5
Experimentos

Este capítulo apresenta o ambiente experimental utilizado para validar o método de predição do CTR proposto neste trabalho. Na seção 5.1, descrevemos a geração do conjunto de dados sintético; na seção 5.2, a metodologia e as métricas de avaliação utilizadas; e na seção 5.3, os resultados dos experimentos realizados.

5.1.
Conjunto de Dados

Foi realizada uma pesquisa por conjuntos de dados contendo CTR's reais, que pudessem ser utilizados nos experimentos com o algoritmo proposto, mas constatou-se que esse tipo de informação não é disponibilizado publicamente. Dessa forma, a solução encontrada para conduzir os experimentos foi a geração de dados sintéticos.

Para gerar esses dados sintéticos, obtivemos o recurso WPT 03[7] (Cardoso et al., 2007), que contém uma coleta da Web Portuguesa de 2003, juntamente com os registros de seis meses de consultas submetidas à máquina de busca TUMBA![8]. Em seguida, submetemos cada uma das palavras-chave dessas consultas à máquina de busca Google, obtendo os anúncios do AdWords a elas associados, e a ordem em que eles são exibidos. Assim, construímos uma matriz P na qual cada linha i representa uma palavra-chave, cada coluna j representa um anúncio, e cada elemento p_{ij} representa a posição do anúncio i para a palavra-chave j, estando preenchidos apenas os elementos cujo anúncio é retornado pelo Google como resultado da submissão da consulta.

Como não temos os CTR's reais desses anúncios, substituímos cada elemento p_{ij} conhecido por uma estimativa desses CTR's. Para tanto, utilizamos

7 O WPT 03 é um recurso criado pelo Grupo XLDB (http://xldb.di.fc.ul.pt – último acesso em março de 2008), e disponibilizado pela Linguateca-XLDB (http://www.linguateca.pt – último acesso em março de 2008).

os valores apresentados na Tabela 4 (Brooks, 2004), que indica o CTR esperado dos anúncios exibidos nas 10 primeiras posições, em relação ao CTR do anúncio na primeira posição, no Google AdWords. Após essa conversão, passamos a ter uma matriz *C*, cujos elementos c_{ij} são os CTR's relativos correspondentes aos elementos p_{ij} conhecidos.

Tabela 4: CTR esperado, em relação ao CTR da primeira posição, no Google AdWords.

Posição	**CTR relativo**
1	100%
2	77,4%
3	66,6%
4	57,4%
5	52,9%
6	50,2%
7	39,7%
8	34,3%
9	26,0%
10	26,3%

Essa matriz *C* possui 55.747 linhas, representando as palavras-chave, 50.608 colunas, representando os anúncios, e 182.090 elementos conhecidos, correspondentes aos CTR's relativos, aos quais este texto se refere apenas por "CTR" daqui por diante. Note que *C* é extremamente esparsa, sendo conhecidos apenas 0,006% de seus elementos, o que resulta nas médias de apenas 3,27 CTR's por palavra-chave e 3,6 CTR's por anúncio.

A partir dos dados contidos no *query log* do WPT 03, é possível calcular o total de consultas submetidas ao TUMBA! durante os 6 meses registrados, considerando as repetições de uma mesma palavra-chave como consultas distintas: 139.219. Combinando tais registros com os elementos conhecidos da

8 Disponível em: http://www.tumba.pt (último acesso em março de 2008).

matriz *C*, também podemos inferir o total de impressões de anúncios associadas a essas consultas, caso as mesmas fossem submetidas ao Google: 463.174.

Observe que a quantidade de palavras-chave e anúncios somados, 106.355, equivale à quantidade mínima de pesos aprendidos pelo algoritmo descrito na seção 3.2, quando o parâmetro de treinamento $K = 1$. Dessa forma, a razão entre a quantidade de elementos c_{ij} conhecidos e o número mínimo de pesos a serem aprendidos é apenas 1,71, o que dificulta muito a geração de um modelo com um bom nível de generalização. Por esse motivo, extraímos da matriz original uma submatriz mais densa, conforme descrito a seguir.

A fim de obter uma submatriz densa de *P*, primeiramente eliminamos todos os elementos p_{ij} cuja linha i ou coluna j contivesse menos de três elementos conhecidos. Após isso, eliminamos todas as linhas e colunas da matriz que ficaram vazias em decorrência da eliminação de seus elementos, reduzindo assim as dimensões da matriz. Em seguida, corrigimos os valores dos elementos da matriz, de forma a refletir as novas posições dos anúncios. Por exemplo, se um elemento $p_{ij} = 2$ é eliminado, então todo $p_{ik} \geq 2$ será atualizado para $p'_{ik} = p_{ik} - 1$. Dessa forma, obtemos a matriz *P'*, à qual aplicamos a mesma conversão que transformou *P* em *C*.

