4
Experimentos

Com o objetivo de avaliar o desempenho do algoritmo *PLSA* em sistemas de recomendação foram realizados experimentos em dois domínios, a recomendação de anúncios na web e a recomendação de filmes. Neste capítulo descrevemos na seção 4.1 o conjunto de dados e as configurações utilizadas para os experimentos. Na seção 4.2 apresentamos os resultados obtidos, e, na seção 4.3 a conclusão retirada dos experimentos.

4.1. Descrição

Para que seja possível recomendar anúncios na web através de métodos de filtragem colaborativa é necessário ter uma base com dados compostas de consultas realizadas e anúncios que foram exibidos ou clicados para cada consulta. Foi realizada uma pesquisa por conjuntos de dados com posicionamento de anúncios na web, mas não foram encontradas bases de dados públicas com estas informações. 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 (Wpt, 03), que contém uma coleta da Web Portuguesa de 2003, juntamente com os registros de seis meses de consultas submetidas à máquina de busca TUMBA! (Tumba, 2008). Em seguida, submetemos cada uma das palavras-chave dessas consultas à máquina de busca Google (Google, 2008), obtendo os anúncios do AdWords (AdWords, 2008) 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.

Essa matriz possui 55.747 linhas, representando as palavras-chave, 50.608 colunas, representando os anúncios, e 182.090 elementos conhecidos, correspondentes aos posicionamentos relativos. Note que a matriz é extremamente esparsa, sendo conhecidos apenas 0,006% de seus elementos, o que resulta nas médias de apenas 3,27 posicionamentos por palavra-chave e 3,6 posicionamentos por anúncio.

Para efetuar a avaliação do algoritmo no contexto de recomendação de filmes foi utilizado o conjunto de dados no *Netflix Prize* (Netflix Prize, 2008). Este disponibiliza um conjunto de treino com 100 milhões de notas dadas por 480 mil usuários a 18 mil filmes, sendo a escala das notas entre 1 e 5 estrelas. Também é fornecido pelo *Netflix Prize* dois conjuntos de dados para teste, o primeiro é disponibilizado para teste localmente, sem submissão ao prêmio, fornecendo assim um conjunto com 1.4 milhões de pares usuário/filme com notas presentes no conjunto de treino. Já o segundo conjunto de testes tem como

objetivo gerar uma submissão ao prêmio, contendo 2.8 milhões de pares usuário/filme onde suas notas não estão presentes no conjunto de treino.

A fim de prover um sistema de referência para comparação com o algoritmo *PLSA*, foi utilizado o algoritmo de Média dos Itens (MI), que fornece como predição para uma posição desconhecida a média de todos os valores conhecidos para o item em questão. Foi adotado este algoritmo, pois, como mostrado na tabela 4.2, ele se mostrou mais efetivo do que a média dos usuários ou a média das médias para o conjunto de dados de anúncios. Já para o conjunto de dados de filme existem diversos sistemas de referência reportados para o mesmo corpus. Dentre eles se se destacam a média das médias (MM) que consiste na combinação linear da média do usuário e do item em questão, e o sistema atual utilizado pelo Netflix o Cinematch, que é construído com base no coeficiente de correlação de Pearson.

Para o conjunto de dados de anúncios a avaliação foi realizada com uma validação cruzada de 20 iterações. Para cada iteração, o conjunto de dados utilizado é dividido em 95% dos exemplos para treinamento e 5% para teste. A fim de garantir que haja dados suficientes no conjunto de treino para o aprendizado dos atributos latentes, cada um dos exemplos que compõem o conjunto de teste obedece aos seguintes critérios:

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

Com o objetivo de medir a influência da inicialização das probabilidades no resultado do *PLSA* para o corpus de recomendação de anúncios foram utilizadas 3 estratégias. São elas:

i. AL – Aleatória. Cada probabilidade é incializada com um número aleatório, sendo posteriormente normalizado para representar uma probabilidade.
ii. CL1 – É executado o algoritmo de clustering k-means (Steinhaus, 1956) nos dados inicialmente. Após a definição de em que cluster cada exemplo está a probabilidade de um exemplo pertencer um cluster é 2 vezes maior no cluster em que o k-means o classificou do que em qualquer outro cluster.

