\ p_{kj} = 1) \wedge \\ & (\forall_{k \neq i}\ p_{ij} > p_{kj} \vee \\ & (\nexists_{k \neq i}\ p_{kj} > p_{ij} \wedge c_j \text{ is the most frequent class})) \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

The first condition in Equation (5.4) ensures that only one category will be assigned to a document whereas the second condition selects the category with the largest probability value. These conditions define a hard mono-classification setting. The third condition decides ties by selecting the most frequent class, a usual machine learning approach to reduce classification error [86].

Given the set of ads $\mathcal{A}$ and the set of triggering pages $\mathcal{D}$, we now describe how the classification information can be used as a similarity measure between $a_j \in \mathcal{A}$ and $d_i \in \mathcal{D}$. For this, we first extend our previous definition of an ad, given in Section 4.1. We now consider that an ad a_j is composed of a category vector besides their creative, keywords, and hyperlink. Formally, an ad $a_j \in \mathcal{A}$ can be defined as a tuple as given by Equation 5.5:

$$a_j = \langle \vec{c_j}, \mathcal{K}_j, \vec{h_j}, \vec{s_j} \rangle \tag{5.5}$$

where $\vec{c_j}$ represents the ad creative, $\mathcal{K}_j = \{\vec{k_1}, \vec{k_2}, ..., \vec{k_n}\}$ represents the set of keywords associated with a_j, $\vec{h_j}$ represents the landing page pointed to by a_j, and $\vec{s_j} = \{s_{j1}, s_{j2}, ..., s_{jm}\}$ is the category vector associated with a_j. By representing a_j and d_i using their category vectors $\vec{s_j}$ and $\vec{p_i}$, respectively, we define the concept-based similarity function $csim(a_j, d_i) \colon \mathcal{D} \times \mathcal{A} \rightarrow \mathbb{R}$ as:

$$csim(a_j, d_i) = \frac{\sum_{t=1}^{m} s_{jt} \times p_{it}}{\sqrt{\sum_{t=1}^{m} s_{jt}^2} \times \sqrt{\sum_{t=1}^{m} p_{it}^2}} \tag{5.6}$$

In other words, the conceptual similarity of a_j with regard to d_i is given by the cosine value of the angle between their corresponding category vectors.

5.3.2 Combining Conceptual and Syntactical Similarities

The basic idea behind using classification in content-target advertising is to avoid associating ads with triggering pages if they are not classified un-

der the same concepts. A straightforward way to achieve this purpose is by combining the decisions provided by the conceptual and syntactical similarity functions by means of a conjunctive operator, as follows. Let $ssim(a_j, d_i)$ represent a syntactical similarity function to be combined with $csim(a_j, d_i)$. The conjunctive combination of $ssim(a_j, d_i)$ and $csim(a_j, d_i)$ is given by Equation 5.7.

$$andsim(a_j, d_i) = csim(a_j, d_i) \times ssim(a_j, d_i) \tag{5.7}$$

Clearly, function *andsim* has a restrictive nature that occasionally may prevent the selection of relevant ads. This could happen because there are situations in which, in spite of an ad not being classified under the same concept of a triggering page, its placement in that page would be convenient. For example, an ad about hotels in Turin would be certainly appropriate for a page about the olympic winter games in Turin, despite the ad being about tourism and the page, about sports. To deal with this problem, we use a noisy-or combination (cf. Section 5.2.2) as defined by Equation 5.8.

$$norsim(a_j, d_i) = 1 - (1 - csim(a_j, d_i))\ (1 - ssim(a_j, d_i)) \tag{5.8}$$

In this case, the placement of an ad is considered pertinent if either it is classified under the same category of the triggering page or it is related to the triggering page according to *ssim*.

5.4 Experiments

To evaluate the effectiveness of the proposed methods for classifying documents and combining conceptual and syntactical evidence, a set of experiments was performed using Web directory pages and an ad collection. We now describe the experimental setup and discuss the results achieved.

5.4.1 Test Collections

In this section we present the ad collection and the Web directory collection used in our experiments.

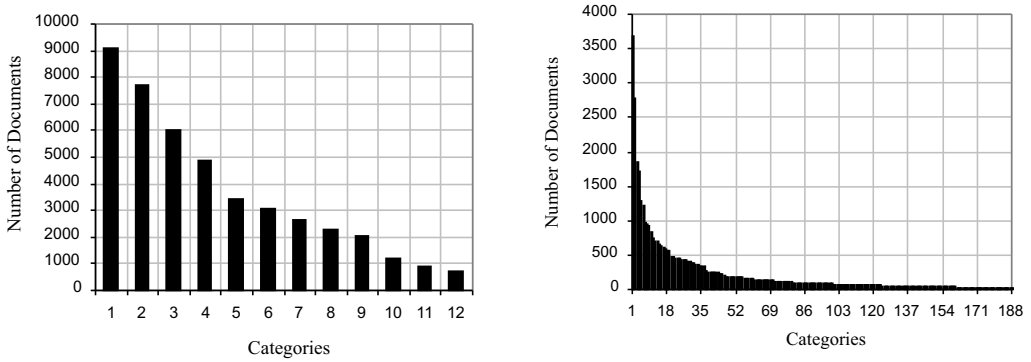
Web Directory Collections

The experiments presented here were performed using a set of classified Web pages extracted from the Cadê directory (`http://www.cade.com.br/`). These pages were used to test our classification approaches and as triggering pages for our ad placement methods. Cadê is a Brazilian Web directory

pointing to Web pages that were classified by human experts. To obtain the contents of the classified pages we used the database of Web pages crawled by the TodoBR search engine [115] (`http://www.todobr.com.br/`). All the text contained in the body of the HTML documents, plus the text between the title tags was used to index the pages.

Two sub-collections were constructed using the data available on Cadê, Cade12 and Cade188. Cade12 is a set of 44,099 pages labeled using the 12 first level categories of Cadê directory (*Computers*, *Culture*, *Education*, *Health*, *Internet*, *Leisure & Entertainment*, *News*, *On-line Shopping*, *Science*, *Services*, *Society*, and *Sports*). Cade188 corresponds to a set of 42,123 pages labeled using the 188 second level categories of Cadê (*Science/Biology*, *Science/Chemistry*, *Culture/Dance*, *Culture/Music*, *Education/Schools*, *Education/Universities*, etc.). Each Web page is classified into only one category.

Figures 5.2(a) and 5.2(b) show the category distributions for the Cade12 and Cade188 collections. Notice that both have skewed distributions. In Cade12, the three most popular categories represent more than 50% of all documents. The most popular category, *Services*, has 9,081 documents while the least popular, *On-line Shopping*, has 715 documents. In Cade188, 50% of the documents are in just 10% of the categories. The most popular category, *Society/People*, has 3,675 documents while the least popular, *Internet/Tutorials*, has 24 documents. Cade12 and Cade188 have vocabularies of 191,962 and 168,257 unique words, respectively, after removing stop words.



(a) Cade12 (b) Cade188

Figure 5.2: Category distribution for the Cade12 and Cade188 collections.

Information about the links related to the Cadê pages was also collected from the TodoBR database. TodoBR provides 40,871,504 links between Web pages (an average of 6.9 links per page). We extracted from this set all the links related to the pages of our two experimental sub-collections. Links

connecting different pages of the same site were discarded because they corresponded mainly to navigational menus. Table 5.1 summarizes the data obtained.

Links were divided into two types: *internal*, which are links between pages classified by Cadê, and *external*, which are links where the target or the source page is in TodoBR, but not in the set of pages classified by Cadê. This distinction is important to verify whether the external information provided by TodoBR can be used to improve the results.

Statistics	Whole Cadê	Cadê without Directory Pages
Internal Links	45,548	3,830
Links from external pages to Cadê pages	570,404	554,592
Links from Cadê pages to external pages	7,584	5,894
Cadê pages with no in-links	2,556	4,392
Cadê pages with no out-links	40,917	40,723

Table 5.1: Link statistics for the Cadê collection.

We call *directory pages* those that belong to the Cadê site itself and are used to compose the directory hierarchy. For instance, the Cadê Science page, which links to science related sites, is a directory page. As can be seen, directory pages represent a great part of the internal links in the Cadê collection. As directory pages provide information on the categories of the remaining pages (for instance, the *Science* page links only to science related pages), they were not used for calculating the link information measures in our experiments. For the same reason, all the pages found in the TodoBR collection that were *similar* to Cadê pages were removed. We consider a page in TodoBR similar to a page in Cadê if they share 70% or more of their out-links. The removed pages were, usually, copies of the Cadê directory. Pages from directories other than Cadê were also discarded because they share many out-links with Cadê. Despite these directories being different from Cadê, the large number of links in common, particularly in very specific categories, indicates that they might have started as copies of Cadê which evolved independently. Thus, our methods might be benefited by using this additional information[3]. However, the limitation of this strict definition of copy is obvious since many pages may have been unnecessarily removed.

Table 5.1 also shows that links from external pages provide the richest source of information. About 90% of the Cadê pages are linked to by external pages while less than 4% link to external pages. This was an important

[3]In fact, these pages had gone unnoticed in previous works, which caused a slight increase in the precision of the results for the link based classifiers.

reason for using Cadê in our experiments. With Cadê, we can obtain information about external links extracted from TodoBR and verify how useful this information can be for the classification process. This is only possible because Cadê is a subset of TodoBR, which is a large collection containing most of the link information available in Brazilian Web pages. This is not the case with most other classification collections where, in order to obtain more link information, it would be necessary to collect a huge amount of Web pages, or to have access to another search engine database, as we did with TodoBR.

Finally, it should be noted that most pages have no out-links at all (neither internal nor external), but the majority does have in-links (either internal or external).

Ad Collection

To evaluate our methods for combining conceptual and syntactical evidence, we extracted from the ad collection presented in Section 4.3.1 a subset of ads in which the landing pages were classified in the Brazilian Yahoo Web directory[4]. All these ads were manually reviewed to minimize misclassifications, which resulted in a taxonomy with 508 categories and an ad collection with 72,097 ads, 1,152 advertisers, and 54,882 keywords. In average, each ad was classified under 1.11 concepts.

Figure 5.3 shows the category distribution for the ad taxonomy. Compared to Cade12 and Cade188, it has a much more skewed distribution. The six most popular categories (*Computers*, *Culture/Music*, *Tourism*, *Culture/Movies*, *Health & Care*, and *Computers/Internet*) represent more than 50% of all documents. The most popular category, *Computers*, has 20,133 ads.

5.4.2 Experimental Methodology and Setup

In this section, we describe the methodology and setup used in our experiments in Web classification and combination of syntactical and conceptual matching strategies.

Web Document Classification

To perform the experiments, we used the 10-fold cross validation method [122]. Each dataset was randomly split in ten parts, such that, in each run, a different part was used as a test set while the remaining were used as a training

[4] `http://www.yahoo.com/`

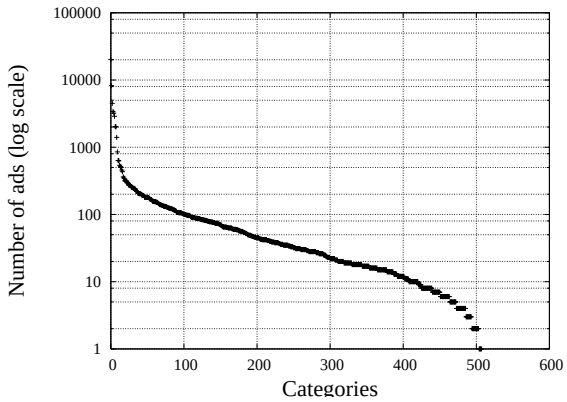


Figure 5.3: Category distribution for advertisement taxonomy.

set. The split on training and test sets was the same in all experiments. The final results of each experiment represent the average of the ten runs. The 10-fold cross-validation method was chosen because it minimizes the likelihood that the differences in the results of the algorithms being compared is due to a bias in the training set. Since we are only interested in comparing the performance of different classification strategies, this method is appropriate. In a different application, however, other training set selection methods can be used [65].

To make sure that the results are not biased by an inappropriate choice of parameters, several experiments were performed and, in all cases, we report the best results obtained. Thus, for the *kNN* classifier, the value of k was set to 30 and 15,000 features were considered. For the SVM classifier, a linear kernel was used and 10,000 features were considered. For the Naive Bayes classifier, 15,000 features were considered. For all algorithms, the features were selected using the information gain method [86]. A feature corresponds to an individual word in a document, after stop word removal.

