Um Estudo sobre Categorização Hierárquica de uma Grande Coleção de Textos em Língua Portuguesa

Silvia Maria Wanderley Moraes[1], Vera Lúcia Strube de Lima

PUCRS – Programa de Pós-Graduação em Ciência da Computação
Av. Ipiranga, 6681 – Prédio 32 – 90619-900 Porto Alegre, Brasil
{silvia.moraes,vera.strube}@pucrs.br

Abstract. *This paper presents an investigation about hierarchical text categorization using the k-Nearest Neighbor algorithm to classify a large corpus in Portuguese, the PLN-BR CATEG corpus. Besides describing the main difficulties to carry on the experiments and to analyze the obtained results, this work empirically studies the influence of some of the parameters in the categorization process.*

Resumo. *Este artigo apresenta um estudo sobre categorização hierárquica de documentos que utiliza o algoritmo k-Nearest Neighbor para classificar uma grande coleção de textos escritos em língua portuguesa, o corpus PLN-BR CATEG. Além de descrever as principais dificuldades encontradas para a realização dos experimentos e para a análise dos resultados, este trabalho estuda de forma experimental a influência de determinados parâmetros no processo de categorização.*

1. Introdução

A categorização (ou classificação[2]) de textos é o processo de automaticamente atribuir uma ou mais categorias predefinidas a documentos textuais [Sun e Lim 2001]. O surgimento de grandes bibliotecas digitais e a própria Internet, se vista como um imenso repositório de textos, têm motivado pesquisas em categorização, dado que o volume de documentos armazenados torna proibitiva a classificação manual. O foco dessas pesquisas tem sido, principalmente, melhorar o processo de recuperação de informações desenvolvendo métodos computacionais capazes de classificar textos de forma rápida e precisa, envolvendo técnicas de aprendizagem de máquina e lingüística computacional. A diversidade dos textos, no entanto, pode levar à definição de um grande número de categorias, o que dificulta a navegação e a busca de informações sobre essas categorias, exigindo mais recursos computacionais. Uma solução comumente adotada é a organização das categorias na forma de uma hierarquia.

O processo de classificação hierárquica inicia na categoria genérica da hierarquia (a raiz) e se repete para as subcategorias, até atingir as categorias mais específicas (as folhas). Reduz os recursos computacionais visto que são executados apenas os

[1] Sílvia M. W. Moraes também é professora da Universidade Luterana do Brasil.
[2] Cabe ressaltar que neste documento os termos categoria e classe são usados como sinônimos. O mesmo vale para categorização e classificação.

classificadores dos ramos da hierarquia que foram ativados durante o processo [Sun e Lim 2001]. Em nosso trabalho, foi utilizada a hierarquia definida por Langie em [Langie 2004], que estrutura as categorias na forma de uma árvore. Langie descreve experimentos que servem como base a nosso trabalho, no entanto com um volume bem menor de documentos. Ele desenvolveu e testou um categorizador hierárquico formado por vários classificadores multicategoriais, que implementam o algoritmo *k-Nearest Neighbor* (*k-NN*), e analisa de forma experimental a influência de determinados parâmetros no processo de classificação sobre uma coleção de 2.896 textos jornalísticos escritos em língua portuguesa. Entre os parâmetros analisados estão o número de vizinhos considerados pelo algoritmo (*k*) e sua relação com a quantidade de documentos usados no treinamento dos classificadores, o número de características escolhidas durante a seleção de atributos, bem como a própria estratégia de classificação.

No presente trabalho são usados 26.606 textos jornalísticos também escritos em língua portuguesa do *corpus*[3] PLN-BR CATEG[4]. Diferente do trabalho de Langie, que usou uma coleção de textos previamente classificada, nosso experimento trabalha a categorização sobre uma coleção de documentos não rotulados, que se encontram apenas organizados sob títulos de 29 seções da Folha de São Paulo. Nosso trabalho convive com um dos grandes problemas na área de categorização automática de textos, que é a ausência de coleções já classificadas [Sebastiani 2006]. Construir tais coleções, além de exigir um grande esforço manual, requer um tempo considerável. Esse é um problema freqüente em aplicações de classificação envolvendo a língua portuguesa. Geralmente, os documentos não estão classificados ou o estão, mas em um número muito reduzido. O objetivo deste trabalho é experimentar a categorização hierárquica de textos em língua portuguesa em uma escala maior, analisando os resultados obtidos e as principais dificuldades encontradas. Faz parte do nosso estudo, também, analisar a influência dos parâmetros usados por Langie no processo de classificação, bem como a eficácia do processo de categorização em relação ao pequeno número de documentos usados para treinar os classificadores. Este artigo está organizado em 5 seções seguidas de bibliografia. A Seção 2 apresenta alguns trabalhos recentes na área de categorização de textos. A Seção 3 apresenta a coleção de documentos usada nos experimentos, incluindo o seu pré-processamento. A Seção 4 descreve os experimentos realizados. A Seção 5 apresenta uma análise dos experimentos realizados e dos resultados obtidos. E a Seção 6, por fim, apresenta as considerações finais.