iii. CL2 – É executado o algoritmo de clustering k-means (Steinhaus, 1956) nos dados inicialmente. Após a definição de em que cluster cada exemplo está a probabilidade de um exemplo pertencer a um cluster é *n* vezes maior no cluster em que o k-means o classificou do que em qualquer outro cluster, onde *n* representa o número de clusters gerados no *k-means*.

Já para a análise dos resultados no conjunto de dados de filmes não foi realizada validação cruzada, uma vez que o conjunto de treino e de teste é sempre o mesmo, e, apesar de ocorrerem variações nos valores inicialização estes não causam grande influência no resultado dos uso das métricas para cada experimento individualmente. Nos experimentos foram utilizados 2 tipos de inicialização, o *AL* descrito anteriormente e o *CL3* descrito abaixo:

iv. CL3 – É executado o algoritmo de clustering k-means (Steinhaus, 1956) nos dados inicialmente. Após a definição de em que cluster cada exemplo está, a probabilidade de um exemplo pertencer a um cluster é proporcional à distância do centróide do cluster em questão.

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. Tal métrica indica o quanto os valores preditos, em média, estão distantes dos valores reais, conferindo maior peso a erros grandes.

Na instância de recomendação de anúncios a ordenação dos mesmos é um fator importante, pois influencia diretamente na qualidade da recomendação. Dessa forma, para cada palavra-chave, os anúncios foram ordenados decrescentemente por posição predita, 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 posições preditas, 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 na predição e a posição que eles ocupavam originalmente na ordenação real.

Todos os experimentos foram rodados em uma máquina com processador AMD X2 4200+ (Freqüência 2.2 Ghz) e 2GB de memória RAM.

4.2. Resultados

No domínio de recomendação de anúncios na web tem-se como objetivo recomendar uma série de anúncios ordenados por relevância dada uma palavra chave. Com o intuito de simplificar os experimentos o corpus gerado sinteticamente obteve apenas 10 anúncios para cada palavra chave. Esta característica permite medir o erro de predição em cada faixa de posição. Tal medição é importante, pois se espera que em uma aplicação real os usuários se interessem mais pelos anúncios das primeiras posições, sendo essencial uma boa recomendação nestas. Com isto realizamos experimentos utilizando as métricas EMAP e PP para as seguintes faixas de posicionamento Top1, Top3, Top5 e Top10, onde estas representam respectivamente a consideração das seguintes posições: Somente o anúncio recomendado na primeira posição, os 3 primeiros anúncios, os 5 primeiros e todas as 10 posições. A tabela 4.1 mostra a distribuição dos exemplos do corpus utilizado para treinamento nas faixas descritas.

Tabela 4.1 – Distribuição dos exemplos por posição

Nível	Número de Exemplos	Percentual dos dados
Top 1	55747	30,62%
Top 3	114383	62,82%
Top 5	145721	80,03%
Top 10	182090	100,00%

Com o objetivo de medir a influência do número de grupos latentes na predição das posições dos anúncios foram realizados experimentos variando o número de grupos latentes de 5 até 1000. Além disto, foram experimentadas as estratégias AL, CL1 e CL2 para inicialização das probabilidades do *PLSA*. Os resultados dos experimentos realizados são mostrados nos gráficos das figuras 4.1 a 4.6.