O processo descrito acima resultou na matriz *C'*, que possui 23.248 linhas, 13.355 colunas e 108.766 elementos c'_{ij} conhecidos, correspondentes a 0,04% do total. Além disso, a razão entre a quantidade de elementos c'_{ij} conhecidos e o número mínimo de pesos a serem aprendidos aumentou para 2,97. Vale observar que o número de consultas submetidas representadas em *C'* é 58.074, ou 42% do total, e que as impressões de anúncios correspondentes são 283.701, ou 61% do total, o que indica que os dados contidos em *C'* são uma parcela representativa dos dados contidos em *C*.

Nas seções a seguir, nos referimos às matrizes *C* e *C'* por "conjunto completo" e "conjunto reduzido", respectivamente.

5.2. Metodologia e Métricas de Avaliação

A fim de prover um **sistema de referência** para comparação com o algoritmo proposto na seção 3.2, utilizamos o algoritmo de **Média das Médias** (**MM**), que fornece como predição para um CTR desconhecido c'_{ij} a média das médias de todos os valores c_{ik} conhecidos contidos na mesma linha e de todos os valores c_{kj} conhecidos contidos na mesma coluna. Dessa forma, adotamos para os dois algoritmos a mesma metodologia e métricas de avaliação, descritas a seguir.

A avaliação foi realizada com uma validação cruzada de 20 iterações. Para cada iteração, o conjunto de dados utilizado é dividido em 95% dos CTR's para treinamento e 5% para teste. Ainda que tal divisão resulte em subconjuntos distintos a cada iteração, ela não pode ser realizada de forma totalmente aleatória, pois a fim de garantir que haja dados suficientes no conjunto de treino para o aprendizado dos atributos latentes, cada um dos CTR's que compõem o conjunto de teste obedece aos seguintes critérios:

i. Deve existir pelo menos outro CTR na mesma linha e outro CTR na mesma coluna que façam parte do conjunto de treino.
ii. Não deve existir outro CTR na mesma linha e nem outro CTR na mesma coluna que façam parte do conjunto de teste.

A fim de comparar o desempenho do algoritmo com diferentes parâmetros de treinamento, e nos diferentes conjuntos de dados, foram utilizadas três métricas de avaliação. A primeira delas é o ***Rooted Mean Squared Error*** (**RMSE**, do inglês, raiz do erro quadrático médio) da predição do CTR, que é equivalente à função objetivo do algoritmo de aprendizado utilizado, o qual busca minimizar o erro quadrático. Tal métrica indica o quanto os valores preditos, em média, estão distantes dos valores reais, conferindo maior peso a erros grandes.

Conforme descrito na seção 2.2, a ordenação de anúncios feita pelas máquinas de busca leva em consideração não apenas o CTR, mas também os lances dos anunciantes pelas palavras-chave. Ainda assim, é conveniente avaliar, além do erro de predição do algoritmo proposto, a qualidade de uma ordenação construída a partir dos CTR's preditos, já que os próprios CTR's utilizados nos

experimentos não são reais, mas foram sintetizados a partir de ordenações de anúncios.

Dessa forma, para cada palavra-chave, os anúncios foram ordenados decrescentemente por CTR predito, e duas métricas foram utilizadas para avaliar o quanto essa nova ordenação se aproxima da ordenação real. A **Precisão do Posicionamento** (**PP**) indica a fração dos anúncios que foram recolocados, na ordenação baseada em CTR's preditos, exatamente na mesma posição que ocupavam originalmente na ordenação real. Já o **Erro Absoluto Médio da Posição** (**EAMP**) indica a média dos valores absolutos das diferenças entre a posição em que os anúncios foram recolocados na ordenação baseada em CTR's preditos e a posição que eles ocupavam originalmente na ordenação real.

5.3. Resultados

A Tabela 5 mostra os resultados dos experimentos do sistema de referência, permitindo uma comparação de seu desempenho com os conjuntos de dados completo e reduzido. Note que, em todas as métricas de avaliação, o sistema de referência obteve um melhor resultado no conjunto de dados reduzido. Tal fato indica que uma matriz de dados mais densa facilita o processo de predição de valores desconhecidos, mesmo quando utilizamos um algoritmo bastante simples.

Tabela 5: Resultados dos experimentos do sistema de referência com os conjuntos de dados completo e reduzido.

Conjunto de dados	**RMSE**	**PP**	**EAMP**
Completo	0,1583	40,53%	1,0949
Reduzido	0,1489	45,46%	0,914

Nos experimentos utilizando o algoritmo de fatoração de matrizes (FM), os quais são descritos a seguir, a taxa de aprendizado η foi fixada em 0,1. Apesar de esse valor ser alto quando comparado com outros trabalhos que utilizam aprendizado por descida de gradiente, ele foi escolhido para compensar a pequena quantidade de exemplos de treinamento disponíveis. Além disso, o fator de regularização λ foi fixado em 0,01, valor que garante um bom nível de generalização sem gerar ruído no aprendizado, conforme verificamos empiricamente.

