6. Trabalhos Relacionados

Durante a pesquisa desenvolvida para a confecção deste trabalho encontramos uma vasta literatura relativa a ALOs. Muitos trabalhos já foram desenvolvidos utilizando a abordagem de armazenamento de LOs (Pereira et. al., 2003), (Melo, et al., 2005), (Silvia, et al., 2005), (Leal et al., 2006), (Gomes et al., 2006), (Baruque et. Al., 2006). Mas nesta estrutura que se criou para a manipulação e armazenamento destes objetos, existe uma lacuna relativa ao processo de geração de conteúdo a ser utilizado na criação dos LOs. O único trabalho encontrado que abordou este problema foi o (Gomes et al., 2006).

Em paralelo pesquisamos diversos projetos que objetivavam a especulação em torno da melhor técnica de aprendizado de máquina aplicada na classificação de textos. Muitos dos trabalhos que estudamos visavam à comparação entre diversos algoritmos distintos utilizados em um mesmo problema. Outros visavam especificamente à criação ou o estudo aprofundado de um processo de classificação, utilizando para isso algoritmos de classificação específicos e propondo aprimoramentos.

Os problemas tratados nestes trabalhos foram os mais diversos, cada um envolvendo um domínio diferente. Cada um almejava reconhecer um tipo diferente de classificação dos textos e por isso utilizava um Corpus específico ao domínio em questão.

Alguns trabalhos focalizaram as pesquisas na comparação de resultados obtidos quando variavam as *features* utilizadas pelos algoritmos na criação dos modelos. Nestes casos, opções de parametrização eram oferecidas no pré-processamento que era feito nos textos ao transformá-los em tabelas atributo-valor a ser lida pelos algoritmos. Um exemplo de opções de parametrização é a utilização de n-gramas ou a utilização de recursos como *stemming*.

Outros trabalhos preferiram investigar o processo de aprendizado ocorrido no algoritmo à medida que novas amostras foram sendo acrescentadas ao modelo, permitindo concluir a utilidade do aprendizado semi-supervisionado. Nestes trabalhos variou-se também a quantidade de exemplos utilizados no Corpus inicial permitindo avaliar o tamanho ideal necessário do Corpus de exemplos para se atingir o melhor desempenho neste tipo de aprendizado.

A seguir descreveremos sucintamente cada um dos trabalhos escolhidos e faremos algumas comparações levando em consideração cada um destes aspectos descritos anteriormente.

6.1. Descrição dos trabalhos

O primeiro trabalho que descreveremos foi o propulsor do presente (Gomes et al., 2006). É uma tese de doutorado desenvolvida no nosso grupo de pesquisa que teve como motivação principal pesquisar, estudar, comparar e propor um modelo, uma arquitetura, um ambiente de software, capaz de integrar dados de Bibliotecas Digitais - DLs e de Sistemas de Aprendizagem. Utilizou-se para isso várias tecnologias, tal como: mediadores para fazer integração, mineração de texto para extração do conteúdo dos documentos da biblioteca digital e uma ontologia para a integração semântica dos mesmos.

O processo proposto neste trabalho para a mineração de texto e extração de conteúdo enfocava a utilização de uma metodologia de raciocínio dedutivo. Foram definidas regras para extrair automaticamente as definições contidas nos documentos digitais, com base em regras genéricas. As regras geralmente foram definidas através de uma representação comum na forma situação/ação. Um exemplo deste tipo de representação são regras do tipo “Se [condição] Então [ação]”.