2. Trabalhos Relacionados

Vários trabalhos recentes na área de categorização de textos têm demonstrado que os desafios apontados em [Sebastiani 2006] ainda são problemas em aberto. Os desafios referentes à necessidade de "garantia" de precisão do método de categorização, em qualquer que seja o contexto de classificação, e a redução do esforço humano para gerar bases pré-classificadas para treinar os classificadores, têm sido alvo de muitas pesquisas, principalmente em função da crescente e volumosa quantidade atual de documentos digitais. Para categorização hierárquica, há trabalhos que apresentam

[3] Neste texto, o termo *corpus* refere-se a uma coleção de documentos e não a uma coleção de referência.
[4] Esta coleção foi obtida através do projeto *Recursos e Ferramentas para a Recuperação de Informação em Bases Textuais em Português do Brasil* (PLN-BR), apoio CNPq #550388/2005-2.

propostas que visam melhorar a precisão da classificação, combinando técnicas ou associando classificadores a tesauros e ontologias.

Em [Cesa-Bianchi *et. al* 2006], por exemplo, se combinam classificadores SVM e classificadores Bayesianos para categorizar documentos das coleções jornalísticas RCV1[5] e médica OHSUMED[6], ambas em língua inglesa, em uma taxonomia definida pelos autores, que consiste em uma floresta de árvores ternárias de altura dois. Já em [Bang *et. al* 2006] os classificadores *k-NN* são combinados com um tesauro que contém conceitos sobre os relacionamentos existentes entre as categorias da hierarquia. O objetivo dos autores é melhorar a precisão reduzindo a ambigüidade quando há mais de uma categoria em que o documento pode ser classificado. Os autores conseguiram melhorar a precisão em 13,86% sem prejudicar o *recall* em seus experimentos com documentos em inglês obtidos de um *site* do Yahoo coreano sobre computadores e Internet. Ainda com o objetivo de melhorar a precisão, mas com foco voltado à técnica de classificação, têm surgido trabalhos que procuram melhorar o desempenho do *k-NN*, visto que é um algoritmo muito usado dada a sua simplicidade. Olsson em [Olsson 2006] estuda o relacionamento entre o tamanho do conjunto de treino e o parâmetro *k* dos classificadores *k-NN*. De acordo com Olsson, quando poucas amostras estão disponíveis para o treinamento, a precisão torna-se sensível ao valor de *k* e os melhores valores encontrados para *k* tendem a crescer acompanhando o tamanho do conjunto de treino. Já em grandes conjuntos de treino, a precisão se estabiliza em relação ao valor de *k*. Olsson usou em seus experimentos conjuntos de treino selecionados randomicamente do *corpus* jornalístico RCV1-v2[5] em língua inglesa. Foram gerados conjuntos com 100, 500, 1.000, 10.000, 50.000 e 100.000 documentos.

Por outro lado, [Hoi *et. al* 2006] apresenta um trabalho sobre categorização de documentos em grande escala que tem como objetivo reduzir os esforços humanos no processo de pré-classificação. De acordo com os autores as pesquisas nesse sentido têm apontado basicamente para duas direções: aprendizagem semi-supervisionada e aprendizagem ativa (*active learning*). Na primeira a meta é aprender o modelo de classificação a partir de textos rotulados e não rotulados. Apesar de tal tipo de aprendizagem ter apresentado bons resultados, envolve um alto custo computacional o qual tem limitado o seu uso. Já a aprendizagem ativa consiste normalmente em escolher, a cada iteração, "um texto" para ser classificado manualmente. Os autores, no entanto, propõem um método de aprendizagem ativa que a cada iteração seleciona um "lote de textos" a serem rotulados manualmente, reduzindo a redundância entre os textos escolhidos. Eles utilizam o modelo estatístico de regressão logística como classificador binário probabilístico, e a matriz de informação de Fisher[7] para selecionar os textos. Os resultados de seus experimentos de categorização sobre a coleção Reuters-21578 e algumas coleções Web foram promissores.