A Tabela 6 exibe os resultados de experimentos com o algoritmo FM, utilizando apenas o conjunto de dados completo. Nesses experimentos, o algoritmo foi avaliado diversas vezes, variando-se a quantidade de atributos latentes a serem aprendidos, controlada pelo parâmetro *K*, entre 1, 2 e 3, bem como o número de épocas a serem utilizadas no treinamento, controlado pelo parâmetro *T*, entre 1, 2, 5, 10, 20, 50, 100, 200, 500 e 1.000. Os valores em negrito indicam o melhor resultado para cada uma das três métricas de avaliação utilizadas.

Tabela 6: Resultados dos experimentos do algoritmo FM com o conjunto de dados completo, variando os parâmetros de treinamento *K* e *T*.

Épocas de treinamento (T)	**Atributos latentes (K)**								
	1			**2**			**3**		
	RMSE	**PP**	**EAMP**	**RMSE**	**PP**	**EAMP**	**RMSE**	**PP**	**EAMP**
1	0,2115	32,95%	1,4934	0,6166	25,33%	1,7836	0,9638	24,41%	1,839
2	0,1785	37,91%	1,2817	0,4768	27,81%	1,66	0,7376	26,21%	1,7368
5	0,1542	42,73%	1,0678	0,34	33,42%	1,3864	0,5185	31,00%	1,4905
10	**0,1481**	45,57%	0,9586	0,275	38,25%	1,1771	0,4159	36,26%	1,2409
20	0,1526	47,75%	0,8939	0,2396	43,18%	1,0152	0,3548	41,32%	1,0756
50	0,1686	48,94%	0,8598	0,2236	46,12%	0,9315	0,3118	44,98%	0,9764
100	0,1783	49,13%	0,8604	0,2177	46,99%	0,9016	0,2802	46,28%	0,9301
200	0,1834	49,25%	0,857	0,2083	47,67%	0,8883	0,248	46,67%	0,9094
500	0,1858	49,41%	**0,8509**	0,1976	48,26%	0,8856	0,2089	47,28%	0,8999
1.000	0,1867	**49,55%**	0,8534	0,1927	48,11%	0,8914	0,1962	47,54%	0,9029

Observe que, em todas as métricas de avaliação, os resultados obtidos pioram conforme aumentamos *K*. Esse fato se deve à pequena quantidade de exemplos em relação à quantidade de pesos a serem aprendidos, que é de apenas 1,71 exemplos por peso para $K = 1$, e se reduz ainda mais para valores maiores de *K*. Dessa forma, as análises relativas à Tabela 6 apresentadas a seguir levam em consideração apenas os experimentos para os quais $K = 1$.

O experimento com 10 épocas de treinamento produziu o menor RMSE, 0,1481, que é 6,4% menor do que o RMSE do sistema de referência para o conjunto de dados completo. Já no caso das métricas baseadas nas posições da ordenação reconstruída, os experimentos com as maiores quantidades de épocas de treinamento produziram os melhores resultados. Mais especificamente, o experimento com 1.000 épocas de treinamento obteve a maior PP, 49,55%, que é 22,3% maior do que a PP do sistema base, e o experimento com 500 épocas de treinamento obteve o menor EAMP, 0,8509, que é 22,3% menor do que o EAMP do sistema de referência. O gráfico da Figura 8 mostra que, apesar desses resultados, a melhoria no valor dessas duas métricas é relativamente pequena após a execução de 100 épocas de treinamento.