The performance of the presented methods was evaluated using the conventional precision, recall and F_1 measures. Micro-averaging and macro-averaging [142] were applied to get single performance values over all classification tasks. For macro-averaging, recall, precision, and F_1 scores were first computed for individual categories and then averaged over all categories. For micro-averaging, the decisions for all categories were counted in a joint pool. We note that, since the datasets used in the experiments have a single label per document, micro-averaged recall, precision and F_1 are the same.

Combination of Syntactical and Conceptual Approaches

As in the previous chapter, we evaluate our methods essentially by considering their performance in terms of relevance. For this, we adopt the same pooling method described in that chapter with two differences. First, we selected the top *ten* (instead of three) ranked ads provided by each of our ad placement strategies. Second, the ads in each pool were submitted to a manual evaluation by a group of 21 users (instead of 15). As a consequence, the resulting pools contained an average of about 41.3 ads and the average number of relevant ads per pool was 24.54.

As before, we present the results considering that a triggering page offers three ad slots and the studied methods always attempt to fill all the slots. The performance of the methods is quantified through the metrics PAVG and PAVG@3, and the precision at the first, second, and third top-ranked ad (cf. Section 2.6.1). All the results to be presented were found statistically significant, when tested with the two-tailed paired t-test, at least at the 90% confidence level.

As the ad and the triggering page collections were classified under different taxonomies, a manual mapping of these taxonomies was carried out to allow the conceptual matching of pages and ads. In particular, to minimize possible problems related to the high heterogeneity of first-level classes *Services* and *Society*, we selected for mapping the second-level Cadê categories (Cade188 taxonomy). The resulting map is presented in Appendix A.

5.4.3 Web Classification Evaluation

The Web classification experiments were divided into two main parts. In the first, we evaluate the performance of the proposed link-based similarity measures when classifying a set of Web pages. In the second, we combine the results obtained by the link-based similarity measures with those of traditional text-based classifiers and study the effects of such combination. To end this section, we present a short summary and discussion of the empirical findings achieved through our experiments.

Link-based Similarity Measures

Table 5.2 shows the F_1 figures for the five different similarity measures, obtained in the Cade12 and Cade188 collections, using the *kNN* algorithm described in Section 5.2.1. Different values of k were tested and only the best results obtained are shown. In particular, a value of $k = 30$ was used for the Cade12 and a value of $k = 10$ was used for Cade188. Only internal

links were considered. As a baseline for comparison we show the results for the three text-based classifiers described in Section 5.2.3.

Classifier	Cade12		Cade188	
	micF_1	macF_1	micF_1	macF_1
Amsler	**23.03**	**8.31**	**9.32**	**2.90**
Bibliographic Coupling	21.80	4.67	8.29	1.66
Co-citation	22.40	6.94	9.01	2.68
Companion-authority	22.65	7.85	8.94	2.62
Companion-hub	22.33	6.28	8.67	2.43
kNN	50.03	44.50	37.45	**28.56**
SVM	**54.18**	48.41	**38.08**	26.08
Naive Bayes	53.09	**49.78**	36.48	18.61

Table 5.2: Micro-averaged and macro-averaged F_1 measures obtained with Cade12 and Cade188 collections, using different link-based similarity measures. Only internal links were used.

We observe that all the results were below the text-based baseline values. By considering only internal links, much of the link structure information of the collection is lost since, as shown in Table 5.1, about 98% of the link information in the collection comes from external pages.

Table 5.3 shows the F_1 figures for the Cade12 and Cade188 collections using both internal and external links. In both collections, the best results were achieved with values of $k = 30$ for the bibliographic coupling and Companion-hub classifiers and $k = 100$ for the co-citation, Amsler and Companion-authority similarities.

Classifier	Cade12		Cade188	
	micF_1	macF_1	micF_1	macF_1
Amsler	**68.77**	75.62	60.62	67.82
Bibliographic Coupling	22.09	5.39	8.57	2.03
Co-citation	68.76	**75.75**	**60.77**	**67.85**
Companion-authority	65.85	71.07	56.69	62.79
Companion-hub	26.95	16.10	12.32	6.58
kNN	50.03	44.50	37.45	**28.56**
SVM	**54.18**	48.41	**38.08**	26.08
Naive Bayes	53.09	**49.78**	36.48	18.61

Table 5.3: Micro-averaged and macro-averaged F_1 measures obtained with Cade12 and Cade188 collections, using different link-based similarity measures. Both internal and external links were used.

The figures for Amsler, co-citation, and the Companion algorithm using authority degrees are well above the baseline. We observe gains over the best text-based classifier (SVM) as high as 26.9% and 59.5% points in micro-averaged F_1, for the Cade12 and Cade188 collections, respectively. Gains in macro-averaged F_1 were even higher, reaching a maximum of 56.4% and 160.1% for the Cade12 and Cade188 collections, respectively. Bibliographic coupling and the Companion algorithm using hub degrees, on the other hand, are still below the text-based baseline.

Since most of the links are *from* external pages *to* pages in the collection, i.e., they are in-links to the pages in the collection, we can expect measures that make use of in-links to perform the best. Thus, co-citation benefits greatly from such information. The same happens for the Companion algorithm using authority degrees and for the Amsler similarity. These last two, however, suffer from the fact that they also rely on out-links, which are not so widely available.

It is interesting to note that, for all link-based measures with significant gains, macro-average F_1 is always higher than micro-averaged F_1. The inverse happens for the text-based classifiers, where micro-averaged F_1 values are the highest. We also observe that, for the text-based classifiers, F_1 values are much higher in the Cade12 collection than the Cade188 collection, whereas for the link-based similarity measures, values are similar in both collections. This indicates that link-based classification is much less affected by the class distribution than text-based classification. All link-based measures were capable of distributing documents among all classes, independently of their size (as indicated by the macro-averaged values) and achieved a similar performance on two collections with different class distributions.

Combination of Results

We now verify the effect of combining both types of evidence using the proposed Bayesian Network model. To study how different weights affect the classification results, we applied Equation (5.3) and varied the weights W_t and W_ℓ for the text-based and link-based types of evidence. Since using only internal link information yielded very poor results, we now proceed by using both internal and external links. For the same reason, we will also disconsider the bibliographic coupling and Companion-hub similarity measures.

Figures 5.4, 5.5, and 5.6 show the resulting F_1 values for the combination of each link-based similarity measure on the Cade12 collection. In all graphs, the xx axis shows the ratio between the link-based evidence weight and the text-based evidence weight (W_t/W_ℓ), on a logarithmic scale. The baselines are shown as horizontal lines.

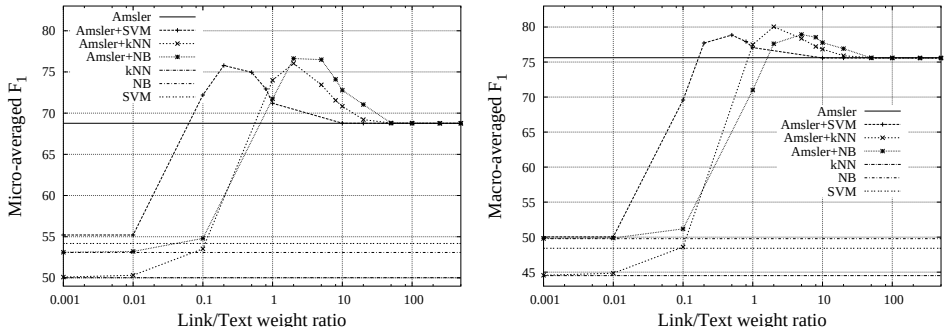


Figure 5.4: Effects of weighted combination for the Amsler similarity in the Cade12 collection.

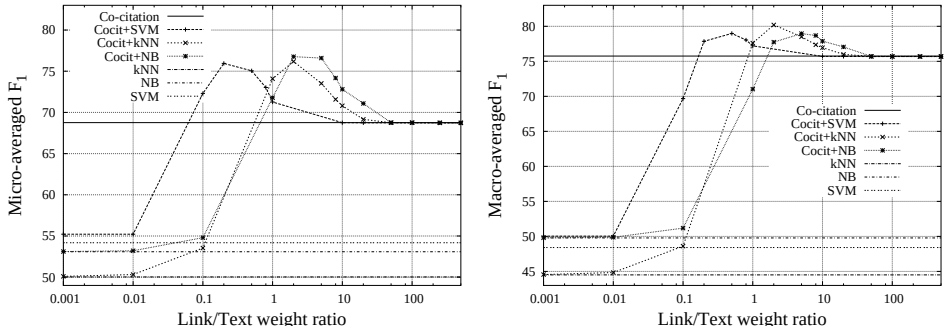


Figure 5.5: Effects of weighted combination for the co-citation similarity in the Cade12 collection.

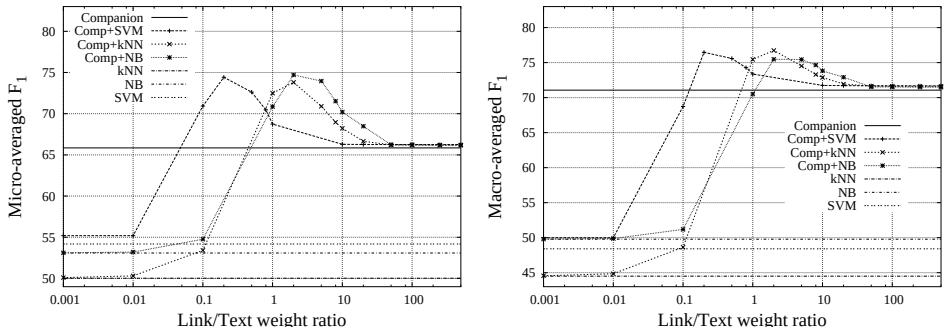


Figure 5.6: Effects of weighted combination for the Companion-authority similarity in the Cade12 collection.

As expected, results improve when the weight given to the link-based evidence is increased. In all cases, combination results always show improvements over the link baseline. When combined with the *kNN* and Naive Bayes classifiers, gains in F_1 achieve the highest values for link weights between 2 and 5 times greater than content weights. Combination with the SVM classifier yields the highest F_1 values at link weights from 2 to 5 times lower than content weights.

The use of different weights can be seen as having two separate effects. First, it balances the importance of the sources of evidence, guaranteeing that each one will have the appropriate influence on the final classification decision. Second, it normalizes the probability values given by each classifier, thus eliminating the problems usually derived from combining values with different orders of magnitude.

In terms of micro-averaged F_1, the evidence combination was able to achieve gains up to 13.44% over the best results of link-based classification alone. In terms of macro-averaged F_1, however, gains were less expressive, going only up to 7.96%. This indicates that the combination, although being capable of further improvements, favors the classification of documents in the most popular classes. This can be seen as the result of mixing characteristics from both the text-based and the link-based algorithms—the text-based classifiers are capable of correctly classifying more documents in popular classes, while the link-based classifiers avoid the less popular classes from being impaired.

Results for the Cade188 collection are shown in the graphs of Figures 5.7, 5.8, and 5.9. For the *kNN* and Naive Bayes classifiers, results follow a similar pattern, with the highest F_1 values being reached at link weights from 2 to 8 times greater than content-weights. Gains in micro-averaged F_1 are as high as 13.18%. Due to the highly skewed class distribution, gains in macro-averaged F_1 were very small, achieving at most 5.73%.

For the combination with the SVM classifier, however, gains are insignificant or non-existent. When classifying a document, an effect of the voting scheme used by the SVM classifier is that the probabilities given to different classes are very similar. Thus, even a very small probability value yielded by the link-based classifier is enough to determine the final combined classification decision. As a result, the link-based classifier dominates the combination. This effect cannot be compensated by the use of weighted evidence because the weights can change the order of magnitude of the probabilities but not the differences between them.

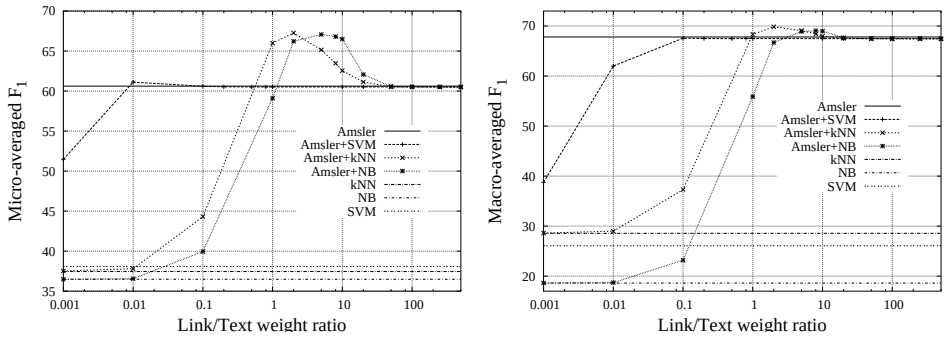


Figure 5.7: Effects of weighted combination for the Amsler similarity in the Cade188 collection.