Os resultados alcançados neste trabalho são descritos na tabela a seguir:

No	Referência	Def. Reais	Definições extraídas	Lixo	Definições Corretas	Precisão	*Recall*	F1
1	(Portugal, 2004)	5	16	11	4	0,25	0,8	0,38
2	(Ochi, 2004)	8	16	12	4	0,25	0,50	0,33
3	(Silva, 2004a)	6	2	1	1	0,5	0,16	0,25
4	(Lopes, 2004)	0	0	0	0	0	0	0
5	(Silva, 2004b)	9	5	0	5	1	0,55	0,71
6	(Penha, 2004)	8	13	8	5	0,38	0,62	0,47
7	(Wedemann,2004)	1	3	1	2	0,66	2	0,71
8	(Barbosa, 2004)	10	2	1	1	0	0,1	0
					Média	0,38	0,59	0,35

Tabela 9 – Resultados do primeiro trabalho relacionado

O segundo trabalho analisado (Matsubara e Monard, 2006) fez parte de um projeto desenvolvido em um grupo de pesquisa de Inteligência Computacional do ICMC-USP que teve como objeto de pesquisa principal um algoritmo de aprendizado semi-supervisionado chamado de CO-TRAINING, proposto por (Blum e Mitchell, 1998) com o objetivo de utilizá-los em problemas de classificação de texto. Adicionalmente foi desenvolvido um ambiente computacional para o pré-processamento de textos, denominado PréText.

O foco principal deste trabalho foi a analise de resultados obtidos utilizando-se o aprendizado semi-supervisionado que introduz a idéia de utilizar um pequeno número de exemplos rotulados, já que é geralmente caro e difícil de obter, e um grande número de exemplos não-rotulados, os quais se encontram facilmente disponíveis, com o objetivo de rotular mais destes exemplos para melhorar o desempenho de algoritmos de Aprendizado de Máquina.Assim como este, existem outros trabalhos que tiveram como objetivo principal a análise do aprendizado semi-supervisionado, um exemplo é o artigo que já referenciamos diversas vezes (Nigam, Mccallum, Thrun e Mitchell, 2000).

O CO-TRAINING desenvolveu três bases de textos que foram utilizadas nos experimentos. O objetivo era a classificação dos textos em duas classes relativas aos temas abordados nos textos e o último Corpus dividiu os textos em duas classes: course e non-course separando os textos que eram referentes a cursos oferecidos em universidades. Para a avaliação do processo de aprendizado semi-supervisionado podemos destacar os seguintes resultados:

	% de exemplos iniciais	Erro médio	Accuracy	Dif
Primeira iteração	1 %	20.1	79.9	
Última iteração		1.6	98.4	23
Primeira iteração	2 %	12.4	87.6	
Última iteração		1.4	98.6	12,5
Primeira iteração	5 %	6	94	
Última iteração		1.4	98.6	4,9
Primeira iteração	7 %	2.7	97.3	
Última iteração		1.4	98.6	1,3
Primeira iteração	10 %	4.7	95.3	
Última iteração		1.1	98.9	3,8

Tabela 10 – Resultados do segundo trabalho relacionado

Um terceiro trabalho que gostaríamos de citar foi desenvolvido na PUC-Rio (Steinbruch, 2006). Ele propõe dois algoritmos de classificação automática de textos baseados no algoritmo Multinomial Naive Bayes e sua aplicação em um ambiente on-line de classificação automática de notícias com realimentação de relevância pelo usuário, combinando técnicas de aprendizado de máquina e mineração de textos.

Particularmente este trabalho foca, diferente da maioria das abordagens propostas, não apenas a resolução de um problema de classificação binária de textos, mas sim a consideração do caso mais geral onde existe a possibilidade de um documento estar associado a mais de um rótulo.

Para testar a eficiência dos algoritmos propostos, foram realizados testes na base de dados da Reuters composta de um conjunto de 21.578 notícias que foram publicadas na rede de notícias Reuters, em 1987, e classificadas de acordo com 135 categorias, a maioria sobre economia e negócios. Adicionalmente utilizou-se a base de dados Ohsumed que é um subconjunto de 348.566 documentos da base MEDLINE (uma base on-line de textos sobre medicina), compilada por William Hersh e proveniente de 270 jornais médicos por um período de cinco anos (1987- 1991).