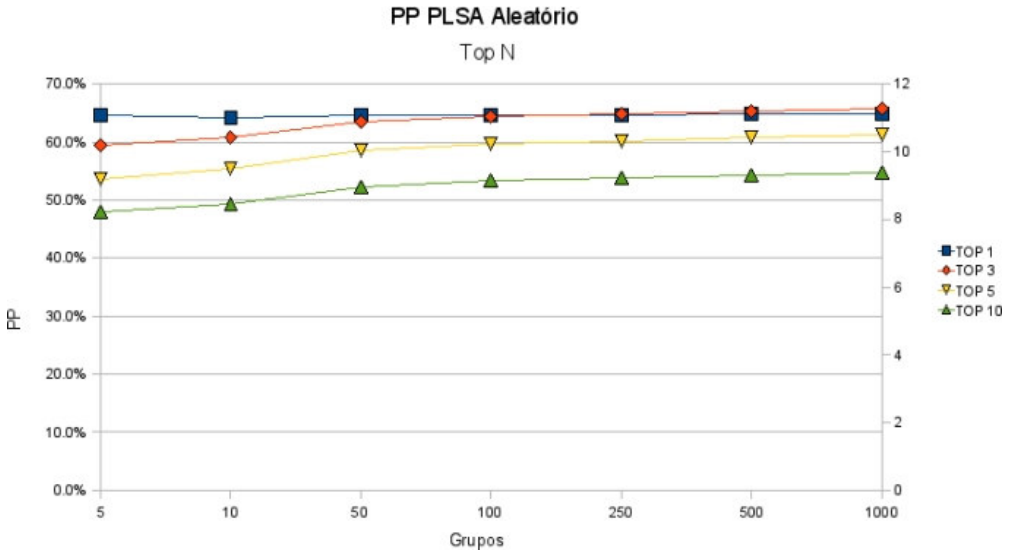

Figura 4.1 – Métrica PP pelo número de grupos latentes no experimento PLSA com inicialização AL.

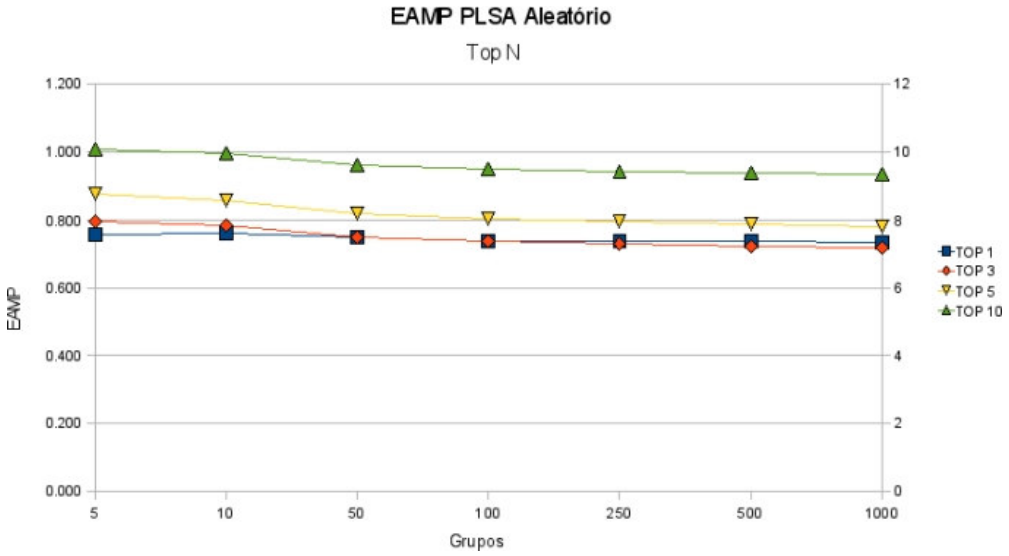

Figura 4.2 – Métrica EAMP pelo número de grupos latentes no experimento PLSA com inicialização AL.