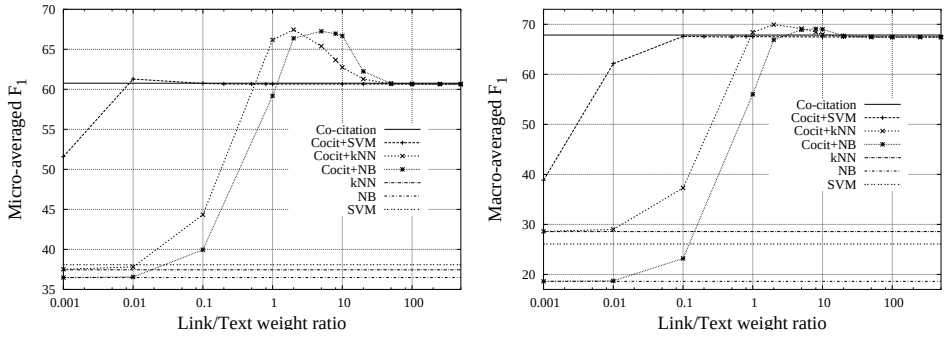


Figure 5.8: Effects of weighted combination for the co-citation similarity in the Cade188 collection.

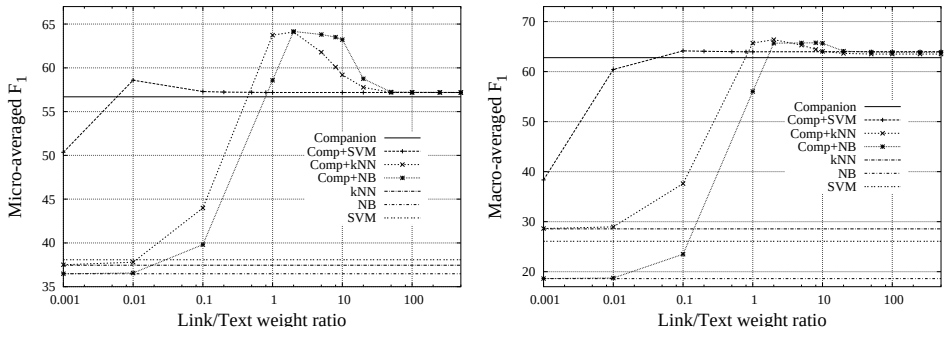


Figure 5.9: Effects of weighted combination for the Companion-authority similarity in the Cade188 collection.

Summary and Discussion

Table 5.4 summarizes the best F_1 values obtained by each of the considered evidence combinations, in the Cade12 collection. Table 5.5 summarizes the best F_1 values obtained by each combination in the Cade188 collection.

	Sim. Measures	*kNN*	SVM	NB
micF_1	Amsler	76.04	75.79	76.63
	Co-citation	**76.18**	**75.95**	**76.77**
	Companion	73.80	74.42	74.70
macF_1	Amsler	80.05	77.71	77.62
	Co-citation	**80.20**	**77.86**	**77.74**
	Companion	76.73	76.47	75.48

Table 5.4: Best F_1 values obtained in all the experiments, in the Cade12 collection.

	Sim. Measures	*kNN*	SVM	NB
micF_1	Amsler	67.26	61.12	67.07
	Co-citation	**67.45**	**61.28**	**67.25**
	Companion	64.10	58.59	64.16
macF_1	Amsler	69.85	62.00	68.90
	Co-citation	**69.92**	**62.15**	**68.90**
	Companion	66.39	60.43	65.68

Table 5.5: Best F_1 values obtained in all the experiments, in the Cade188 collection.

All link-based similarity measures were capable of capturing pertinent information regarding the classes in both collections. The co-citation measure performed consistently better than the remaining, providing the highest F_1 values, independently of the text-based classifier used. For any given link-based similarity measure, results show little variation when combined with different text-based classifiers, which confirms the importance of link information on the classification process.

Increases in F_1 over the use of link-based classification alone are quite small, indicating that there is little to gain by combining it with the results of the text-based classifiers. However, text-based classification should not be discarded. The small improvements obtained may still be useful in applications where high precision and recall are essential. Further, we can expect that the use of more effective text-based classification algorithms will achieve higher gains than those obtained in this work.

Text-based classifiers tend to be highly dependent on the document per class distribution. The link-based classifiers here proposed, however, were shown to perform equally well on collections with different distributions. Although further experimentation should be performed before reaching a definitive conclusion, this clearly indicates that link information is useful to compensate the bias towards popular classes introduced by text-based algorithms.

Finally, one major issue that should be considered when combining classification results is that of normalization. Values provided by different sources of evidence can be of a very distinct nature. They can be in different scales or of a different order of magnitude. Further, even similar values might have totally different meanings. In particular, in the case of SVM, a possible solution could be the employment of methods for estimating class pertinence probabilistically. Authors in [139] have compared several such methods and suggested new approaches based on pairwise coupling for multi-classification settings using SVM. They concluded that their methods are more stable than voting in providing probability estimates.

5.4.4 Evaluation of Classification in Content-targeted Advertising

We now evaluate the use of classification in content-targeted advertising. We start by defining the information to be combined through Equations (5.7) and (5.8), that is, the syntactical similarity denoted by function *ssim* and the conceptual similarity denoted by function *csim*. As *ssim* function, we use method `AAK`, the baseline in Chapter 4. As for the *csim* function, we use the category vectors provided by the human experts[5] as well as the soft and hard category vectors obtained by means of the automatic text-based, link-based, and combined-based classifiers. By using these four classifiers we are able to analyze the performance of the proposed combination approaches in scenarios where a manual classification is either available or not. For the case where only automatic classification is available we study the impact of classifiers of progressively better quality.

In particular, we selected *kNN* as the text-based classifier because SVM and Naive-Bayes provide less reliable probability estimates. In the case of SVM, the voting scheme used yields very similar probabilities for all categories making it hard to distinguish them. In the case of Naive Bayes, the required normalizing scheme bears probabilities very close to either 0 or 1 and, consequently, not appropriate for combination. Since co-citation and co-

[5]Notice that these vectors comprise hard mono-classification decisions.

Method	Description
$\texttt{AAK\_TXT}_{h,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{hard,text}(a_j, d_i)$
$\texttt{AAK\_LNK}_{h,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{hard,link}(a_j, d_i)$
$\texttt{AAK\_CMB}_{h,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{hard,combined}(a_j, d_i)$
$\texttt{AAK\_MAN}_{h,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{hard,manual}(a_j, d_i)$
$\texttt{AAK\_TXT}_{s,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{soft,text}(a_j, d_i)$
$\texttt{AAK\_LNK}_{s,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{soft,link}(a_j, d_i)$
$\texttt{AAK\_CMB}_{s,a}(a_j, d_i)$	$\texttt{AAK}(a_j, d_i) \times csim_{soft,combined}(a_j, d_i)$
$\texttt{AAK\_TXT}_{h,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{hard,text}(a_j, d_i))$
$\texttt{AAK\_LNK}_{h,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{hard,link}(a_j, d_i))$
$\texttt{AAK\_CMB}_{h,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{hard,combined}(a_j, d_i))$
$\texttt{AAK\_MAN}_{h,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{hard,manual}(a_j, d_i))$
$\texttt{AAK\_TXT}_{s,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{soft,text}(a_j, d_i))$
$\texttt{AAK\_LNK}_{s,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{soft,link}(a_j, d_i))$
$\texttt{AAK\_CMB}_{s,n}(a_j, d_i)$	$1 - (1 - \texttt{AAK}(a_j, d_i))\ (1 - csim_{soft,combined}(a_j, d_i))$

Table 5.6: Combination methods.

citation combined with *kNN* provided the better results in the classification experiments, we use them as the link-based and combined-based classifiers, respectively.

To reflect the different kinds of *category vectors* and *classifiers* taken into consideration by function $csim(a_j, d_i)$, from now on, we will refer to this function as $csim_{cv,classifier}(a_j, d_i)$. The *cv* subscript will assume the values "hard" or "soft" to indicate hard or soft classification, respectively. Similarly, the $classifier$ subscript will assume the values "text", "link", "combined", or "manual" to indicate the *kNN*, co-citation, cocitation-plus-*kNN*, or manual classifiers used to obtain the category vectors. Table 5.6 describes all the resulting combination methods to be evaluated in the following sections.

Impact of the Classification

We start our analysis by verifying if the concept-based similarity, taken in isolation, can be used as an ad ranking as good as that provided by baseline method AAK. For this, Table 5.7 presents the performance figures for the ad rankings obtained through the manual and the automatic classifiers using hard and soft classification.

We notice that there is no case where the ad ranking provided by the baseline is worse than those provided by the automatic classifiers. In fact, the gain of the baseline is not statistically significant only in the case of the PAVG@3 metric for the two best classifiers, i.e., manual with hard classification ($csim_{hard,manual}$) and co-citation with soft classification ($csim_{soft,cocit}$).

Methods	Precision			PAVG			PAVG@3		
	@1	@2	@3	score	gain(%)	cl(%)	score	gain(%)	cl(%)
baseline									
AAK	0.640	0.620	0.600	0.173	-	-	0.552	-	-
hard classification									
$csim_{hard,kNN}$	0.448	0.410	0.398	0.106	-38.7	99	0.346	-37.3	99
$csim_{hard,cocit}$	0.610	0.555	0.533	0.141	-18.5	99	0.484	-12.3	99
$csim_{hard,cocit+kNN}$	0.600	0.550	0.533	0.138	-20.2	99	0.481	-12.9	95
$csim_{hard,manual}$	0.628	0.574	0.560	0.147	-15.0	95	0.504	-8.7	-
soft classification									
$csim_{soft,kNN}$	0.361	0.423	0.399	0.100	-42.2	99	0.345	-37.5	99
$csim_{soft,cocit+kNN}$	0.520	0.505	0.510	0.131	-24.3	98	0.459	-16.8	90
$csim_{soft,cocit}$	0.545	0.520	0.532	0.137	-20.8	95	0.475	-13.9	-

Table 5.7: Performance for the ad rankings obtained through classifiers taken in isolation. Note that *cl* stands for t-test confidence level.

We also observe that the gains obtained with the text classifier are smaller than those obtained with the link-based and the combined-based classifiers that, by their turn, are smaller than those obtained with manual classification. This leads us to conclude that the more accurate the classifier, the greater the ranking precision. Notice that there is no significant difference neither between the link-based and the combined-based classifiers nor between hard and soft classification.

This situation changes, however, when we combine the conceptual information with method AAK. Table 5.8 presents the performance figures for the ad rankings obtained through the combination of the classification evidence with baseline method AAK. As we can see, several combination methods present performance gains over AAK. In particular, when considering the PAVG@3 metric, the best method involving manual classification, $\texttt{AAK\_MAN}_{h,a}$, presents a gain of 32.8% over the baseline whereas the best method involving automatic classification, $\texttt{AAK\_CMB}_{s,a}$, presents a gain of 22.8%.

As previously observed in Table 5.7, gains tend to be better as the combined classifiers are more accurate, at least for PAVG@3. The noticeable exception is method $\texttt{AAK\_MAN}_{h,n}$. This happened because, in that case, AAK had almost no impact on the combination. Since the manual classifier only provides hard decisions (probability values equal to 1), after the noisy-or combination, the top ranked ads are exactly those top ranked by the manual classifier[6]. In other words, the $\texttt{AAK\_MAN}_{h,n}$ ranking is practically reduced to the $csim_{hard,manual}$ which, as we can see in Table 5.8, presents no gain over the baseline.

Table 5.8 also shows that combination methods involving the text-based

[6]In fact, there is no statistically significant difference between the performances of methods $\texttt{AAK\_MAN}_{h,n}$ and $csim_{hard,manual}$.