A maioria dos trabalhos mencionados realiza experimentos usando grandes bases de textos previamente classificadas. Para a língua inglesa, realizar experimentos que permitam medir a precisão da categorização é bem menos complicado visto que

[5] RCV1 contém 100.000 notícias ordenadas cronologicamente da coleção *Reuters*, disponível em http://about.reuters.com/. A RCV1-v2 é uma versão desta coleção e pode ser encontrada no mesmo site.
[6] OHSUMED contém 55.503 documentos e pode ser encontrada em http://medir.ohsu.edu/pub/ohsumed/.
[7] A matriz de informação de *Fischer* é muito usada em estatística para medir a incerteza de modelos.

existem vários *benchmarks* com volume expressivo de textos rotulados. Tal recurso é praticamente inexistente quando se trata de aplicações voltadas à língua portuguesa. Apesar de já existirem trabalhos como [Hoi *et. al* 2006], que tenta minimizar o esforço manual de classificar os documentos, é difícil comprovar que os métodos usados para a língua inglesa obterão resultados igualmente satisfatórios para língua portuguesa, sem *benchmarks* de tamanho considerável em português. Embora existam iniciativas nesse sentido como o *corpus Lácio-Web*[8], que reúne seis coleções, em língua portuguesa, com características diferenciadas e direcionadas para desenvolvimento de ferramentas computacionais para o processamento lingüístico, não é comum encontrar na literatura uma coleção com o volume de textos usados neste trabalho, especialmente em experimentos de categorização. E é justamente nesse ponto que se concentra nossa contribuição. Além disso, a análise realizada pode, inclusive, ser útil para trabalhos de categorização que visam melhorar a fase operacional[9] dos categorizadores automáticos, visto que diversos problemas foram identificados e são discutidos neste artigo. O trabalho pode também ser usado como ponto de partida para a classificação manual da coleção usada, já que oferece a sugestão de uma categorização "inicial" para os documentos.

3. Coleção PLN-BR CATEG

A coleção usada foi a PLN-BR CATEG que reúne 30 mil textos da Folha de São Paulo dos anos de 1994 a 2005. Para o uso em nossos experimentos, foram retirados desta coleção os textos referentes a 1994, visto que um bom número desses textos fazia parte da Folha-Hierarq [Langie 2004], coleção utilizada no treinamento do categorizador hierárquico. A Folha-Hierarq é um subconjunto da Folha-Ricol[10] e é composta por 2.896 documentos. A Tabela 1 apresenta as 29 seções da coleção e a distribuição dos textos por seção.

Tabela 1. Seções dos textos da coleção PLN-BR CATEG

Seção	#Textos	Seção	#Textos	Seção	#Textos
Agrofolha	166	***Empregos***	211	***Informática***	362
Brasil	4.788	***Esporte***	4.133	Mais!	207
Caderno Especial 2	48	FolhaInvest	156	Mundo	2.099
Caderno Especial	444	***Folha Negócios***	36	Primeira Página	152
Ciência	175	Folha Sinapse	11	Revista da Folha	3
Construção	7	FolhaTeen	216	Tudo	75
Cotidiano	5.831	Folhinha	73	***Turismo***	429
Dinheiro	3.790	FolhaVest	73	Tv Folha	207
Entrevistada 2	4	Ilustrada	2.594	***Veículos***	179
Equilíbrio	27	***Imóveis***	110		

Os textos da PLN-BR CATEG foram lematizados[11] pelas ferramentas CHAMA e FORMA, desenvolvidas por Marco Gonzalez e discutidas em [Gonzalez *et al*. 2006], cuja precisão quanto ao processo de lematização é, conforme relata a fonte, de aproximadamente 97%. Possivelmente por algum problema de formatação, 1.105

[8] *Lácio-web* pode ser encontrado em http://www.nilc.icmc.usp.br/lacioweb/.
[9] Fase *operacional* corresponde à fase posterior ao *treino* e *teste*. É a aplicação do classificador na resolução de problemas reais.
[10] Disponível em http://www.linguateca.pt/Repositorio/Folha-RIcol/.
[11] Verbos no infinitivo e substantivos na forma masculina singular.

documentos da coleção não puderam ser processados pelas ferramentas CHAMA e FORMA. Já que o problema não foi identificado a tempo de permitir-se uma correção, esses documentos também foram retirados. Desta forma, foram usados no processo de categorização apenas 26.606 textos. Como os documentos da coleção não estavam previamente rotulados com suas classes e as seções do jornal não correspondiam exatamente à hierarquia de categorias usada (definida por Langie em [Langie 2004] e ilustrada na Figura 1 deste artigo), sem mencionar o fato de o número de documentos ser demasiamente grande para realizar uma classificação manual, foi estabelecida uma equivalência entre algumas seções do jornal e as categorias da hierarquia. Essa equivalência foi obtida por assunto. A Tabela 2 exibe as seções e categorias consideradas equivalentes. Cabe ressaltar ainda que os documentos da coleção foram preparados da mesma forma que os da coleção Folha-Hierarq, sendo usada inclusive a mesma lista de *stopwords*[12]. A preparação dos textos é detalhada na próxima seção.