Como o domínio e o Corpus utilizados foram muito diferentes dos nossos fica complicado comparar os valores das métricas alcançados. Podemos

destacar como resultados interessantes para comparação os parâmetros e quantidade de atributos que alcançaram melhores resultados.

Utilizando a base Reuters(10) onde são considerados apenas as 10 classes que possuem a maior quantidade de exemplos, chegou-se aos resultados que se segue utilizando quase 7000 exemplos para o treinamento do modelo multinomial.

Micro *Recall*	94,26
Micro *Precision*	92,11
Micro F1	93,17
Macro *Recall*	88,98
Macro *Precision*	83,90
Macro F1	85,86

Tabela 11 - Resultados do terceiro trabalho relacionado

Ainda neste trabalho são listadas como sugestões para trabalhos futuros dois tópicos que foram abordados aqui:(1) “Possibilitar que o aprendizado seja semi-supervisionado, reduzindo a necessidade de o usuário associar para cada documento um conjunto de categorias, tarefa que pode ser bastante trabalhosa e suscetível a erros.” (2) “Inclusão de relações entre categorias, ou seja, permitir que o usuário apresente ao sistema relações entre categorias e o sistema agregue tais informações ao aprendizado.

Utilizando o algoritmo EM acrescentamos o aprendizado semi-supervisionado ao modelo multinomial reduzindo a tarefa de etiquetagem dos exemplos. Sobre a segunda sugestão sugerimos aqui a utilização de uma ontologia para a representação das relações entre as classes. Procuramos também mostrar que o uso dela com apoio a definição das classes do modelo favoreceu o processo de aprendizado de máquina. Ainda nesta linha de pesquisa outros trabalhos futuros serão sugeridos mais adiante.

6.2.Comparação com outros trabalhos

A tarefa de obter LOs a partir de textos diversos é um problema razoavelmente antigo, mas ainda não muito explorado. Diversos trabalhos relacionados ao tratamento de LOs podem ser encontrados, mas todos se baseiam em bancos de dados de LOs já desenvolvidos, não focando no

processo de extração de textos para a criação destes LOs, mas nos processos de geração manual, reutilização, troca, recuperação e armazenamentos deles.

O domínio do problema que estamos analisando trata de uma área extremamente abstrata, onde o foco principal é transmissão de um conceito. Queremos encontrar textos que transmitam conhecimento, que é algo bastante complexo de se descrever. Na verdade não estamos à procura de um tema como a maioria dos trabalhos na área de classificação de textos, mas sim de uma forma de texto, de um propósito. Estamos querendo capturar o objetivo instrucional de cada parte de um texto que pertença a um conteúdo didático qualquer. Por este motivo, as classes que propusemos para a classificação do texto tinham como atributos diferenciadores características muito difíceis de serem capturadas.

Sob este aspecto o único trabalho que encontramos com um propósito parecido com o nosso foi o primeiro trabalho abordado no tópico anterior e que foi o empregado no primeiro Corpus que desenvolvemos. Conseguimos alcançar um desempenho maior, no máximo 75% de *recall* contra 100% do presente. Os valores máximos de 66% de *precision* contra 8% do nosso, por outro lado pode querer condenar o nosso processo, mas no caso de classificação binária a medida mais importante a ser analisada é o *recall*, pois ela indica mais claramente o desempenho obtido quando a classificação se aproxima mais de uma filtragem que é o caso da seleção de LOs de definição.

Refizemos também um dos experimento de (Gomes et al., 2006) onde são extraídas de um artigo todas as definições existentes. Para isso utilizamos o Corpus de Definição desenvolvido no nosso trabalho. Este Corpus não utilizou nenhum exemplo parecido com os textos encontrados neste artigo, mas apesar disso, para o texto (Ochi, 2004) encontramos um *Recall* de 61%, ou seja, 61% das definições que foram encontradas numa análise manual foram selecionadas pelo algoritmo. No trabalho (Gomes et al., 2006) foi indicado um *Recall* de 50%. A precisão alcançou 37% e a medida F1 ficou em 46% ambas as medidas também maiores.