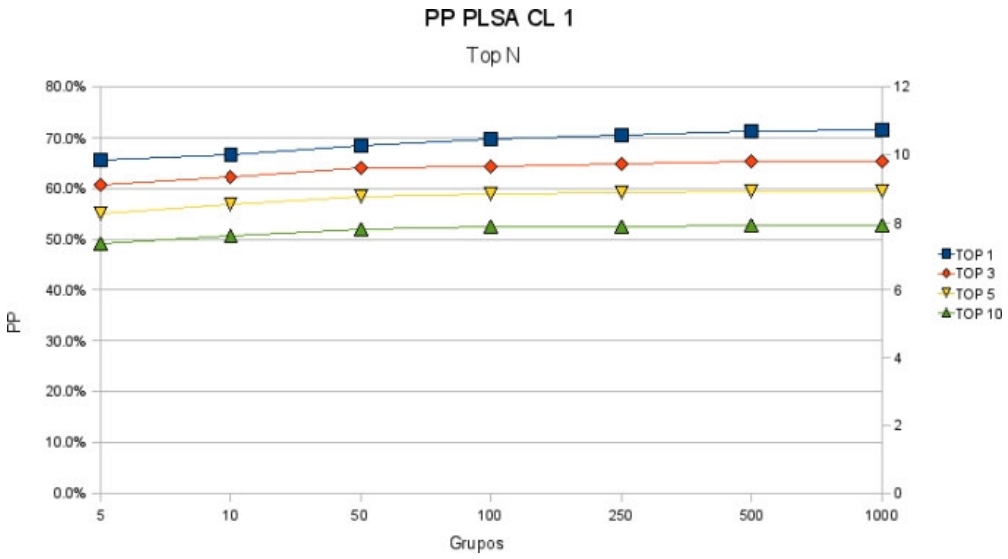

Figura 4.3 – Métrica PP pelo número de grupos latentes no experimento PLSA com inicialização CL1.

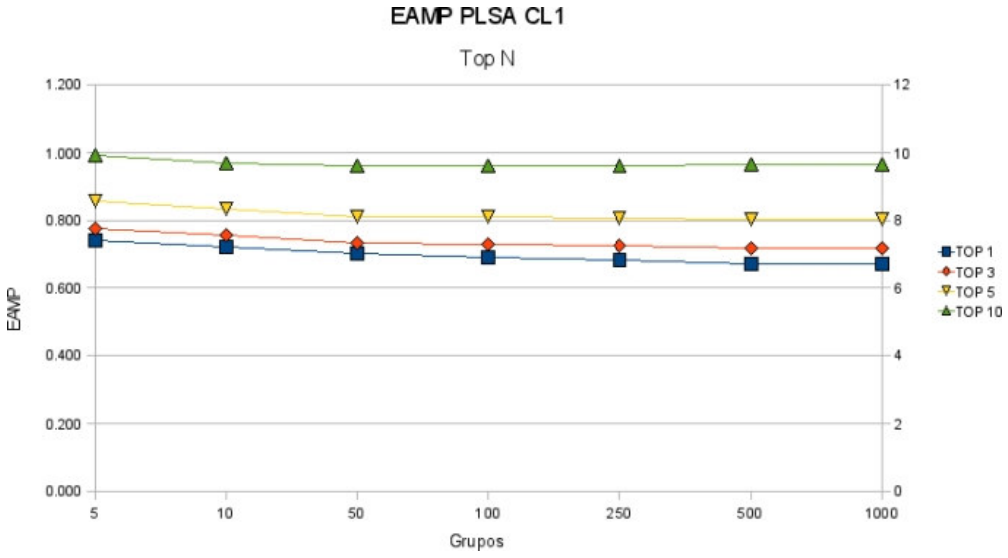

Figura 4.4 – Métrica EAMP pelo número de grupos latentes no experimento PLSA com inicialização CL1.