Tabela 2. Seções e Categorias consideradas equivalentes

Seção	Categorias Equivalentes	Seção	Categorias Equivalentes
Agrofolha	Área Rural	***Imóveis***	Finanças, Imóveis
Ciência	Ciência	***Informática***	Informática
Empregos	Finanças, Empregos	***Turismo***	Turismo
Esporte	Esportes	***Veículos***	Veículos
Folha Negócios	Finanças, Negócios		

4. Configuração dos Experimentos

Os experimentos documentados neste artigo foram baseados em [Langie 2004], trabalho que relata o desenvolvimento de um categorizador hierárquico que utiliza 7 classificadores *k-NN* multicategoriais. A Figura 1 apresenta a árvore de categorias construída pelo autor. Os classificadores são aplicados apenas nos nós cujos títulos estão em negrito. A *Raiz* representa o nó mais genérico da hierarquia e não é uma categoria.

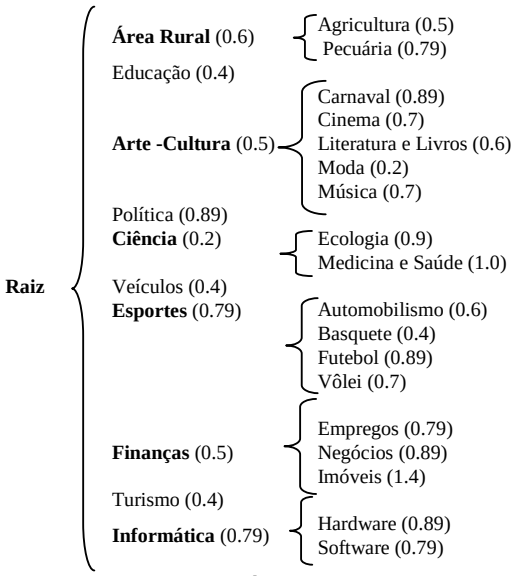

Figura 1. Árvore de categorias

[12] As *stopwords* são termos considerados de pouca valia à classificação, incluindo preposições, conjunções e alguns advérbios. Foi utilizada uma lista com 365 *stopwords.*

Em seus experimentos, Langie usou a Folha-Hierarq, composta por 2.896 documentos, sendo que 1.737 foram usados para treinar os classificadores e os 1.159 restantes, para testá-los. A preparação dos textos foi feita por classificador e, como os documentos da coleção já se encontravam lematizados, consistiu apenas em identificar os termos (únicos), selecionar os termos mais relevantes e remover as *stopwords*. Os documentos foram representados segundo a abordagem *bag-of-words* na qual um documento d_c é definido como um vetor de pesos $d_c = \{w_{1c}, w_{2c}, ..., w_{nc}\}$, onde n é o número de atributos selecionados e w_{jc} é um valor normalizado, entre *0* e *1* [Lavelli *et. al* 2004]. A técnica *Tfidf* foi utilizada para calcular os pesos dos termos. As equações (1) e (2) apresentam, respectivamente, a forma de cálculo e normalização dos pesos.

$$w_{jc} = ft(t_j, d_c) \times \log\left(\frac{q}{\#d(t_j)}\right) \quad (1) \qquad w_{jc-normalizado} = \frac{w_{jc}}{\sqrt{\sum_{i=1}^{n} w_{ic}}} \quad (2),$$

onde:

- q é o número de documentos da coleção.
- $\#(t_j, d_c)$: número de ocorrências do termo t_j no documento d_c.
- $\#d(t_j)$: número de documentos da coleção nos quais o termo t_j ocorre.
- $ft(t_j, d_c)$ é definido como: $\begin{cases} 1 + \log \#(t_j, d_c) & \text{se } \#(t_j, d_c) > 0 \\ 0 & \text{caso contrário} \end{cases}$