Methods	Precision			PAVG			PAVG@3		
	@1	@2	@3	score	gain(%)	cl(%)	score	gain(%)	cl(%)
baseline									
AAK	0.640	0.620	0.600	0.173	-	-	0.552	-	-
and combination, hard classification									
AAK_TXT$_{h,a}$	0.672	0.641	0.594	0.160	-7.5	99	0.559	+1.3	99
AAK_CMB$_{h,a}$	0.788	0.753	0.714	0.195	+12.7	-	0.684	+23.9	-
AAK_LNK$_{h,a}$	0.821	0.769	0.739	0.209	+20.8	-	0.710	+28.6	-
AAK_MAN$_{h,a}$	0.822	0.806	0.767	0.218	+26.0	-	0.733	+32.8	99
and combination, soft classification									
AAK_TXT$_{s,a}$	0.688	0.667	0.639	0.183	+5.8	-	0.596	+8.0	-
AAK_LNK$_{s,a}$	0.753	0.706	0.691	0.200	+15.6	-	0.654	+18.5	95
AAK_CMB$_{s,a}$	0.750	0.745	0.720	0.207	+19.7	99	0.678	+22.8	99
noise-or combination, hard classification									
AAK_TXT$_{h,n}$	0.680	0.650	0.633	0.182	+5.2	-	0.583	+5.6	-
AAK_CMB$_{h,n}$	0.730	0.705	0.697	0.206	+19.1	98	0.652	+18.1	98
AAK_LNK$_{h,n}$	0.750	0.685	0.690	0.206	+19.1	99	0.641	+16.1	98
AAK_MAN$_{h,n}$	0.520	0.525	0.537	0.149	-13.9	-	0.457	-17.2	90
noise-or combination, soft classification									
AAK_TXT$_{s,n}$	0.670	0.665	0.640	0.179	+3.5	-	0.591	+7.1	-
AAK_LNK$_{s,n}$	0.750	0.695	0.680	0.206	+19.1	99	0.633	+14.7	95
AAK_CMB$_{s,n}$	0.710	0.710	0.707	0.206	+19.1	98	0.655	+18.7	98

Table 5.8: Performance of ad rankings obtained through the combination of conceptual (classifier decisions) and syntactical information (method AAK). Note that *cl* stands for t-test confidence level.

classifier never outperformed the baseline. In general, this is not the case for combination involving the best classifiers, especially in the case of the noisy-or combination. Methods involving the link-based and the combined-based classifiers present only one significant difference when compared: AAK_CMB$_{s,a}$ is better than AAK_LNK$_{s,a}$ with a gain of 3.5% at a 95% confidence level in PAVG and 3.7% at a 98% confidence level in PAVG@3. This implies that, at least for content-targeted advertising, the gain in accuracy of the combined-based classifier over the link-based classifier is not large enough to make the first a better choice than the second.

And versus noisy-or combination, Hard versus Soft Classification

An intriguing issue in Table 5.8 is the performance presented by the link-based classifier in the top of the ranking when *and* and *noisy-or* combinations are compared. For instance, the gain of AAK_LNK$_{h,a}$ in PAVG@3 is about 29% while the gain of AAK_LNK$_{h,n}$ is only 16.1%. Despite the gain obtained with AAK_LNK$_{h,a}$ (*and* combination) being better, it is not statistically significant. The reverse is observed for noisy-or combination. At first, this seems to indicate that noisy-or combination is a better alternative for automatic classifiers. However, the situation changes when we analyze the performance of the combination methods in a collection composed only by

Methods	All documents			Only hits		
	score	gain(%)	cl(%)	score	gain(%)	cl(%)
baseline						
AAK	0.552	-	-	0.583	-	-
combinations						
AAK_LNK$_{h,a}$	0.710	+28.6	-	0.795	+36.4	99
AAK_LNK$_{s,a}$	0.654	+18.5	95	0.771	+32.2	99
AAK_LNK$_{h,n}$	0.641	+16.1	98	0.732	+25.5	99
AAK_LNK$_{s,n}$	0.633	+14.7	95	0.737	+26.4	99

Table 5.9: PAVG@3 figures obtained through the combination of link-based classifier decisions and baseline method AAK. Note that *cl* stands for t-test confidence level.

correctly classified pages (which we call *hits*), as shown in Table 5.9.

Table 5.9 presents the gains in PAVG@3 over AAK obtained by the combination methods involving link-based classifiers. The gains were calculated for the entire collection as well as for a collection composed only by pages which were correctly classified (hits). As expected, the gains increased for the hits. Further, we note that the confidence levels also increased making significant even the gain for AAK_LNK$_{h,a}$. This indicates that the combination methods perform very differently for correctly classified and misclassified pages. We also notice that this difference is smaller for AAK_LNK$_{h,a}$ which suggests that soft classification and noisy-or combination yield lower quality rankings. This is reinforced by Tables 5.10 and 5.11 which show the difference between *and* and noisy-or combination and hard and soft classification.

Table 5.10 shows that, for hard classification and PAVG@3 metric, *and* combination is worse than noisy-or combination only for the text-based classifier. This is due to the poor performance of the classifier. When we consider the PAVG metric, we notice that *and* combination is also worse for the combined-based classifier. For all other cases, however, the better the classifier, the larger the gain obtained through *and* combination. For soft classification, there is no significant difference between *and* or noisy-or combinations. This is due to the fact that soft classification introduces many irrelevant ads.

Table 5.11 shows that, for *and* combination and PAVG@3 metric, soft classification is clearly worse than hard classification for the text-based classifier and clearly better for the link-based classifier. Notice that, for noisy-or combination, no significant difference between soft and hard classification is observed. Similarly to the soft classification in Table 5.10, the noisy-or combination introduces many irrelevant ads.

In summary, *and* combination and hard classification seem to be better options when we are dealing with very accurate classifiers, such as the manual one. For automatic classifiers, the performance of *and* combination and hard

Methods	Precision			PAVG			PAVG@3		
	@1	@2	@3	score	gain(%)	cl(%)	score	gain(%)	cl(%)
hard classification									
AAK_TXT$_{h,n}$	0.680	0.650	0.633	0.182	-	-	0.583	-	-
AAK_TXT$_{h,a}$	0.672	0.641	0.594	0.160	-12.1	99	0.559	-4.1	99
AAK_CMB$_{h,n}$	0.730	0.705	0.697	0.206	-	-	0.652	-	-
AAK_CMB$_{h,a}$	0.788	0.753	0.714	0.195	-5.3	99	0.684	+4.9	98
AAK_LNK$_{h,n}$	0.750	0.685	0.690	0.206	-	-	0.641	-	-
AAK_LNK$_{h,a}$	0.821	0.769	0.739	0.209	+1.5	99	0.710	+10.8	99
AAK_MAN$_{h,n}$	0.520	0.525	0.537	0.149	-	-	0.457	-	-
AAK_MAN$_{h,a}$	0.822	0.806	0.767	0.218	+46.3	99	0.733	+60.4	99
soft classification									
AAK_TXT$_{s,n}$	0.670	0.665	0.640	0.179	-	-	0.591	-	-
AAK_TXT$_{s,a}$	0.688	0.667	0.639	0.183	+2.2	-	0.596	+0.8	-
AAK_LNK$_{s,n}$	0.750	0.695	0.680	0.206	-	-	0.633	-	-
AAK_LNK$_{s,a}$	0.753	0.706	0.691	0.200	-2.9	-	0.654	+3.3	-
AAK_CMB$_{s,n}$	0.710	0.710	0.707	0.206	-	-	0.655	-	-
AAK_CMB$_{s,a}$	0.750	0.745	0.720	0.207	+0.5	-	0.678	+3.5	-

Table 5.10: Comparison between *and* and *noisy-or* combination strategies. Note that *cl* stands for t-test confidence level.

Methods	Precision			PAVG			PAVG@3		
	@1	@2	@3	score	gain(%)	cl(%)	score	gain(%)	cl(%)
and combination									
AAK_TXT$_{h,a}$	0.672	0.641	0.594	0.160	-	-	0.559	-	-
AAK_TXT$_{s,a}$	0.688	0.667	0.639	0.183	+14.4	99	0.596	+6.6	99
AAK_CMB$_{h,a}$	0.788	0.753	0.714	0.195	-	-	0.684	-	-
AAK_CMB$_{s,a}$	0.750	0.745	0.720	0.207	+6.2	99	0.678	-0.9	99
AAK_LNK$_{h,a}$	0.821	0.769	0.739	0.209	-	-	0.710	-	-
AAK_LNK$_{s,a}$	0.753	0.706	0.691	0.200	-4.3	99	0.654	-7.9	99
noise-or combination									
AAK_TXT$_{h,n}$	0.680	0.650	0.633	0.182	-	-	0.583	-	-
AAK_TXT$_{s,n}$	0.670	0.665	0.640	0.179	-1.6	-	0.591	+1.4	-
AAK_CMB$_{h,n}$	0.730	0.705	0.697	0.206	-	-	0.652	-	-
AAK_CMB$_{s,n}$	0.710	0.710	0.707	0.206	0.0	-	0.655	+0.5	-
AAK_LNK$_{h,n}$	0.750	0.685	0.690	0.206	-	-	0.641	-	-
AAK_LNK$_{s,n}$	0.750	0.695	0.680	0.206	0.0	-	0.633	-1.2	-

Table 5.11: Comparison between hard classification and soft classification strategies. Note that *cl* stands for t-test confidence level.

classification may be very irregular for misclassified pages. Therefore, in that case, soft classification and noisy-or combination are better options since they can lead to smaller, but stable gains.

5.5 Conclusions

In this chapter we studied how to improve automatic Web document classification and how to apply this classification to content-targeted advertising. In particular, we compared five different similarity measures based on link structure and explored their combination with traditional text-based classifiers. Finally, we used the conceptual evidence provided by our best classifiers to improve a syntactical matching method designed for content-targeted advertising.

Experiments performed on two collections of Web directory pages allow us to conclude that links are indeed a valuable source of information for Web classification. However, links that are internal to the collection do not provide sufficient information. In order to achieve expressive results, links to and from pages outside the directory should be used. Also, most external pages are parents of the pages in the collection, i.e., they have a link to the pages in the collection. Thus, similarity measures that make use of in-link information are expected to be the most appropriate.

Of the tested link-based similarity measures, co-citation and Amsler yielded the best results, with values of about 69% and 61% in micro F_1, in the Cade12 and Cade188 collections, respectively. The Companion algorithm using authorities showed a slightly inferior performance, with about 66% and 57% in micro F_1, in the Cade12 and Cade188 collections, respectively. These values represent gains over the best text-based classifier (SVM) of up to 27% in Cade12 and 62% in Cade188. Bibliographic coupling and the Companion algorithm using hubs showed a much inferior performance, due to the fact that they rely on out-link information, whereas most pages in the collection have no out-links.

By combining text-based and link-based sources of evidence, micro F_1 values of up to 77% and 67% were achieved, in the Cade12 and Cade188 collections, respectively. Further, even though the gain in macro-averaged F_1 has been quite small, the increase observed in both F_1 metrics indicates that more documents were correctly classified, while the trend to impair classification on the least popular classes, expected in the text-based classifiers, was avoided.

Other important conclusions were drawn from the observed results. First, link-based similarity measures achieved a similar performance on two collec-

tions with different document distributions. This indicates that, unlike text-based classifiers, they are more independent of class size. Second, weighted evidence combination was essential, not only to balance the importance of each source of evidence, but also to normalize the values to be combined. In the case of the SVM classifier a more complex normalization procedure is still required, showing that normalization is an essential issue when combining distinct sources of evidence.

Experiments with a real ad collection showed that a content-targeted advertising method based only on a conceptual similarity metric is not able to outperform one based only on a syntactical metric. However, a method obtained through the combination of conceptual and syntactical metrics yielded gains in precision of about 23% and 33% when using an automatic and a manual classifier, respectively. As expected, the best results were obtained when the best classifiers were used. We also learned that the gain obtained by the combined-based over the link-based classifier was not significant enough to make the former a better source of conceptual information than the latter.

Finally, the study in how to treat the classifier decisions and how to combine conceptual and syntactical information led us to conclude that *and* combination and hard classification are strategies more appropriate to accurate classifiers, such as the manual classifier. In the case of the automatic classifiers, soft classification and noisy-or combination will provide smaller, but more stable gains.

Chapter 6

Conclusions and Future Work

In this chapter, we summarize our accomplishments and present final conclusions. We also suggest new research venues which address open questions left by our study as well as ideas arisen during the course of the work.

6.1 Conclusions

In this work, we studied how to improve the precision of matching algorithms used in content-targeted advertising. For this, (a) we analyzed the impact on syntactical matching of the different evidence sources already available to information gatekeepers that operate keyword targeted advertising systems, (b) we studied how to deal with the vocabulary mismatch found between the Web pages and the ads, and (c) we explored the user trend, characteristic of content-targeted advertising, to consider relevant associations of broader nature.