Através dos gráficos apresentados observa-se uma melhoria nas métricas PP e EAMP em todos os experimentos conforme é aumentado do número de grupos latentes. No experimento com a inicialização aleatória das probabilidades (AL) observou-se que tanto a métrica PP quanto EAMP tiveram os melhores resultados para a faixa Top 3. Isto se deve do algoritmo não conseguir aprender os resultados

da primeira posição tão bem quanto das outras com esta inicialização. Já para os experimentos realizados com as inicializações CL1 e CL2 foram obtidos os melhores resultados nas posições Top 1, e, foi observada uma melhoria proporcional em todos os níveis de posição. Isto mostra que efetuar uma clusterização a priori proporciona um aprendizado melhor distribuído em todas as posições.

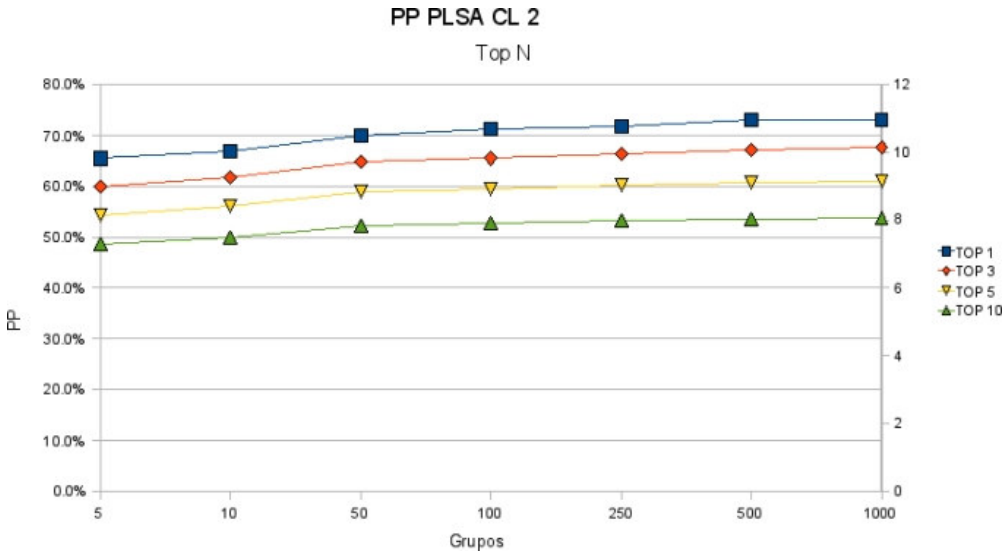

Figura 4.5 – Métrica PP pelo número de grupos latentes no experimento PLSA com inicialização CL2.

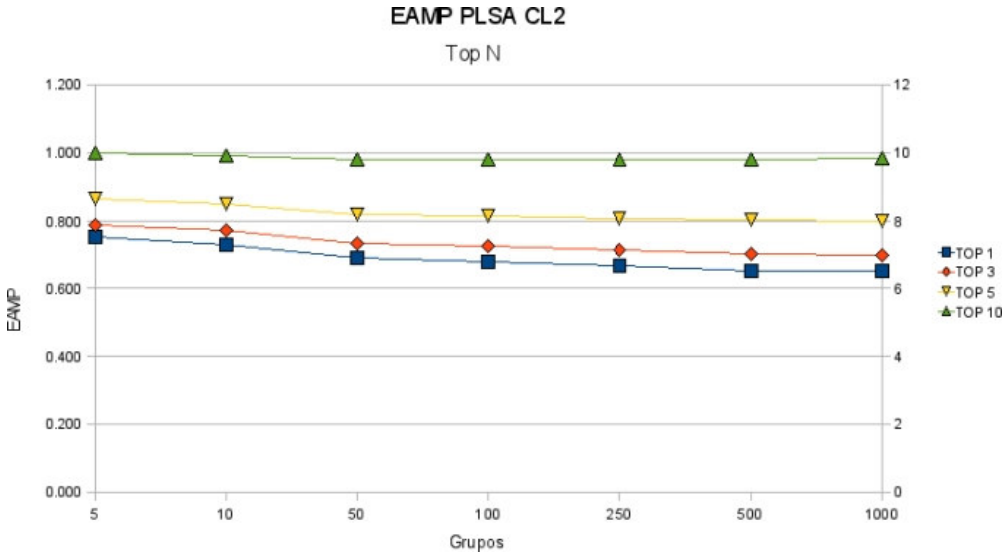

Figura 4.1 – EAMP pelo número de grupos latentes no experimento PLSA com inicialização CL2.