Os classificadores implementam o algoritmo *k-NN* de Yang e Liu [Yang e Liu 1999]. No *k-NN*, dado um texto d_i do conjunto de documentos de teste (não rotulados), o algoritmo encontra os k documentos vizinhos a esse texto que pertencem ao conjunto de treino. Os vizinhos são determinados a partir de uma métrica que avalia a similaridade entre os termos dos documentos. Yang e Liu usam a métrica co-seno apresentada na equação (3). Os escores de similaridade assim calculados são, então, usados para definir a relevância das categorias dos *k-vizinhos* (v_z) para o texto sob processo de classificação. A equação (4) apresenta o cálculo de relevância que determina o quanto uma categoria c_m é relevante para um documento d_i.

$$\cos(d_i, d_j) = \frac{d_i^T \times d_j}{\| d_i \| \times \| d_j \|} \quad (3) \qquad relevancia(d_i, c_m) = \sum_{z=1}^{k} \cos(d_i, v_z), \text{ onde } v_z \in c_m \quad (4)$$

Langie utilizou duas estratégias de classificação: *limiar baseado em rank* e *limiar baseado em relevância*. Na primeira estratégia, as categorias dos *k-vizinhos* são ordenadas conforme os seus valores de relevância e as n categorias melhor colocadas são escolhidas como as classes do documento d_i. Em seus experimentos, Langie utilizou sempre $n = 1$, ou seja, a categoria de maior relevância é a classe do documento. Já na segunda estratégia, são definidos valores limiares por categoria: $limiar(c_m)$. Durante o processo de classificação, as categorias dos *k-vizinhos* cujo fator de relevância atingir esses limiares, $relevância(d_i, c_m) \geq limiar(c_m)$, serão definidas como as classes de d_i. Na Figura 1, ao lado do nome das categorias, estão os limiares usados. Para cada estratégia de classificação, Langie realizou 5 experimentos (ver Tabela 3) com objetivo de analisar parâmetros como o valor de corte usado na seleção de atributos (*min_df*) e número k de documentos vizinhos que são considerados durante a execução do algoritmo *k-NN*. O categorizador hierárquico de Langie foi por nós reimplementado e seus experimentos refeitos. Nossa versão do categorizador foi implementada em linguagem C na plataforma Windows XP. Os resultados dos experimentos foram avaliados segundo as

medidas Precisão, *Recall* e F1, incluindo micro e macro-médias. Mais detalhes sobre essas medidas de avaliação podem ser encontrados em [Sebastiani 2002].

Tabela 3. Configuração dos Experimentos de Langie

Config.	Min_df	k	Config.	Min_df	k
1	4	13	**4**	3 se *1<=q <=500* 4 se *q >500*	13
2	4	7 se *1<=q <=250* 13 se *251<= q <= 500* 17 se *q > 500*	**5**	4 se *1<=q <=250* 5 se *251<=q <=500* 6 se *q>500*	13
3	4	13 se *1<=q <=250* 17 se *251<= q <= 500* 23 se *q > 500*			

5. Análise dos Resultados

Na presente análise foram considerados apenas os resultados obtidos para as categorias e subcategorias definidas como equivalentes às seções da coleção PLN-BR CATEG. As Tabelas 4 e 5 apresentam, respectivamente, os melhores resultados de categorização da coleção PLN-BR CATEG, ambos da configuração 3 (vide Tabela 1), para as estratégias de *limiar por rank* e *limiar por relevância*. Para facilitar a comparação foram colocados nas tabelas também os resultados de categorização da coleção Folha-Hierarq para a mesma configuração. São apresentadas as medidas Precisão (Pr), *Recall* (Re) e F1, além dos escores que contabilizam o número de documentos classificados corretamente (TP), o número de documentos rejeitados incorretamente (FN) e o número de documentos classificados "incorretamente" (FP'). No caso das subcategorias, as medidas foram calculadas a partir do TP de suas categorias-pai. Cabe frisar que, como os documentos da coleção PLN-BR CATEG não foram previamente rotulados, o valor FP' contabiliza muito mais do que os *falsos-positivos*, ou seja, há documentos que de fato pertencem a essas categorias, mas que estão em seções que não foram definidas como equivalentes. No entanto, não há como distinguir tais documentos sem um processo manual que, dada a quantidade de textos, não foi realizado. Uma conseqüência imediata deste fato é a baixa precisão de classificação na maioria das categorias escolhidas.