Regarding the use of different evidences, we studied the impact on matching of the title and description of the ads, the contents of the pages pointed to by the ads, and the keywords used by the advertisers to describe their ads. Regarding the vocabulary mismatch problem, we proposed how to expand Web pages with new terms to facilitate the task of matching ads and triggering pages. In both cases, we used the traditional vector space ranking strategy to match ads and triggering pages. Experiments performed on a real ad collection, using an online newspaper as source of triggering pages and a sample Web collection as source of general Web pages, allowed us to draw the following conclusions:

1. Ad keywords are essential to improve ranking accuracy. In particular, by enforcing their appearance we were able to improve precision. In

fact, these observations were also supported by experiments performed on a collection of pages extracted from a Web directory;

2. Despite being of less importance, the text of the ad contributes to improve ranking accuracy;

3. The contents of the pages pointed to by the ads had a negligible impact on the ranking;

4. Better rankings are possible by expanding the contents of the triggering pages with terms gathered from Web pages of similar content.

Regarding the associations of broader nature, we studied how to improve Web page classification and how to use this classification information, as well as the category information available to the ads, for enhancing syntactical matching algorithms.

As for the classification problem, we studied how link information could be used to improve the results of traditional content-based classifiers because we believe that the hypertextual nature of the Web is a useful resource to indicate page's topic and importance. Four different linkage similarity measures, bibliographic coupling, co-citation, Amsler, and Companion, were used in combination with three traditional classifiers, Naive Bayes, *kNN*, and Support Vector Machine. The combination was achieved through a Bayesian network model. Experiments were performed in two different collections extracted from a Web directory. Results allowed us to draw the following conclusions:

1. Links are very effective to indicate if Web pages are similar which makes them very useful for Web classification. In fact, links alone performed better than any of the content-based classifiers on our test collections;

2. Since in Web pages that have links, in-links are more common than out-links, similarity measures should make use of in-link information. Measures which do not use in-links are likely to perform poorly in Web collections;

3. Combination of link-based and content-based information is useful to improve classification accuracy. Furthermore, the larger the importance given to the more reliable source of evidence, the larger the improvement obtained. Since in our test collections the more reliable source of evidence was the link-based classifier, our best results were achieved when we gave more weight to them.

Finally, by using the conceptual information provided by our best classifiers, we improved the precision of our methods for ranking ads. In particular, our probabilistic classifiers allowed us to investigate the impact of progressively better quality category information on the matching of ads to a Web page and determine a better form to treat and combine classification decisions. Experiments performed on a real ad collection, using `AAK` as baseline, the Cadê directory as source of triggering pages and a manually-built ad taxonomy, allowed us to draw the following conclusions:

1. Taking into consideration the conceptual information is useful to improve the precision of the ad ranking. When taken in isolation, however, the conceptual information performs, in general, worse than the syntactical information;

2. In general, the higher the accuracy of the classifier, the larger the ranking improvement obtained in the combination. In our experiments, the classifier with lowest accuracy – the text-based classifier which was able to classify correctly less than 40% of the test pages – was not able to improve the ranking quality. The best classifiers – link-based and combined-based classifiers which were able to classify correctly, at least, 67% of the test pages – presented significant gains over the baseline in most cases. Finally, in general, the best results in combination were obtained with manual classification;

3. While significant for the classification task, the gain in precision of the combined-based classifier over the link-based classifier was not large enough to make the combined-based classifier a better option for ad ranking combination;

4. For accurate classifiers, such as in the scenario where manual classification is available, *and* combination and hard classification should be used. For automatic classifiers, the performance of *and* combination and hard classification may be very irregular for misclassified pages. Therefore, in that case, soft classification and noisy-or combination could be better options since they may lead to more stable performance.

The methods proposed and the experiments performed allow us to provide answers, at least partially, to the research questions that motivated this work, as follows:

- The careful consideration of additional evidence, such as those already available to information gatekeepers that operate keyword-targeted advertising systems, can lead to enhancements in content-targeted advertising. In particular, except by the contents of the landing pages, all the other ad fields were useful for improving results. As expected, the keywords were the most valuable source of information especially when they were required to appear in the triggering page;

- At least for informational pages whose contents can be obtained with minimal noise, the mismatch between the vocabulary of an ad and the vocabulary of a Web page, frequently observed in content-targeted advertising, can be reduced by expanding the contents of these pages with terms extracted from other Web pages of similar content;

- We can improve classification accuracy of Web pages by combining link-based and content-based information, since links were found to be very effective to indicate if Web pages are similar. Notice that we drew this conclusion from experiments on Web directories. However, a Web directory is not representative of the whole Web. Thus, we do not claim that our methods will perform well on any Web page. However, we expect that they can be applied to other Web directories and any collection of popular pages because these pages, as those in a Web directory, probably have a high number of in-links [10, 64];

- Conceptual information is useful to improve the precision of ad placement. In particular, we can enhance matching algorithms in content-targeted advertising by combining the context-based ranking obtained through manual and automatic classification with the ranking provided by syntactical matching methods.

To conclude, besides being a very successful way of monetization in the Web, content-targeted advertising has introduced several challenging problems and raised some intriguing questions. Our work here addressed some of these questions, that is, how to improve the precision of ad placement systems in content-targeted advertising. Our results indicate that, by considering additional evidence, high quality content-targeted advertising is feasible and practical.

6.2 Future Work

In this section, we suggest how to continue this work addressing open questions left by our research and new ideas arisen during our study.

Investigation on other Evidence and Combination Approaches

Our study on matching algorithms took advantage of some pieces of evidence and some methods for combining them. In the following paragraphs, we suggest other evidence and combination approaches to be studied.

An evidence to be exploited in future works is the amount the advertiser is willing to pay for the placement of their ads. By taking into consideration this information, we will be able to study the main aspects of the content-targeted advertising problem. In fact, in order to deal with the way in which content-targeted advertising is sold today, we have to determine which keywords to consider after finding the best ads for a page. Notice this is particularly challenging for the cases in which the matching involved expansion strategies. Other evidence to take into account are the demographic, the contextual, and the user's click-through information. Demographic and contextual information are useful to filter out ads, to characterize the users to the advertisers, and to indicate the probable purpose of the users. For instance, ads can be selected according to the age, gender, life style, geographical location, and language of the user. Regarding the user's click-through information, it can be taken as a useful relevance judgement about the ads, the advertisers, and the quality of the traffic offered by the publishers.

Alternative ways to combine the evidence should also be tested. In this work, we studied the effects of weighted combination for document classification. However, our study was restricted to a linear combination of the pieces of evidence, as described in Section 5.2.2. A different evaluation of when and how much a piece of evidence should be favored over another must be performed. Similarly, despite we have shown that the use of different evidence can impact on the ad ranking, we were not able to determine how much importance should be given to each piece of evidence. In fact, for both document classification and ranking function design, experiments are already under way to explore alternative combination strategies. For instance, for document classification, we have tested a combination strategy in which the classifier to be used to determine the document category is defined according to the degree of belief expressed by the most reliable classifier. Preliminary results with this strategy have shown gains over our best combination method described here [25]. We also have investigated an evidence combination framework based on genetic programming to automatically learn combination functions. For document classification, preliminary results show gains of about 15% over our best combination function presented here and even over a combination framework based on composite kernels [143, 144]. For ad ranking function design, we evolved functions to take into consideration only the ad creative, their keywords, the content of their landing pages, and statis-

tics for ads alone or clustered in campaigns. Our results show that the best evolved function is able to perform 60% better than the `AAK` method when we consider the PAVG@3 metric [69]. We also intend to compare our GP method to other machine learning strategies such as that presented in [126].

Improvement of Keyword Processing in Matching Systems

During the course of this work, we studied the impact of different weighting schemes on matching algorithms. Preliminary results obtained showed that, amongst the several schemes analyzed, those that favor long keywords are the best. Such gains were greater when text segmentation was applied to the triggering pages even for very simple strategies, such as, paragraph segmentation. These results suggest us that to favor long keywords improves ranking especially when the words that compose such keywords occur close to each other in the triggering page. Motivated by these findings, we intend to study a matching system in which the weighting scheme favors long keywords and word distances are taken into account.

We also intend to exploit sub-phrases. As mentioned before (see Section 3.1.1), several content-targeted advertising systems in the industry offer to their users sophisticated options in their matching systems. This includes, among others, the use of exact and approximated matching and even the definition of sub-phrases. All these systems, however, require that the users define the matching method to be applied. For instance, the keywords *cars new york* and *cars "new york"* have different meanings. In second case, the advertiser clearly defined that *new york* should be treated as a sub-phrase. On the contrary of these approaches, we intend to use information already provided by the advertisers to automatically recognize patterns such as sub-phrases in the keywords. For that, we intend to use techniques based on maximal termsets [46] and query segmentation [105].

Evaluation

To evaluate the methods presented in this work, we essentially took into account their performance in terms of relevance. In spite of this approach being appropriate to compare the performance of these methods, it provides incomplete information about their actual effectiveness. Thus, alternative ways of evaluating these methods should be considered. More precisely, other information sources more indicative of conversion rates, such as click-through information, should be used.

Further, our evaluation was mainly based on collections composed of informational pages. However, to analyze the performance of the proposed

methods in more general contexts, evaluation with pages of distinct nature should be conducted. In particular, we intend to study our methods in collections composed of pages such as those found in blogs, forums, and emails.

Further Studies

Some aspects of the content-targeted advertising problem were not considered at all or require further investigation. Among them, we cite:

- Given the negative impact of irrelevant ads on the brand of publishers and advertisers, it should be interesting to investigate how to design functions that minimize the placement of irrelevant ads. This is especially important in cases where relevant ads are not available;
- Category affinity information could be exploited to facilitate the association between pages and ads belonging to conceptually close categories (e.g., *Internet* and *Computing*) or related to compatible services (e.g., *Tourism* and *Transports*);
- An evaluation of how much the syntactical evidence should be favored over the conceptual evidence must be performed because, in this work, we always treated these pieces of evidence as equivalent. This, however, is not the case since from our experiments with these pieces of evidence taken in isolation we concluded that the conceptual one was inferior;
- Similarly, an evaluation of how important is each structural part of the ad should be conducted since, in this work, we also treated these structural parts as equivalents;
- In this work, we studied the combination of conceptual matching strategies only with method `AAK`. It would be interesting to expand this investigation to the impedance coupling methods proposed in Chapter 5;
- To improve the syntactical matching methods, we expanded the contents of the triggering pages with terms extracted from similar pages. In a similar way, expansion strategies could be applied to other components of the content-targeted advertising problem. For instance, the ads could be expanded using terms extracted from similar ads or query logs. Further, different ways to cluster ads could be studied to take advantage of natural ad organizations, such as campaigns and groups of ads related to a same topic.

Bibliography

[1] Silvia Acid, Luis M. de Campos, Juan M. Fernández-Luna, and Juan F. Huete. An information retrieval model based on simple bayesian networks. *International Journal of Intelligent Systems*, 18(2):251–265, January 2003.

[2] Robert Amsler. Application of citation-based automatic classification. Technical report, The University of Texas at Austin, Linguistics Research Center, Austin, TX, December 1972.

[3] Vinod Anupam, Alain Mayer, Kobbi Nissim, Benny Pinkas, and Michael K. Reiter. On the security of pay-per-click and other web advertising schemes. In *WWW '99: Proceeding of the eighth international conference on World Wide Web*, pages 1091–1100, New York, NY, USA, 1999. Elsevier North-Holland, Inc.

[4] Giuseppe Attardi, Andrea Esuli, and Maria Simi. Best bets: thousands of queries in search of a client. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 422–423, New York, NY, USA, 2004. ACM Press.

[5] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley-Longman, 1st edition, 1999.

[6] Hemant K. Bhargava and Juan Feng. Preferential placement in Internet search engines. In *Proceedings of the eleventh international conference on World Wide Web*, pages 117–123. ACM Press, 2002.

[7] Julie Bichtler and Edward A. Eaton III. The combined use of bibliographic coupling and cocitation for document retrieval. *Journal of the American Society for Information Science*, 31(4):278–282, July 1980.

[8] David C. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), March 1985.

[9] Eric A. Brewer. The consumer side of search – introduction. *Communications of ACM*, 45(9):40–41, 2002.

[10] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, April 1998.

[11] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC-3. In *Proc. of the 3rd Text Retrieval Conference, TREC-3*, pages 69–80, Gaithersburg, MD, USA, 1995.