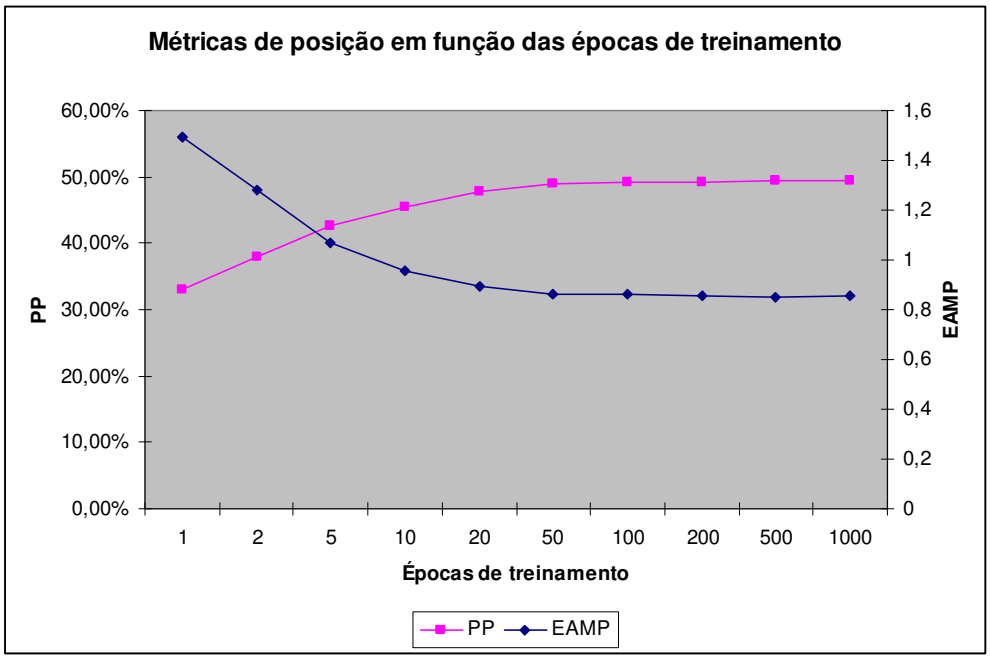

Figura 8: Métricas de posição em função das épocas de treinamento.

A Tabela 7 exibe os resultados de mais experimentos do algoritmo FM, dessa vez utilizando o conjunto de dados reduzido. Tais experimentos foram similares àqueles realizados com o conjunto completo, variando-se a quantidade de atributos latentes e de épocas de treinamento pelos mesmos intervalos de valores utilizados anteriormente.

Tabela 7: Resultados dos experimentos do algoritmo FM com o conjunto de dados reduzido, variando os parâmetros de treinamento *K* e *T*.

Épocas de treinamento (T)	Atributos latentes (K)								
	1			2			3		
	RMSE	PP	EAMP	RMSE	PP	EAMP	RMSE	PP	EAMP
1	0,1989	34,98%	1,3312	0,5935	25,98%	1,707	0,927	24,70%	1,7741
2	0,167	40,16%	1,1149	0,4558	29,22%	1,5467	0,6987	27,26%	1,651
5	0,1417	46,54%	0,8918	0,3161	36,17%	1,2598	0,4751	33,36%	1,3854
10	**0,1383**	49,56%	0,8005	0,2505	42,34%	1,036	0,3669	39,45%	1,1506
20	0,1461	51,40%	0,7603	0,218	46,97%	0,8985	0,3059	44,62%	0,983
50	0,1573	51,99%	0,75	0,2005	49,77%	0,8122	0,2641	48,18%	0,8719
100	0,1641	52,06%	0,7422	0,1928	50,22%	0,7901	0,2351	49,73%	0,8167
200	0,1666	52,21%	0,7436	0,1852	50,53%	0,7804	0,2092	50,06%	0,7958
500	0,1671	52,22%	**0,7412**	0,1765	50,31%	0,7844	0,1846	49,00%	0,8046
1.000	0,1671	**52,26%**	0,7451	0,1737	50,68%	0,7798	0,1773	49,07%	0,8094

Note que, para os mesmos valores de K e T, os experimentos com o conjunto de dados reduzido obtiveram, em todas as métricas de avaliação, melhores resultados do que os experimentos com o conjunto de dados completo. Essa melhora indica que uma matriz de dados mais densa também facilita o aprendizado de um modelo preditivo quando utilizamos o algoritmo FM. Além disso, de forma análoga ao que ocorreu com o conjunto de dados completo, os resultados obtidos com o conjunto reduzido pioram conforme aumentamos K, fato pelo qual as análises apresentadas a seguir levam em consideração apenas os experimentos com $K = 1$.

Mais uma vez, o experimento com 10 épocas de treinamento produziu o menor RMSE, 0,1383, que é 7,1% menor do que o RMSE do sistema de referência para o conjunto de dados reduzido. Também no caso das métricas baseadas nas posições da ordenação reconstruída, os experimentos com as maiores quantidades de épocas de treinamento novamente produziram os melhores resultados. Mais especificamente, o experimento com 1.000 épocas de treinamento obteve a maior PP, 52,26%, que é 14,96% maior do que a PP do sistema base, e o experimento com 500 épocas de treinamento obteve o menor EAMP, 0,7412, que é 18,91% menor do que o EAMP do sistema de referência.

Os experimentos mostram que, para os dois conjuntos de dados e as três métricas de avaliação utilizadas, o método de predição do CTR proposto neste trabalho obtém resultados comparáveis aos do sistema de referência. Assim, constatamos a validade de tal método, o que significa que ele pode ser aplicado de forma satisfatória à predição do CTR em conjuntos de dados reais.