Exceto para as categorias *Esportes*, *Empregos* e *Imóveis***,** os índices de precisão apresentados na Tabela 4 foram muito baixos. A categoria *Finanças***,** por exemplo, teve um desempenho muito inferior ao das suas subcategorias, em especial se comparada a *Empregos* e *Imóveis*. Analisando-se os documentos categorizados pelo programa para a estratégia *limiar por rank (configuração 3)*, detectou-se que 76,35% dos textos classificados como *Finanças* eram da seção *Dinheiro*. Ao se definir a seção *Dinheiro* como equivalente a *Finanças*, e repetir alguns dos experimentos, a precisão alcançou 0,55. Outro fato que reduziu a precisão foi a presença de seções muito abrangentes como *Brasil*, *Cotidiano* e *Mundo* na nova coleção. Em média, para cada uma das categorias do nível 1 (*Área rural*, *Ciências*, *Esportes*, *Finanças*, *Informática*, *Turismo* e *Veículos*), o categorizador atribuiu 239 documentos da seção *Brasil*, 558 da seção *Cotidiano* e 129 da seção *Mundo*. Para a categoria *Finanças***,** os escores ficaram acima da média, atribuindo a essa categoria 1.030 documentos da seção *Brasil*, 1.781 da seção *Cotidiano* e 178 da *Mundo*.

Como a categoria *Esportes* foi a que atingiu os melhores resultados em ambas as estratégias, foi realizada uma análise que procurou medir a similaridade entre os termos

existentes do conjunto de treino *Esportes* (coleção Folha-Hierarq) e da seção *Esporte* (coleção PLN-BR CATEG), ano a ano. Essa mesma análise foi realizada também para as categorias *Ciência*[13] e *Turismo* que obtiveram precisão muito baixa, inferior a 0,20.

Tabela 4. Resultados da Estratégia de *Limiar por Rank* (configuração 3)

	Coleção *Pln-Br*						*Hierarq*		
Categoria	**Pr**	***Re***	**F1**	**TP**	**FP'**	**FN**	**Pr**	**Re**	**F1**
Área Rural	0,27	0,83	0,41	138	372	28	0,82	0,82	0,82
Agricultura	-	-	-	-	62	-	1,00	0,80	0,89
Pecuária	-	-	-	-	76	-	0,94	0,73	0,82
Ciência	0,08	0,59	0,13	103	1.266	72	0,98	0,64	0,78
Ecologia	-	-	-	-	**30**	-	0,62	0,57	0,59
Medicina e Saúde	-	-	-	-	**73**	-	0,61	0,67	0,64
Esportes	**0,84**	**0,95**	**0,90**	3.940	729	193	0,97	0,96	0,97
Automobilismo	-	-	-	-	336	-	0,94	0,80	0,86
Basquete	-	-	-	-	199	-	1,00	0,95	0,98
Futebol	-	-	-	-	3.103	-	0,92	0,97	0,94
Vôlei	-	-	-	-	302	-	1,00	0,83	0,91
Finanças	0,04	0,81	0,08	289	6.421	68	0,84	0,90	0,87
Empregos	**0,97**	**0,54**	**0,69**	113	3	98	0,70	0,81	0,75
Imóveis	**0,80**	**0,82**	**0,81**	90	23	20	0,62	0,89	0,73
Negócios	0,40	0,67	0,50	24	36	12	0,71	0,74	0,73
Informática	0,26	0,91	0,41	329	916	33	0,93	0,95	0,94
Hardware	-	-	-	-	135	-	0,62	0,88	0,73
Software	-	-	-	-	194	-	0,77	0,78	0,77
Turismo	0,19	0,75	0,30	320	1.373	109	0,92	0,82	0,87
Veículos	0,30	0,84	0,44	150	356	29	0,89	0,88	0,89
Macro-médias	0,41	**0,77**	0,47	-	-	-	0,81	**0,80**	0,80
Micro-médias	0,32	**0,89**	0,47	-	-	-	0,84	**0,85**	0,85

Tabela 5. Resultados da Estratégia de *Limiar por Relevância* (configuração 3)