[12] Pável Calado, Altigran Soares da Silva, Rodrigo C. Vieira, Alberto H. F. Laender, and Berthier A. Ribeiro-Neto. Searching web databases by structuring keyword-based queries. In *Proceedings of the 11th International Conference on Information and Knowledge Management CIKM 2002*, pages 346–357, McLean, VA, USA, November 2002.

[13] James P. Callan. Document filtering with inference networks. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 262–269, Zurich, Switzerland, August 1996.

[14] John Joseph Carrasco, Daniel Fain, Kevin Lang, and Leonid Zhukov. Clustering of bipartite advertiser-keyword graph. In *Workshop on Clustering Large Datasets, 3th IEEE International Conference on Data Mining*, Melbourne, Florida, USA, November 2003. IEEE Computer Society Press. Available at `http://research.yahoo.com/publications.xml`.

[15] Soumen Chakrabarti, Byron Dom, and Piotr Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 307–318, Seattle, Washington, June 1998.

[16] Ernest P. Chan, Santiago Garcia, and Salim Roukos. TREC-5 ad hoc retrieval using K nearest-neighbors re-scoring. In *The Fifth Text REtrieval Conference (TREC-5)*. National Institute of Standards and Technology (NIST), November 1996.

[17] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[18] C. Chekuri, M. Goldwasser, Prabhakar Raghavan, and E. Upfal. Web search using automatic classification. In *Proceedings of WWW-97, 6th International Conference on the World Wide Web*, San Clara, CA, USA, 1997. Available at `http://theory.stanford.edu/people/wass/publications/Web_Search/Web_Search.html`.

[19] Ramnath K. Chellappa and Shivendu Shivendu. A model of advertiser–portal contracts: Personalization strategies under privacy concerns. *Inf. Tech. and Management*, 7(1):7–19, 2006.

[20] Hao Chen and Susan T. Dumais. Bringing order to the Web: Automatically categorizing search results. In *Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems*, pages 145–152, Hague, The Netherlands, April 2000.

[21] Tatiana Coelho, Pável Calado, Lamarque Souza, and Berthier Ribeiro-Neto. Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):408–417, April 2003.

[22] David Cohn and Thomas Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 430–436. MIT Press, 2001.

[23] James H. Coombs. Hypertext, full text, and automatic linking. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 83–98, Brussels, Belgium, September 1990.

[24] Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks (research note). *Artificial Intelligence*, 42(2-3):393–405, 1990.

[25] Thierson Couto, Marco Cristo, Marcos André Gonçalves, Pável Calado, Nivio Ziviani, and Berthier Ribeiro-Neto. A comparative study of citations and links in document classification. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, New York, NY, USA, June 2006. ACM Press.

[26] W. Bruce Croft, T. J. Lucia, J. Cringean, and P. Willett. Retrieving documents by plausible inference: an experimental study. *Information Processing & Management*, 25(6):599–614, 1989.

[27] W. Bruce Croft and Howard Turtle. A retrieval model for incorporating hypertext links. In *Proceedings of the 2nd Annual ACM Conference on Hypertext*, pages 213–224, Pittsburgh, PA, USA, November 1989.

[28] Luis M. de Campos, Juan M. Fernández-Luna, and Juan F. Huete. Query expansion in information retrieval systems using a bayesian network-based thesaurus. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI–98)*, pages 53–60, San Francisco, CA, July 1998.

[29] Maria de Lourdes da Silveira, Berthier Ribeiro-Neto, Rodrigo de Freitas Vale, and Rodrigo Torres Assumpção. Vertical searching in juridical digital libraries. In *Advances in Information Retrieval, 25th European Conference on IR Research, ECIR 2003, Proceedings*, pages 491–501, Pisa, Italy, April 2003.

[30] Jeffrey Dean and Monika Rauch Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11–16):1467–1479, May 1999. Also in Proceedings of the 8th International World Wide Web Conference.

[31] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[32] Susan T. Dumais, John Platt, David Hecherman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management CIKM 98*, pages 148–155, Bethesda, Maryland, USA, November 1998.

[33] Elena Eneva. 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, Twentieth International Conference on Machine Learning*, Washington DC, USA, August 2003. AAAI Press.

[34] Juan Feng, Hemant Bhargava, and David Pennock. Implementing paid placement in Web search engines: computational evaluation of alternative mechanisms. *INFORMS Journal on Computing*, 2006. To be published.

[35] Michelle Fisher and Richard Everson. When are links useful? Experiments in text classification. In F. Sebastianini, editor, *Proceedings of the 25th annual European conference on Information Retrieval Research, ECIR 2003*, pages 41–56. Springer-Verlag, Berlin, Heidelberg, DE, 2003.

[36] E. Fix and J. L. Hodges. Discriminatory analysis — nonparametric discrimination:consistency properties. Technical Report Technical Report 4, Proj. N. 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX, USA, 1951.

[37] Mark E. Frisse and Steve B. Cousins. Information retrieval from hypertext: Update on the dynamic medical handbook project. In *Proceedings of the 2nd Annual ACM Conference on Hypertext*, pages 199–212, Pittsburgh, PA, USA, November 1989.

[38] Johannes Furnkranz. Exploiting structural information for text classification on the WWW. In *Intelligent Data Analysis*, pages 487–498, 1999.

[39] Eugene Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.

[40] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring Web communities from link topology. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space - Structure in Hypermedia Systems*, pages 225–234, June 1998.

[41] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the 3rd ACM International Conference on Digital Libraries*, pages 89–98, Pittsburgh, Pennsylvania, USA, June 1998.

[42] David Gleich and Leonid Zhukov. SVD based term suggestion and ranking system. In *Proceedings of the 4th IEEE International Conference on Data Mining*, pages 391–394, Brighton, UK, November 2004. IEEE Computer Society.

[43] Eric Glover, Gary Flake, Steve Lawrence, William P. Birmingham, Andries Kruger, Lee L. Giles, and David Pennock. Improving category specific web search by learning query modifications. In *Symposium on Applications and the Internet, SAINT*, pages 23–31, San Diego, CA, January August–December 2001. IEEE Computer Society, Los Alamitos, CA.

[44] Eric J. Glover, Kostas Tsioutsiouliklis, Steve Lawrence, David M. Pennock, and Gary W. Flake. Using Web structure for classifying and describing Web pages. In *Proceedings of WWW-02, International Conference on the World Wide Web*, 2002.

[45] Joshua Goodman and Vitor R. Carvalho. Implicit queries for email. In *CEAS 2005 - Second Conference on Email and Anti-Spam, July 21-22, 2005, Stanford University*, 2005.

[46] Karam Gouda and Mohammed Javeed Zaki. Efficiently mining maximal frequent itemsets. In *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 163–170, Washington, DC, USA, 2001. IEEE Computer Society.

[47] Norbert Gövert, Mounia Lalmas, and Norbert Fuhr. A probabilistic description-oriented approach for categorizing web documents. In *Proceedings of the 8th International Conference on Information and Knowledge Management CIKM 99*, pages 475–482, Kansas City, Missouri, USA, November 1999.

[48] David Green. Search Engine Marketing: why it benefits us all. *Business Information Review*, 20(4):195–202, December 2003.

[49] Robyn Greenspan. How to cheat Google AdWords select. SitePoint, December 2003. Available at `http://www.sitepoint.com/article/adwords-select-parts-1-4`.

[50] David Haines and W. Bruce Croft. Relevance feedback and inference networks. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2–11, Pittsburgh, PA, USA, June 1993.

[51] D. K. Harman. Overview of the fourth text retrieval conference TREC-4. In D. K. Harman, editor, *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 1–24, Gaithersburg, Maryland, USA, November 1996. NIST Special Publication 500-236.

[52] Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM Press. Available at `http://www2002.org/CDROM/refereed/127/`.

[53] David Hawking, Nick Craswell, and Paul B. Thistlewaite. Overview of TREC-7 very large collection track. In *The Seventh Text REtrieval Conference (TREC-7)*, pages 91–104, Gaithersburg, Maryland, USA, November 1998.

[54] Xiaofeng He, Hongyuan Zha, Chris H. Q. Ding, and Horst D. Simon. Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis*, 41(1):19–45, November 2002.

[55] Gord Hotchkiss. Contextual text ads: the next big thing? SEO Today, August 2003. Available at `http://www.searchengineposition.com/info/netprofit/contextads.asp`.

[56] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, March 2002.

[57] IAB and PricewaterhouseCoopers. IAB internet advertising revenue report, April 2006. Available at `http://www.iab.net/2004adrevenues`.

[58] Lucas Introna and Helen Nissenbaum. Defining the web: The politics of search engines. *Computer*, 33(1):54–62, 2000.

[59] Greg Jarboe. Report: Most search marketers are unsophisticated. Search Engine Watch, February 2005. Available at `http://searchenginewatch.com/searchday/article.php/3469301`.

[60] Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, Germany, April 1998.

[61] Thorsten Joachims, Nello Cristianini, and John Shawe-Taylor. Composite kernels for hypertext categorisation. In Carla Brodley and Andrea Danyluk, editors, *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 250–257, Williams College, US, 2001. Morgan Kaufmann Publishers, San Francisco, US.

[62] Rob Kaiser. Web words become a lucrative market. Chicago Tribune, February 2005. Available at `www.chicagotribune.com/business/chi-0501240002jan24,1,2431400.story?coll=chi-business-hed&ctrack=1&cset=true`.

[63] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, January 1963.

[64] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[65] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143, Montreal, Canada, August 1995.

[66] Youji Kohda and Susumu Endo. Ubiquitous advertising on the WWW: merging advertisement on the browser. *Computer Networks and ISDN Systems*, 28(7-11):1493–1499, 1996.

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

[68] S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. The Web as a graph. In *Proceedings of the 19th Symposium on Principles of Database Systems*, pages 1–10, Dallas, Texas, USA, May 2000.

[69] Anísio Lacerda, Marco Cristo, Marcos André Gonçalves, Weiguo Fan, Nivio Ziviani, and Berthier Ribeiro-Neto. Learning to advertise. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, August 2006. ACM Press.

[70] Marc Langheinrich, Atsuyoshi Nakamura, Naoki Abe, Tomonari Kamba, and Yoshiyuki Koseki. Unintrusive customization techniques for Web Advertising. *Computer Networks*, 31(11-16):1259–1272, 1999.

[71] Kevin Lee. Context is king, or is it? ClickZ Experts, March 2003. Available at `http://www.clickz.com/experts/search/strat/article.php/2077801`.

[72] Kevin Lee. Google regional targeting power. ClickZ Experts, October 2003. Available at http://www.clickz.com/experts/search/strat/article.php/3101731.

[73] Kevin Lee. Making sense of AdSense... et al. ClickZ Experts, September 2003. Available at http://www.clickz.com/experts/search/strat/article.php/3076061.

[74] Kevin Lee. The SEM content conundrum. ClickZ Experts, July 2003. Available at http://www.clickz.com/experts/search/strat/article.php/2233821.

[75] Kevin Lee. AdSense makes much more sense. ClickZ Experts, May 2004. Available at http://www.clickz.com/experts/search/strat/article.php/3350001.

[76] Kevin Lee. Separating search and contextual inventory. ClickZ Experts, January 2004. Available at http://www.clickz.com/experts/search/strat/article.php/3305651.

[77] Kevin Lee. MSN's adCenter: more control and better results. ClickZ Experts, May 2005. Available at http://www.clickz.com/experts/search/strat/article.php/3490876.

[78] David D. Lewis. Evaluating text categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 312–318, February 1991.

[79] Hairong Li and John Leckenby. Internet Advertising formats and effectiveness. Center for Interactive Advertising, 2004. Available at http://www.ciadvertising.org/studies/reports/measurement/ad_format_print.pdf.

[80] Kate Maddox. Forrester reports advertising shift to online, May 2005. Available at http://www.btobonline.com/article.cms?articleId=24191.

[81] Fredrick Marckini. Contextual Advertising. ClickZ Experts, October 2003. Available at http://www.clickz.com/experts/search/results/article.php/3087311.

[82] Andrew McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. Available at http://www.cs.cmu.edu/~mccallum/bow, 1996.

[83] Andrew McCallum and Kamal Nigam. A comparison of event models for Naive Bayes text classification. In *Proceedings of AAAI/ICML-98, Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press, 1998.

[84] Sally McMillan. Internet Advertising: one face or many? In editors David W. Schumann & Esther Thorson, editor, *Internet Advertising: Theory and Research*. Lawrence Erlbaum, 2005.