Com o objetivo de analisar o desempenho global do método apresentamos na tabela 4.3 os melhores resultados obtidos com o PLSA em comparação com o sistema de base Média dos Itens (MI), nesta são mostrados: o melhor valor encontrado de cada métrica por inicialização e faixa de posicionamento, a quantidade de grupos latentes no qual foi este foi encontrado, e a melhoria obtida sobre o sistema de base. Também apresentamos um comparativo dos resultados obtidos para o nível de posicionamento Top 10 com os resultados apresentados por Cavalcante (2008) utilizando o algoritmo SVD (*Singular Vector Decomposition*) para o mesmo corpus.

Tabela 4.2 – Sistemas de base Média dos Usuários (MU), Média dos Itens (MI) e Média das Médias (MM) para o corpus de anúncios na web.

	EAMP			PP		
	MU	MY	MM	MU	MY	MM
Top 1	1,263	0,710	0,713	0,356	0,683	0,680
Top 3	1,246	0,827	0,828	0,325	0,568	0,567
Top 5	1,261	0,900	0,901	0,298	0,515	0,513
Top 10	1,333	0,998	0,998	0,270	0,464	0,463

Tabela 4.3 – Comparação dos modelos do PLSA com sistemas de base para o problema de recomendação de filmes.

Modelo	EAMP	PP	EAMP – Grupo	PP – Grupo
Top 1				
PLSA AL	0,734	64.9%	1000	1000
PLSA CL 1	0,670	71.5%	1000	1000
PLSA CL 2	0,652	73.2%	500	1000
MI	0,710	68.3%	8,2%	7,2%
Top 3				
PLSA AL	0,718	65.8%	1000	1000
PLSA CL 1	0,718	65.5%	1000	1000
PLSA CL 2	0,700	67.6%	1000	1000
MI	0,827	56.8%	15,4%	19,0%
Top 5				
PLSA AL	0,780	61.3%	1000	1000
PLSA CL 1	0,803	59.5%	500	1000
PLSA CL 2	0,799	61.0%	1000	1000
MI	0,900	51.5%	11,2%	18,4%
Top 10				
PLSA AL	0,934	54.8%	1000	1000
PLSA CL 1	0,961	52.8%	50	500
PLSA CL 2	0,980	53.8%	50/100	1000
SVD	0,850	49.5%	-9.0%	10,7%
MI	0,998	46.4%	6,4%	18,1%

Observamos que tanto na PP quanto no EAMP o PLSA superou o sistema de base MU em todos os níveis de posicionamento. O PLSA se destaca claramente na métrica de precisão no posicionamento atingindo 73,2 % de precisão na primeira posição, o que representa uma melhoria de 7,2 % sobre a média dos usuários. Além do índice absoluto de precisão considerável na primeira posição o PLSA apresenta uma melhoria de 18,1% sobre a média dos usuários na precisão do posicionamento levando em consideração as 10 posições, e, apresenta também uma melhoria de 10,7% sobre o algoritmo SVD para a mesma medida. Vale a pena destacar que as melhorias encontradas na métrica EAMP foram também significativas, porém o algoritmo SVD se mostrou mais eficiente nesta métrica do que o PLSA apresentando 9,0% de melhoria sobre este.

Efetuando a comparação entre os métodos de inicialização do PLSA evidenciamos que a inicialização aleatória (AL) apresenta melhores resultados para o Top 5 e Top 10 já a inicialização CL2 apresentou os melhores resultados para o Top 1 e Top 3. Aparentemente a informação a priori da clusterização forneceu ao modelo dados para que ele pudesse aprender melhor nas primeiras posições. Porém, esta mesma informação impediu uma generalização tão boa quanto a utilização de dados aleatórios nas últimas posições, onde, justamente o problema é mais difícil.

Foram realizados 4 experimentos com o corpus do Netflix Prize para o problema de recomendação de filmes. Tais experimentos visam avaliar o desempenho do algoritmo com a variação do número de grupos latentes e a inicialização das probabilidades do PLSA. Para isto foram experimentados os modelos com 5 grupos e 40 grupos latentes, tanto com a inicialização aleatória (AL) quanto com uma clusterização *K-means* a priori (CL3). Os resultados dos experimentos são mostrados na tabela 4.4 e 4.5, nos experimentos foram realizadas 10 iterações no PLSA.

Foi observada a convergência do algoritmo para um mesmo erro quadrático em todos os experimentos, havendo apenas variações no RMSE a partir da 4ª casa decimal. Este fato é justificado pela distribuição dos dados do Netflix Prize tornar o problema eminentemente difícil, o dificulta o aprendizado do PLSA sem uma boa inicialização a priori. Em ambas inicializações não é fornecida semântica aos grupos, o que intuitivamente levaria da convergência para um mesmo RMSE em todos os experimentos. Apesar desta dificuldade apresentamos no quadro 4.6 o

erro quadrático de alguns sistemas de bases para o conjunto de dados do Netflix, e, no quadro 4.7 a comparação do RMSE de convergência obtido. Observa-se que o PLSA obteve mesmo sem uma inicialização satisfatória resultados superiores a 5 dos 6 sistemas de base, sendo a 3,78% melhor que o sistema Média das Médias é inferior somente que o Cinematch, o qual é 1,58% melhor que o PLSA. Tais resultados demonstram a qualidade do modelo *PLSA* para sistemas de recomendação mesmo quando o conhecimento a priori é pequeno.

Tabela 4.4 – Experimentos com o *PLSA* no corpus *Netflix* – 10 iterações Inicialização aleatória (CL2)

Grupos Latentes	Treino	Teste
5	0,921	0,966
40	0,921	0,966

Tabela 4.5 – Experimentos com o *PLSA* no corpus *Netflix* – 10 iterações por Clusterização K-Means (CL3)

Grupos Latentes	Treino	Teste
5	0,921	0,966
40	0,921	0,966

Tabela 4.6 – Resultado dos sistemas de base para o corpus do *Netflix*

Sistema de Base	RMSE
3 estrelas	1.313
Média (3.6)	1.130
Média dos Filmes	1.052
Média dos Usuários	1.043
Média das Médias	1.004
Cinematch	0.951

Tabela 4.7 – Melhorias obtidas sobre os sistemas de base

Sistema de Base	RMSE
3 estrelas	26,43%
Média (3.6)	14,51%
Média dos Filmes	8,17%
Média dos Usuários	7,38%
Média das Médias	3,78%
Cinematch	-1,55%

4.3. Sumário

Foram apresentados experimentos utilizando o PLSA para dois domínios de sistemas de recomendação, a recomendação de anúncios na web e a recomendação de filmes. No primeiro a ordenação dos anúncios é muito importante para a aceitação do usuário, levando-se a utilizar métricas baseadas na ordenação das predições como a PP e EAMP. O PLSA mostrou um excelente desempenho nestes experimentos apresentando 18,1% na PP de melhoria sobre a média dos usuários e 10,7% sobre o SVD. Já no corpus para a recomendação de filmes a ordenação não é o fator mais importante, sendo necessária a otimização do RMSE. Nos resultados apresentados foi obtida uma melhoria de 3,78% sobre a médias das médias, porém o corpus proporciona uma maior dificuldade no aprendizado do modelo, tornando necessária uma pesquisa melhor sobre o comportamento dos dados para que o PLSA possa otimizar o seu aprendizado.