	Coleção *Pln-Br*						*Hierarq*		
Categoria	**Pr**	***Re***	**F1**	**TP**	**FP'**	**FN**	**Pr**	**Re**	**F1**
Área Rural	0,17	0,84	0,28	140	707	26	0,62	0,95	0,75
Agricultura	-	-	-	-	66	-	0,86	0,95	0,90
Pecuária	-	-	-	-	75	-	0,95	0,86	0,90
Ciência	0,04	0,86	0,07	151	4.159	24	0.40	0.88	0.55
Ecologia	-	-	-	-	**2**	-	1,00	0,50	0,67
Medicina e Saúde	-	-	-	-	**22**	-	0,84	0,64	0,72
Esportes	**0,92**	**0,92**	**0,92**	3.785	312	348	0,93	0,99	0,96
Automobilismo	-	-	-	-	318	-	0,90	0,90	0,90
Basquete	-	-	-	-	278	-	0,85	1,00	0,92
Futebol	-	-	-	-	2.697	-	0,89	0,97	0,93
Vôlei	-	-	-	-	233	-	1,00	0,83	0,91
Finanças	0,04	0,89	0,07	316	7.956	41	0,61	0,97	0,75
Empregos	**1,00**	**0,46**	**0,63**	97	0	114	0,74	0,65	0,69
Imóveis	**0,89**	**0,65**	**0,75**	71	9	39	0,85	0,74	0,79
Negócios	0,48	0,39	0,43	14	15	22	0,73	0,67	0,70
Informática	0,25	0,88	0,39	320	948	42	0,85	0,99	0,92
Hardware	-	-	-	-	131	-	0,55	0,98	0,71
Software	-	-	-	-	196	-	0,67	0,90	0,77
Turismo	0,12	0,85	0,21	363	2.709	66	0,65	0,94	0,77
Veículos	0,17	0,88	0,29	158	763	21	0,80	0,98	0,88
Macro-médias	0,40	**0,76**	0,40	-	-	-	0,76	0,87	0,79
Micro-médias	0,24	**0,88**	0,37	-	-	-	0,71	0,92	0,80

[13] Na coleção PLN-BR CATEG, a seção Ciência só aparece entre os anos de 2000 a 2005.

Como se pode observar no gráfico B da Figura 2, o índice de termos em comum dos documentos da coleção PLN-BR CATEG em relação à Folha-Hierarq para a categoria *Esportes*, por ano, é em média de 80%. Embora as categorias *Turismo* e *Ciência*, respectivamente, consigam atingir 65% (1997) e 70% (2005) de termos em comum, isso não é suficiente, pois a freqüência desses termos nos documentos é muito baixa. No gráfico A da Figura 2 fica evidente outro fator, o da diferença de freqüência dos termos nos documentos. As seções *Ciência* e *Turismo* da coleção PLN-BR CATEG têm menos de 5 documentos, enquanto que a coleção *Esporte* tem pelo menos 15 documentos por ano que contêm os termos comuns ao treino.

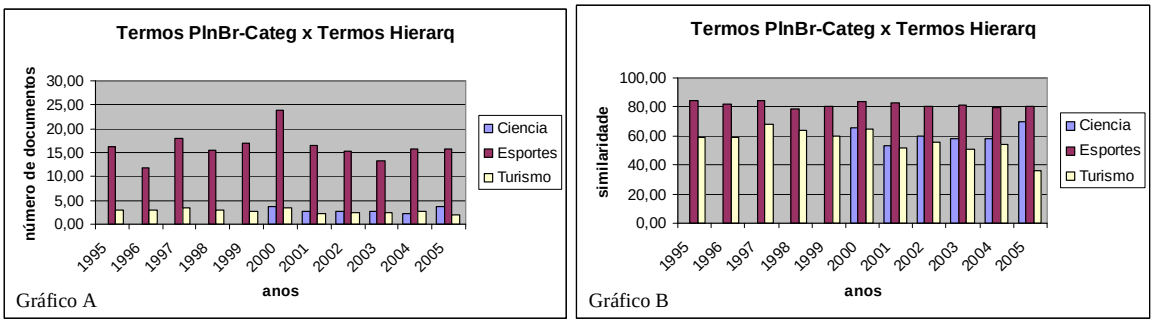

Figura 2. Gráficos comparativos

As categorias *Esportes*, *Ciência* e *Turismo* na coleção Folha-Hierarq continham, respectivamente, 480, 180 e 322 documentos. Embora *Ciência* tenha menos documentos até mesmo em relação a *Turismo* no conjunto de treino, isso não foi provavelmente a principal causa da baixa precisão. A causa parece ter sido a mudança no vocabulário. Ao longo dos 11 anos em que os textos foram produzidos, a ciência certamente evoluiu e não são mais os mesmos assuntos que estão sendo discutidos no jornal. Além disso, a área *Esportes* parece ter um vocabulário bem mais reduzido, visto que nos 11 anos foram encontrados 1.274 termos em comum, ou seja, termos que aparecem em todos os anos em pelo menos 20 documentos de cada ano. Enquanto que em *Ciência* e *Turismo* esse número cai para 280 e 207, respectivamente. Comparando as estratégias de classificação (Tabelas 4 e 5) pode-se observar ainda que foi a categoria *Esportes* quem obteve melhores resultados com a estratégia de classificação *limiar por relevância*. No entanto, analisando-se de uma forma geral esta estratégia de classificação, pode-se perceber que ela não melhorou a precisão da maioria das categorias e isso se deve, em grande parte, aos limiares definidos por Langie. Para que esta estratégia pudesse ser usada em fase operacional de categorização seria necessário algum mecanismo de atualização desses limiares. Já o *recall*, para as categorias equivalentes em ambas as estratégias de classificação, obteve resultados interessantes quanto às suas micro e macro-médias. As médias ficaram sempre acima de 0.75 e, na maioria das vezes, com valores muito próximos aos da coleção Folha-Hierarq. Uma curiosidade é que 59,33% dos documentos da seção **Brasil** foram atribuídos à categoria **política**. E a categoria **arte_cultura** recebeu 87,20% dos textos da seção **Ilustrada**.