[85] Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized on-line matching. In *FOCS '05: Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 264–273, Washington, DC, USA, 2005. IEEE Computer Society.

[86] Tom Mitchell. *Machine Learning*. McGraw-Hill, March 1997.

[87] Kevin Newcomb. FindWhat unveils answer to AdSense. ClickZ Experts, September 2004. Available at `http://www.clickz.com/news/article.php/3415431`.

[88] Kevin Newcomb. Google sues AdSense publisher for click fraud. ClickZ Experts, November 2004. Available at `http://www.clickz.com/news/article.php/3440341`.

[89] Kevin Newcomb. Kanoodle unveils self-service ad tool. ClickZ Experts, October 2004. Available at `http://www.clickz.com/news/article.php/3417811`.

[90] Scott Nicholson, Tito Sierra, U. Yeliz Eseryel, Ji-Hong Park, Philip Barkow, Erika J. Pozo, and Jane Ward. How much of it is real? analysis of paid placement in web search engine results. *J. Am. Soc. Inf. Sci. Technol.*, 57(4):448–461, 2006.

[91] Thomas P. Novak and Donna L. Hoffman. New metrics for new media: toward the development of Web measurement standards. *World Wide Web Journal*, 2(1):213–246, 1997.

[92] Hyo-Jung Oh, Sung Hyon Myaeng, and Mann-Ho Lee. A practical hypertext catergorization method using links and incrementally available class information. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 264–271. ACM Press, 2000.

[93] OneUpWeb. How keyword length affects conversion rates, January 2005. Available at `http://www.oneupweb.com/landing/keywordstudy_landing.htm`.

[94] Satoshi Oyama, Takashi Kokubo, Toru Ishida, Teruhiro Yamada, and Yasuhiko Kitamura. Keyword spices: A new method for building domain-specific web search engines. In Bernhard Nebel, editor, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pages 1457–1466. Morgan Kaufmann, 2001.

[95] Said Mirza Pahlevi and Hiroyuki Kitagawa. Conveying taxonomy context for topic-focused web search: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 56(2):173–188, 2005.

[96] Pamela Parker. The next context. ClickZ Experts, April 2004. Available at `http://www.clickz.com/experts/brand/buzz/article.php/3341161`.

[97] Jeffrey Parsons, Katherine Gallagher, and K. Dale Foster. Messages in the medium: An experimental investigation of Web Advertising effectiveness and attitudes toward Web content. In Jr. Ralph H. Sprague, editor, *Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 6*, page 6050, Washington, DC, USA, 2000. IEEE Computer Society.

[98] J. Pearl and T. Verma. The logic of representing dependencies by directed graphs. In *Proc. of AAAI-87*, pages 374–379, Seattle, WA, 1987.

[99] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann Publishers, 2nd edition, 1988.

[100] Michael Persin. Document filtering for fast ranking. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–348, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[101] Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 160–169, New York, NY, USA, 1993. ACM Press.

[102] Devanand Ravindran and Susan Gauch. Exploiting hierarchical relationships in conceptual search. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 238–239, New York, NY, USA, 2004. ACM Press.

[103] Andy Reinhardt. And you thought that Web ad market was dead. Business Week, May 2003. Available at `http://www.businessweek.com/magazine/content/03_18/b3831134_mz034.htm`.

[104] Berthier Ribeiro-Neto and Richard Muntz. A belief network model for IR. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, Zurich, Switzerland, August 1996.

[105] Knut Magne Risvik, Tomasz Mikolajewski, and Peter Boros. Query segmentation for web search. In *WWW (Posters)*, Budapest, Hungary, May 2003.

[106] JJ Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System Prentice-Hall*, pages 313–323, Englewood NJ, USA, 1971.

[107] Gerard Salton. Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, 10(4):440–457, October 1963.

[108] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal*, 24(5):513–523, 1988.

[109] Gerard Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1st edition, 1983.

[110] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002.

[111] Catherine Seda. Contextual ads: vital to a search marketing campaign? Search Engine Watch, October 2004. Available at `http://searchenginewatch.com/searchday/article.php/3418151`.

[112] Mike Shields. Multi-term searches get high conversion. Media Week, February 2005. Available at `http://www.mediaweek.com/mw/search/article_display.jsp?schema=&vnu_content_id=1000816071`.

[113] Mike Shields. 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`.

[114] Dongwook Shin, Sejin Nam, and Munseok Kim. Hypertext construction using statistical and semantic similarity. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 57–63, Philadelphia, PA, USA, July 1997.

[115] Altigran Silva, Eveline Veloso, Paulo Golgher, Berthier Ribeiro-Neto, Alberto Laender, and Nivio Ziviani. CobWeb - a crawler for the brazilian web. In *Proceedings of the String Processing and Information Retrieval Symposium (SPIRE'99)*, pages 184–191, Cancun, Mexico, September 1999.

[116] Michael Singer. Contextual ad debate rouses critics. ClickZ Experts, August 2003. Available at `http://www.clickz.com/news/article.php/3066971`.

[117] Seán Slattery and Tom Mitchell. Discovering test set regularities in relational domains. In *Proceedings of the 17th International Conference on Machine Learning*, pages 895–902, Stanford, California, USA, June 2000.

[118] Henry G. Small. Co-citation in the scientific literature: A new measure of relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, July 1973.

[119] Henry G. Small and Michael E. D. Koenig. Journal clustering using a bibliographic coupling method. *Information Processing and Management*, 13(5):277–288, 1977.

[120] Alastair G. Smith. Web links as analogues of citations. *Information Research*, 9(4), July 2004.

[121] Richard Stern. Is gatoring unfair or illegal? *IEEE Micro*, 22(1):6–7, 92–93, 2002.

[122] M. Stone. Cross-validation choices and assessment of statistical predictions. *Journal of the Royal Statistical Society*, B36:111–147, 1974.

[123] Danny Sullivan. The content context contest. ClickZ Experts, June 2003. Available at `http://www.clickz.com/experts/search/opt/article.php/2241231`.

[124] Danny Sullivan. Contextual ads and the little guy. ClickZ Experts, June 2003. Available at `http://www.clickz.com/resources/search_reference/contextual_advertising/article.php/2224611`.

[125] Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. Web classification using support vector machine. In *Proceedings of the fourth international workshop on Web information and data management*, pages 96–99. ACM Press, 2002.

[126] Wen tau Yih, Joshua Goodman, and Vitor R. Carvalho. Finding advertising keywords on web pages. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 213–222, New York, NY, USA, 2006. ACM Press.

[127] Loren Terveen, Will Hill, and Brian Amento. Constructing, organizing, and visualizing collections of topically related Web resources. *ACM Transactions on Computer-Human Interaction*, 6(1):67–94, March 1999.

[128] Mike Thelwall and David Wilkinson. Finding similar academic Web sites with links, bibliometric couplings and colinks. *Information Processing & Management*, 2003. (in press).

[129] Denis Turker. The optimal design of a seach engine from an Agency Theory perspective. Working paper, Institute for Broadcasting Economics – University of Koeln, August 2004. Available at `http://www.rundfunkoekonomie.uni-koeln.de/institut/publikationen/arbeitspapiere/ap191.php`.

[130] Howard Turtle and W. Bruce Croft. Inference networks for document retrieval. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–24, Brussels, Belgium, September 1990.

[131] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991.

[132] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Velag, New York, March 1995.

[133] Chingning Wang, Ping Zhang, Risook Choi, and Michael Daeredita. Understanding consumers attitude toward Advertising. In *Eighth*

Americas Conference on Information Systems, pages 1143–1148, August 2002.

[134] Tessa Wegert. Contextual ads: a consumer's point of view. AtNetwork, April 2003. Available at `http://www.atnewyork.com/news/article.php/2196831`.

[135] Melius Weideman. Ethical issues on content distribution to digital consumers via paid placement as opposed to website visibility in search engine results. In *The Seventh ETHICOMP International Conference on the Social and Ethical Impacts of Information and Communication Technologies*, pages 904–915. Troubador Publishing Ltd, April 2004.

[136] Melius Weideman and Timothy Haig-Smith. An investigation into search engines as a form of targeted advert delivery. In *Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*, pages 258–258. South African Institute for Computer Scientists and Information Technologists, 2002.

[137] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, 2nd edition, 1999.

[138] S. K. M. Wong and Y. Y. Yao. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1):38–68, January 1995.

[139] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, August 2004.

[140] Yiming Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In W. B. Croft and editors C. J. van Rijsbergen, editors, *Proceedings of the 17rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 13–22. Springer-Verlag, 1994.

[141] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49, Berkeley, California, USA, August 1999.

[142] Yiming Yang, Seán Slattery, and Rayid Ghani. A study of approaches to hypertext categorization. *Journal of Intelligent Information Systems*, 18(2-3):219–241, 2002.

[143] Baoping Zhang, Yuxin Chen, Weiguo Fan, Edward A. Fox, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Intelligent fusion of structural and citation-based evidence for text classification. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 667–668, New York, NY, USA, 2005. ACM Press.

[144] Baoping Zhang, Yuxin Chen, Weiguo Fan, Edward A. Fox, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Intelligent GP fusion from multiple sources for text classification. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 477–484, New York, NY, USA, 2005. ACM Press.

Appendix A

Mapping of Taxonomies

Table A.1 shows the mapping of the Cade188 taxonomy to the ad taxonomy. As we can see, in general, the Cade188 categories were mapped according to their first-level entries. For instance, each subcategory of *Culture* in the Cade188 taxonomy was mapped to the ad category *Shopping & Services/Culture*. The same is observed to the subcategories of *Education*, *Science*, *Sports*, *Computers*, *Internet*, *Leisure and Entertainment*, and *Health*. Hence ads belonging to the ad category *Shopping & Services/Culture/Music*, for example, can be placed in a triggering page classified under the page category *Culture/Dance*.

On the other hand, given the heterogeneity of the classes *On-line Shopping*, *News*, *Services*, and *Society*, their subcategories were mapped directly to the corresponding ad categories. For instance, the page category *Services/Automobiles* was mapped to *Shopping & Services/Vehicles/Cars*.

Also note that some mappings were made according to the affinity of the classes. For instance, the page category *Leisure and Entertainment/Movies & Video* was mapped to the ad category *Shopping & Services/Electrical and Electronic Appliances/Audio & Video*. Consequently, ads about home theaters and DVD players can be placed in pages about movies.

Table A.1: Mapping of ad and triggering page taxonomies

Cade188 Category	Ad Category
Culture/Plastic Arts	Shopping & Services/Culture Shopping & Services/Culture/Plastic Arts
Culture/Associations	Shopping & Services/Culture
Culture/Dance	Shopping & Services/Culture Shopping & Services/Culture/Dance
Culture/Philosophy	Shopping & Services/Culture
Culture/Folklore	Shopping & Services/Culture
Culture/History	Shopping & Services/Culture
Culture/Literature	Shopping & Services/Culture Shopping & Services/Culture/Books
Culture/Museums	Shopping & Services/Culture
Culture/Music	Shopping & Services/Culture Shopping & Services/Culture/Music Shopping & Services/Education/Training & Development/Music Shopping & Services/Computers/Software/Music
Culture/Countries	Shopping & Services/Culture
Culture/Other	Shopping & Services/Culture
Computers/Computer Graphics	Shopping & Services/Computers
Computers/Education	Shopping & Services/Computers Shopping & Services/Education
Computers/Companies	Shopping & Services/Computers
Computers/Forums & User Groups	Shopping & Services/Computers
Computers/Hackers & Viruses	Shopping & Services/Computers
Computers/Hardware & Supplies	Shopping & Services/Computers Shopping & Services/Documents/Printing/Supplies
Computers/Programming	Shopping & Services/Computers
Computers/Networks & Connectivity	Shopping & Services/Computers

continued...