O algoritmo CO-TRAINING utilizado em (Matsubara, Monard, 2006) apresenta um método de escolha dos exemplos a serem rotulados diferentes dos métodos utilizados pelo algoritmo descrito neste trabalho, mas como análise

do processo de aprendizado semi-supervisionado podemos observar nos nossos resultados a mesma tendência.

% de exemplos iniciais	Accuracy CO-TRAINING	% de Aumento na Accuracy CO-TRAINING	Accuracy SQLLOMining	% de Aumento na Accuracy SQLLOMining
1%	79.9 98.4	19	50 34.1	-32
2%	87.6 98.6	11	50.1 34.2	-32
5%	94 98.6	5	50.2 34.4	-34
7%	97.3 98.6	1	78.8 81.6	4
10%	95.3 98.9	4	88.1 92.7	5
20%			92.3 95.1	3
30%			93.9 96.1	2
40%			93.6 95.4	2

Tabela 12 – Comparação de resultados com segundo trabalho relacionado

Ambos os algoritmos aumentam a Accuracy com a adição de amostras ao modelo inicial gerado a partir dos exemplos. À medida que a quantidade de exemplos iniciais cresce o ganho no aprendizado com as amostras diminui e sempre podemos encontrar uma faixa onde o aprendizado é alto e a quantidade de exemplos ainda se mantém pequena sugerindo ser a melhor escolha.

Podemos também observar que, diferente dos trabalhos mencionados (Matsubara, Monard, 2006) e (Nigam, Mccallum, Thrun e Mitchell, 2000), nos nossos experimentos tivemos que utilizar uma quantidade de exemplos iniciais bem maior para obter ganho no aprendizado com as amostras. Podemos destacar os resultados apresentados na tabela 4 que indicam a necessidade de pelo menos 400 exemplos para obter 12% de ganho com a adição de amostras. Para o trabalho (Nigam, Mccallum, Thrun e Mitchell, 2000) utilizando apenas 40

exemplos o aprendizado ocorre com um ótimo ganho de 60% e na tabela acima podemos ver que para pequenas quantidades de exemplos, os ganhos já são bem significativos também para o trabalho de (Matsubara, Monard, 2006).

Podemos atribuir esta diferença nos resultados principalmente ao fato do domínio do problema ser bem diferente nos três casos. Tanto que podemos encontrar em (Matsubara, Monard, 2006) resultados bem ruins para o experimento mais parecido com o nosso, onde foram classificados em course e non-course textos que eram referentes a cursos oferecidos em universidades. Neste experimento, além da subjetividade na definição das classes o fato da classe course possuir muito mais casos dentre os exemplos, atrapalha demais o processo de aprendizado com este tipo de algoritmo.

Conforme também mencionado em (Steinbruch, 2006), o modelo de misturas é muito sensível ao fato de não termos a quantidade de exemplos balanceada entre as classes. Nos experimentos feitos com o Corpus de Lei e Não-Lei tivemos claramente este mesmo problema, pois dispúnhamos de poucos exemplos. Pudemos observar uma diminuição de valores dos resultados quando a quantidade de exemplos ultrapassou o limite de 100. Nestes casos a quantidade de exemplos de Lei não chegava a 30% do total. Entre os 100 exemplos utilizados apenas 27 eram de Lei. À medida que acrescentamos exemplos os valores foram diminuindo o que não aconteceu com o Corpus de Definição que possuía 342 exemplos. Para o Corpus de Definição o limiar de 30% só foi alcançado quando utilizamos 1000 exemplos.

Exemplos	36	50	100	200	300
Accuracy	79%	81%	95%	88%	80%
Recall - Classe Lei	91%	93%	80%	55%	25%
Precision - Classe Lei	63%	63%	100%	100%	93%
F1 - Classe Lei	72%	74%	88%	71%	41%

Tabela 13 – Resultados obtidos com o Corpus Lei variando a quantidade de exemplos