6. Considerações Finais

Este trabalho mostrou algumas das dificuldades de se lidar com grandes coleções e o uso do algoritmo *k-NN* em categorização hierárquica. Embora a configuração 3, que varia o valor de *k* conforme o número de documentos, tenha obtido um desempenho

melhor que as demais configurações, os resultados apresentados ainda são preliminares. Mais experimentos precisam ser realizados como os feitos em [Olsson 2006] para uma análise mais "abrangente". É necessário também realizar uma avaliação manual dos resultados do categorizador, a fim de estudar sua eficácia principalmente em nível de subclasse: analisar, por exemplo, se os documentos classificados em *Agricultura* de fato pertencem a essa subcategoria. Faz parte também de nossos trabalhos futuros aplicar outros algoritmos de classificação, tais como o SVM, visto que o *k-NN* não é adequado para sistemas que exijam resposta rápida. Quanto maior o conjunto de treino, maior será o tempo de processamento desse algoritmo. Outros aspectos que precisam ser estudados envolvem a "expansão" do conjunto de treino, possivelmente através de aprendizagem semi-automática, e técnicas para definição adequada de limiares, principalmente para uso em fase operacional. Mesmo considerados todos esses pontos, acreditamos que o trabalho possui contribuição científica relevante por ser o primeiro na área de categorização automática que trabalha com uma base de textos de tamanho tão expressivo para a língua portuguesa e pelo fato de os resultados obtidos e análise realizada poderem auxiliar no processo de classificação manual da PLN-BR CATEG.

Referências

Bang, S. L, Yang, J.D e Yang, H. J. (2006) " Hierarchical Document Categorization with k-NN and concept-based thesauri", Information Processing & Management, Nº 42, Elsevier, p. 387-406.

Cesa-Bianchi, N., Gentile, C. e Zaniboni, L. (2006) "Hierarchical Classification: Combining Bayes with SVM", In: 23rd International Conference on Machine Learning, Proceedings**…**, Pittsburgh, PA, p. 177-183.

Gonzalez, M., Lima, V.L.S. e Lima, J.V. (2006) "Tools for Nominalization: an Alternative for Lexical Normalization", In: Workshop on Comp. Proc. Of Portuguese Lang – Written and Spoken, 7; PROPOR, 2006, Proceedings**…**, Springer-Verlag, p.100-109.

Hoi, S.C.H, Jin, R. e Lyu, M.R. (2006) "Large-Scale Text Categorization by Batch Mode Active Learning", In International World Wide Web Conference (WWW 2006), Edinburgh, Scotland, ACM.

Langie, L. C. (2004) "Um Estudo sobre a Aplicação do algoritmo k-NN à Categorização Hierárquica de Textos". Dissertação de Mestrado. Faculdade de Informática, PUCRS, 126 p.

Lavelli, A., Sebastiani, F. e Zanoli, R. (2004) "Distributional Term Representations: An Experimental Comparison" In: ACM International Conference on Information and Knowledge Management, Proceedings**…**, ACM, Washington, USA, p. 615-624.

Olsson, J.S. (2006) "An Analysis of Coupling between Training Set and Neighborhood Sizes for the kNN Classifier", SIGIR'06, Seattle, Washington, USA.

Sebastiani, F. (2002) "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol.34, Nº. 1, p. 1-47.

Sebastiani, F. (2006) "Classification of text, automatic", In Keith Brown (ed.), *The Encyclopedia of Language and Linguistics*, vol. 14, 2ª edição, Elsevier Science Publishers, Amsterdam, NL, p. 457-462.

Sun, A. e Lim, E. (2001) "Hierarchical Text Classification and Evaluation", In: IEEE International Conference on Data Mining, Proceedings**…,** Califórnia, USA, p.521-528.

Yang, Y. e Liu, X. (1999) "A re-examination of text categorization methods". In : International Conference on Research and Development in Information Retrieval (SIGIR'99), Proceedings**…,** Berkeley, CA, USA, p.42-49.