Cade188 Category	Ad Category
Computers/Services	Shopping & Services/Computers
Computers/Software	Shopping & Services/Computers Shopping & Services/Culture/Music/Software Shopping & Services/Real Estate/Engineering/Software Shopping & Services/Real Estate/Services & Materials/Software Shopping & Services/Real Estate/Services & Materials/Software/Financial Shopping & Services/Health & Care/Dentistry/Software Shopping & Services/Kids/Software
Computers/Educational Software	Shopping & Services/Computers Shopping & Services/Education
Computers/Medical Software	Shopping & Services/Computers Shopping & Services/Health & Care
Computers/Tutorials & Information	Shopping & Services/Computers
Computers/Shopping	Shopping & Services/Computers
Education/Courses	Shopping & Services/Food & Beverages/Courses Shopping & Services/Food & Beverages/Courses/Japanese Cooking Shopping & Services/Culture/Music/Course Shopping & Services/Law/Course Shopping & Services/Economy/Courses Shopping & Services/Computers/Courses Shopping & Services/Computers/Hardware/Course Shopping & Services/Computers/Internet/Course Shopping & Services/Health & Care/Courses Shopping & Services/Economy/Education Shopping & Services/Education Shopping & Services/Computers/Software/Kids/Education
Education/Distance Learning	Shopping & Services/Education
Education/Organizations & Associations	Shopping & Services/Education
Education/Schools	Shopping & Services/Education
Education/Information	Shopping & Services/Education

continued...

Cade188 Category	Ad Category
Education/Educational Exchange	Shopping & Services/Education Shopping & Services/Education/Spanish Shopping & Services/Education/Foreign Studies Shopping & Services/Education/Languages/Foreign Studies/ Shopping & Services/Education/English Shopping & Services/Education/Accommodation/Campinas
Education/Languages	Shopping & Services/Education Shopping & Services/Education/Spanish Shopping & Services/Education/Foreign Studies Shopping & Services/Education/Languages/Foreign Studies/ Shopping & Services/Education/English
Education/Professionals	Shopping & Services/Education
Education/Term & Research Papers, Reports & Essays	Shopping & Services/Education
Education/Colleges & Universities	Shopping & Services/Education Shopping & Services/Education/Higher Education
Education/Vestibular Contest	Shopping & Services/Education Shopping & Services/Education/Vestibular Contest
Health/Diagnostic Centers	Shopping & Services/Health & Care
Health/Clinics	Shopping & Services/Health & Care
Health/Diseases & Related Information	Shopping & Services/Health & Care
Health/Organizations	Shopping & Services/Health & Care
Health/Medical Specialties	Shopping & Services/Health & Care
Health/Drugstores	Shopping & Services/Health & Care
Health/Physiotherapy	Shopping & Services/Health & Care
Health/Homeopathy	Shopping & Services/Health & Care
Health/Hospitals	Shopping & Services/Health & Care
Health/Industry	Shopping & Services/Health & Care
Health/Medical Information	Shopping & Services/Health & Care

continued...

Cade188 Category	Ad Category
Health/Institutions	Shopping & Services/Health & Care
Health/Yoga	Shopping & Services/Health & Care
Health/Laboratories	Shopping & Services/Health & Care Shopping & Services/Laboratories
Health/Alternative Medicine	Shopping & Services/Health & Care
Health/Nutrition	Shopping & Services/Health & Care
Health/Dentistry	Shopping & Services/Health & Care Shopping & Services/Computers/Software/Dentistry
Health/Professionals	Shopping & Services/Health & Care
Health/Psychology & Psychoanalysis	Shopping & Services/Health & Care
Health/Senior Health	Shopping & Services/Health & Care
Health/Occupational Safety & Health	Shopping & Services/Health & Care
Health/Health Insurance & Health Care Plans	Shopping & Services/Health & Care Shopping & Services/Insurance
Health/Services	Shopping & Services/Health & Care
Health/Spas	Shopping & Services/Health & Care Shopping & Services/Tourism/Resorts/Spas
Health/Veterinary	Shopping & Services/Health & Care Shopping & Services/Pets & Animals
Internet/Chat	Shopping & Services/Computers/Internet
Internet/Cams	Shopping & Services/Computers/Internet
Internet/Web Development	Shopping & Services/Computers/Internet
Internet/Downloads	Shopping & Services/Computers/Internet
Internet/Web Indexes	Shopping & Services/Computers/Internet
Internet/Information	Shopping & Services/Computers/Internet
Internet/Access Providers	Shopping & Services/Computers/Internet
Internet/Domain Names & Web Hosting	Shopping & Services/Computers/Internet
Internet/Services	Shopping & Services/Computers/Internet
Internet/Tutorials	Shopping & Services/Computers/Internet

continued...

Cade188 Category	Ad Category
Leisure & Entertainment/Camping	Shopping & Services/Entertainment
Leisure & Entertainment/Pets & Animals	Shopping & Services/Entertainment Shopping & Services/Pets & Animals
Leisure & Entertainment/Bars & Restaurants	Shopping & Services/Entertainment Shopping & Services/Food & Beverages Shopping & Services/Computers/Software/Commercial Automation/ Restaurants
Leisure & Entertainment/Nightclubs	Shopping & Services/Entertainment
Leisure & Entertainment/Brazilian Carnival	Shopping & Services/Entertainment
Leisure & Entertainment/Movies & Video	Shopping & Services/Entertainment Shopping & Services/Culture/Movies Shopping & Services/Electrical & Electronic Appliances/Audio & Video Shopping & Services/Electronic Media/Video
Leisure & Entertainment/Clubs & Associations	Shopping & Services/Entertainment
Leisure & Entertainment/Cooking	Shopping & Services/Entertainment Shopping & Services/Food & Beverages Shopping & Services/Culture/Gastronomy
Leisure & Entertainment/Hobbies	Shopping & Services/Entertainment Shopping & Services/Mock-ups
Leisure & Entertainment/Humor	Shopping & Services/Entertainment
Leisure & Entertainment/Kids	Shopping & Services/Entertainment Shopping & Services/Kids Shopping & Services/Computers/Software/Kids
Leisure & Entertainment/Games	Shopping & Services/Entertainment
Leisure & Entertainment/Lotteries & Gambling Games	Shopping & Services/Entertainment Gambling Games
Leisure & Entertainment/Parks	Shopping & Services/Entertainment Shopping & Services/Kids/Playgrounds
Leisure & Entertainment/Personalities	Shopping & Services/Entertainment

continued...

Cade188 Category	Ad Category
Leisure & Entertainment/Meeting Points	Shopping & Services/Entertainment
Leisure & Entertainment/Cultural Events	Shopping & Services/Entertainment Shopping & Services/Culture
Leisure & Entertainment/Comics & Animation	Shopping & Services/Entertainment
Leisure & Entertainment/Radio	Shopping & Services/Entertainment
Leisure & Entertainment/Theater	Shopping & Services/Entertainment Shopping & Services/Culture/Theater Shopping & Services/Education/Training & Development/Theater
Leisure & Entertainment/Tourism	Shopping & Services/Entertainment Shopping & Services/Tourism
Leisure & Entertainment/Gangs	Shopping & Services/Entertainment
Leisure & Entertainment/TV	Shopping & Services/Entertainment
News/Classified	Shopping & Services/Classified
On-line Shopping/Food & Beverages	Shopping & Services/Food & Beverages Shopping & Services/Culture/Gastronomy Industry/Foods/Supplies & Equipments
On-line Shopping/Audio & Video	Shopping & Services/Electrical & Electronic Appliances Shopping & Services/Electronic Media/Video
On-line Shopping/Electronics	Shopping & Services/Electrical & Electronic Appliances Industry/Electronic Industry/Electronic/Voltage Regulators & UPS
On-line Shopping/Flowers	Shopping & Services/Festivals & Events/Gifts & Occasions/Baskets & Flowers
On-line Shopping/Books	Shopping & Services/Culture/Books
On-line Shopping/Fashion, Clothing & Accessories	Shopping & Services/Culture/Fashion Shopping & Services/Culture/Fashion/Clothing Shopping & Services/Culture/Fashion/Clothing/Accessories
On-line Shopping/Sex	Shopping & Services/Sex
On-line Shopping/Virtual Mall	Shopping & Services/Virtual Mall
Science/Chemistry	Shopping & Services/Chemistry

continued...

Cade188 Category	Ad Category
Science/Chemistry	Industry/Chemistry Industry/Chemistry/Equipments & Material
Services/Law	Shopping & Services/Law
Services/Storage	Shopping & Services/Storage
Services/Automobiles	Shopping & Services/Vehicles/Cars
Services/Beauty	Shopping & Services/Computers/Software/Health & Care Shopping & Services/Health & Care
Services/Notary's Office	Shopping & Services/Support Services/Notary's Office
Services/Communication	Shopping & Services/Telecommunication
Services/Consulting	Shopping & Services/Environment/Consulting Shopping & Services/Health & Care/Drug Companies/Consulting Shopping & Services/Corporative Services/Consulting Shopping & Services/Transport/Air/Consulting Shopping & Services/Transport/Consulting Shopping & Services/Retail/Consulting Shopping & Services/Food & Beverages/Consulting Shopping & Services/Economy/Consulting Shopping & Services/Economy/Finance & Investment/Consulting Shopping & Services/Engineering/Consulting Shopping & Services/Computers/Consulting Shopping & Services/Computers/Software/Consulting
Services/Accountancy	Shopping & Services/Economy/Finance & Investment/Accounting
Services/Interior Design & Decoration	Shopping & Services/Decoration & Furniture Shopping & Services/Domestic Utensils
Services/Dispatcher	Shopping & Services/Vehicles/Cars/Services/Dispatcher
Services/Detectives & Investigations	Shopping & Services/Security/Investigation Services
Services/Engineering	Shopping & Services/Engineering
Services/Offices	Shopping & Services/Real Estate/Architecture/Offices Shopping & Services/Real Estate/Offices Shopping & Services/Decoration & Furniture/Offices

continued...

Cade188 Category	Ad Category
Services/Events	Shopping & Services/Culture/Shows & Events Shopping & Services/Festivals & Events
Services/Document Management	Shopping & Services/Computers/Software/Documents Shopping & Services/Documents/Management
Services/Real Estate	Shopping & Services/Real Estate
Services/Condominium & Gated community	Shopping & Services/Real Estate
Services/Laundry	Shopping & Services/Maintenance/Laundry & Cleaning
Services/Auctions	Shopping & Services/Culture/Plastic Arts/Auctions Shopping & Services/Auctions
Services/Washing & Cleaning	Shopping & Services/Maintenance/Laundry & Cleaning Shopping & Services/Maintenance/Cleaning
Services/Landscape Architecture & Garden Design	Shopping & Services/Decoration & Furniture/Home & Garden Shopping & Services/Decoration & Furniture/Home & Garden/Landscape Architecture
Services/Human Resources	Shopping & Services/Computers/Software/Human Resources Shopping & Services/Job Market/Human Resources Shopping & Services/Corporative Services/Human Resources
Services/Security	Shopping & Services/Computers/Security Shopping & Services/Security
Services/Translation	Shopping & Services/Documents/Translation
Services/Transports	Shopping & Services/Transport
Society/Aviation	Shopping & Services/Transport/Air
Society/Cars & Motorcycles	Shopping & Services/Vehicles/Cars
Society/Esotericism	Shopping & Services/Culture/Religion/Esotericism
Society/Events	Shopping & Services/Culture/Shows & Events Shopping & Services/Festivals & Events
Society/Environment	Shopping & Services/Environment Shopping & Services/Energy/Ecological Alternative
Society/Organizations	Society/Organizations
Society/People	Shopping & Services/Computers/Internet/Homepages

continued...

Cade188 Category	Ad Category
Society/Personal Relationships	Shopping & Services/Personal Relationships
Society/Religion	Shopping & Services/Culture/Religion
Society/Sexuality	Shopping & Services/Health & Care/Sex Shopping & Services/Sex
Sports/Health Clubs	Shopping & Services/Sports Shopping & Services/Health & Care/Gymnastics Shopping & Services/Fitness Facilities
Sports/Mountaineering	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Martial Arts	Shopping & Services/Sports Shopping & Services/Culture/Martial Arts Shopping & Services/Fitness Facilities
Sports/Auto Racing	Shopping & Services/Sports Shopping & Services/Vehicles
Sports/Cycling	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Organizations & Associations	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Extreme & Adventure Sports	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Soccer	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Equestrianism	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Business, Shopping & Services	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Scuba Diving	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Motorcycle Racing	Shopping & Services/Sports Shopping & Services/Fitness Facilities

continued...

Cade188 Category	Ad Category
Sports/Swimming	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Off-road	Shopping & Services/Sports Shopping & Services/Fitness Facilities Shopping & Services/Vehicles/Cars/4x4
Sports/Fishing	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Rodeo	Shopping & Services/Sports Shopping & Services/Fitness Facilities Shopping & Services/Farming/Rodeo
Sports/Surfing	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Tennis	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Sailing	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Air Sports	Shopping & Services/Sports Shopping & Services/Fitness Facilities
Sports/Chess	Shopping & Services/Sports Shopping & Services/Fitness Facilities

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo