Um Modelo Temporal-Relacional para Classificação de Documentos

Fernando Mourão, Wagner Meira Jr.

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) - Belo Horizonte, MG, Brasil

{fhmourao,meira}@dcc.ufmg.br

Abstract. *Automatic Document Classification (ADC) is one of the most relevant research problems in information retrieval. Despite the large number of ADC techniques already proposed, there is still a demand for techniques that are effective and efficient in taking into consideration relationships among terms. In this paper we propose a new network-based model for textual documents and introduce a family of relational algorithms for ADC that consider the temporal evolution of documents. Experimental evaluation of these algorithms shows that they achieve results comparable to SVM in four real datasets. Further, its simplicity, efficiency and the absence of a complex parameter tuning are characteristics that make our algorithm an interesting alternative to SVM.*

Resumo. *Classificação Automática de Documentos (CAD) é uma das mais relevantes tarefas em Recuperação de Informação. Apesar do grande número de propostas para CAD, ainda há uma demanda por técnicas eficazes e eficientes que consideram relacionamentos entre termos. Neste trabalho, propomos um novo modelo relacional para documentos textuais e introduzimos uma família de algoritmos relacionais para CAD que consideram a evolução temporal dos documentos. Avaliações experimentais mostram que tais algoritmos alcançam resultados comparáveis ao SVM em quatro coleções reais. Além disso, sua simplicidade, eficiência, bem como a eliminação de um complexo ajuste de parâmetros tornam nosso algoritmo uma alternativa interessante ao SVM.*

1. Introdução

Em meio ao grande volume de dados disponíveis, organizar e encontrar os recursos informacionais apropriados para satisfazer as necessidades dos usuários passou a figurar como um dos problemas mais estudados e desafiadores recentemente. Como grande parte desta informação é organizada como texto, Recuperação de Informação (RI) se tornou uma área de crescente interesse. Dada sua importância na organização e recuperação de dados textuais, Classificação Automática de Documentos (CAD), representa uma das mais relevantes tarefas em RI. CAD é definida como a tarefa de inferir a categoria semântica a qual um documento pertence, dado um conjunto discreto e finito de categorias conhecidas. Dentre as várias aplicações desta tarefa citamos a construção de filtros de *spam* e documentos, bem como o auxílio à navegação e pesquisa na Web. Existem, na literatura, diversas abordagens para CAD, tais como, Vizinhos mais Próximos, Classificadores Bayesianos e *Support Vector Machines (SVM)*, dentre outros [Manning et al. 2008].

Embora haja um grande e crescente número de propostas para CAD, poucas consideram uma importante característica da construção de textos, o relacionamento entre termos. Textos são organizados como sentenças, que são compostas por termos que interagem entre si. Logo, considerar tais relacionamentos pode ser importante para uma modelagem apropriada das várias classes de um domínio [Macskassy and Provost 2007].

Mais ainda, determinados termos tendem a apresentar relacionamentos importantes com subconjunto de termos distintos em cada classe, definindo diferentes vocabulários. A Figura 1 exemplifica este comportamento para o termo DNA, definindo vocabulários distintos quando este co-ocorre com determinados termos em documentos. Assim, antes a simplesmente considerar os termos isoladamente, ou mesmo co-ocorrências mais freqüentes, podemos considerar o vocabulário com o qual cada termo se relaciona. Entretanto, estudos que consideram propriedades da comunicação em CAD, usualmente, ignoram os relacionamentos entre termos. Tais estudos objetivam apenas utilizar medidas estatísticas dos termos (e.g., freqüência de uso), informações sintáticas (e.g., se um termo é substantivo ou verbo), baseado em uma simples análise gramatical [Montejo-Raez et al. 2008].

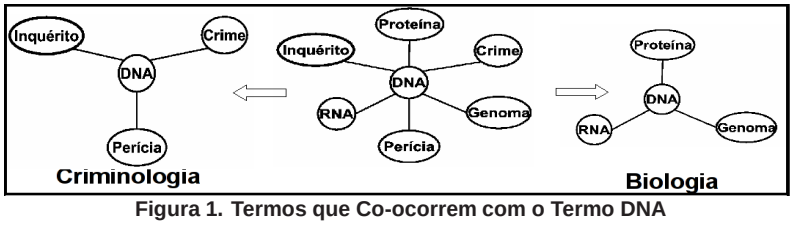

Figura 1. Termos que Co-ocorrem com o Termo DNA

Outra importante observação sobre os textos é que tais relacionamentos podem variar em intensidade ao longo do tempo como conseqüência, por exemplo, do surgimento e desaparecimento de algumas áreas de pesquisa. Redes de termos são, em geral, altamente dinâmicas, crescendo e modificando rapidamente ao longo do tempo. Assim, modelos de classificação relacionais (i.e., que consideram redes de relacionamentos) baseados em todo o histórico da rede podem apresentar um desempenho deteriorado, uma vez que informações importantes sobre mudanças comportamentais da rede são perdidas [Mourão et al. 2009]. Neste contexto, um grande desafio consiste em selecionar a granulação temporal apropriada dos dados a ser considerada para a modelagem. Por exemplo, em uma rede de co-autoria em artigos científicos, a fim de identificar a área de interesse dos pesquisadores, é importante considerar que autores publicam em períodos diferentes, sobre tópicos de pesquisa distintos devido a vários fatores, tais como novos interesses ou surgimento de novas áreas.

Dessa forma, este trabalho objetiva definir uma família de classificadores relacionais para CAD, baseada na análise de relações entre termos presentes em documentos, que seja robusta a mudanças naturais inerentes às coleções de documentos. Para tanto, inicialmente, propomos um novo modelo de classificação relacional baseado na rede de relacionamento entre termos. Nesta rede dois termos se relacionam se co-ocorrem em um mesmo documento. Posteriormente, realizamos uma discussão teórica que visa mostrar que a utilização de informações contidas nestes relacionamentos pode beneficiar CAD.

Em seguida, propomos uma família de algoritmos relacionais para CAD baseada na análise de vizinhança, sobre a rede de termos definida, inspirada em uma simples e intuitiva máxima freqüentemente utilizada: *"me digas com quem andas que direi quem és"*. Esses algoritmos determinam a classe de um documento de teste através de um processo de votação ponderada das classes estimadas para cada termo, a partir de seus relacionamentos. Diferentes funções de ponderação e estratégias de análise de vizinhança são exploradas por esses algoritmos. Apesar de sua simplicidade, mostramos que os algoritmos propostos são capazes de superar métodos simples de CAD baseados em *BAG of words* [1], alcançando resultados comparáveis ao SVM, em quatro coleções de documentos reais. Além disso, mesmo sendo transdutivo[2], nosso algoritmo (MRP-RV) pode apresentar um tempo de classificação até 60% menor que o tempo alcançado pelo SVM.

[1] Métodos baseados na simples ocorrência de atributos assumidamente independentes.

[2] i.e., nossa estratégia 'projeta' uma parte da rede contendo apenas os termos presentes em cada documento de teste.

Por fim, objetivando incorporar a dimensão temporal em nossos algoritmos, atribuímos a cada relacionamento de nossa rede a informação do momento no qual a relação foi construída, que corresponde, para CAD, ao momento no qual o documento, em que os termos ocorrem, foi publicado. Abordar tal dimensão temporal representa uma forma de obter informações mais precisas sobre o comportamento de cada termo. Essa melhoria pode ser vista, inclusive, como um aprimoramento da máxima mencionada para: *"me digas com quem andas,* **quando**, *que direi quem és"*. Utilizando uma estratégia de ponderação temporal dos relacionamentos, conseguimos verificar a hipótese que considerar a evolução temporal pode melhorar o desempenho do nosso algoritmo relacional.

2. Principais Contribuições

As principais contribuições deste trabalho podem ser sumarizadas como segue. Devido à limitação de espaço, não apresentamos uma discussão detalhada sobre cada contribuição. Maiores detalhes sobre conceitos, algoritmos e implementações relacionados a cada contribuição são apresentados no texto da dissertação [Mourão 2009].

1. Discussão teórica sobre propriedades dos termos na construção de textos e sua utilidade para CAD;
2. Proposta de um modelo de representação de documentos textuais baseado nos relacionamentos entre termos;
3. Proposta de uma família de algoritmos simples, intuitivos, eficientes e eficazes, baseado na análise de vizinhança, para CAD;
4. Incorporação do aspecto temporal nos algoritmos relacionais propostos para CAD;
5. Validação dos conceitos e algoritmos propostos em coleções reais.

3. Modelo Relacional

Nesta seção descrevemos formalmente o modelo relacional proposto, bem como uma breve discussão teórica sobre sua implementação. De maneira geral, nosso modelo objetiva expor e beneficiar-se de características associadas à interação entre termos.

O nosso modelo, similarmente ao empregado em outros contextos, consiste em uma rede na qual termos representam nodos e os relacionamentos são definidos entre termos que co-ocorrem no mesmo documento, a qualquer distância. Mais formalmente, seja D o conjunto de documentos de treino e C o conjunto de classes presentes em D. Seja também $T_i = \{t_1, t_2, \cdots, t_k\}$ o conjunto de termos distintos que ocorrem em um documento de treino $D_i \in D$, e T^D o conjunto de todos os termos distintos observados em D. Definimos uma rede $G = (V, A)$, tal que cada termo presente em T^D corresponde a um vértice (i.e., $|V| = \left|T^D\right|$). Dois termos distintos estão conectados por uma aresta se eles co-ocorrem em pelo menos um dos documentos de D. Assim, cada documento $D_i \in D$, individualmente, representa uma *clique*, visto que todos os termos de T_i possuem arestas entre si. Além disso, para cada aresta $A_{t_x - t_y} \in A$, definimos dois atributos. O primeiro consiste na classe $C_i \in C$ na qual os termos t_x e t_y co-ocorrem mais frequentemente em D. O segundo atributo, que mensura a intensidade dos relacionamento, é a *Predominância* de C_i [Rocha et al. 2008], definida como a porcentagem de vezes que $A_{t_x - t_y}$ foi observada em C_i.

A implementação e uso desse modelo, na prática, envolve a avaliação de dois grandes desafios. O primeiro deles decorre do número de relações que podem aparecer em uma rede ser eventualmente muito grande, uma vez que este número é quadrático com relação ao tamanho do vocabulário T^D. O segundo desafio, consiste no fato da qualidade das informações providas por cada relacionamento, individualmente, variar significativamente, como conseqüência de sua relevância. Assim, a fim de efetivamente implementar

o modelo, necessitamos determinar critérios que suportam a escolha de relacionamentos mais apropriados para a tarefa de classificação.

De forma a identificar tais critérios, inicialmente, avaliamos como os termos são utilizados na construção textual. Através dessa avaliação, demonstramos a existência de dois tipos distintos de termos sobre coleções reais, considerando a freqüência de aparição deles em textos. Enquanto um grupo é empregado indiscriminadamente em variados contextos (i.e., classes), o outro grupo, usualmente, está relacionado a classes específicas. O uso de ambos grupos de termos em CAD resulta em um compromisso entre capacidade discriminativa, quanto às classes, e generalidade das informações consideradas. Ou seja, considerar os termos isoladamente suscita duas questões: (1) Como podemos identificar a classe de um termo em um documento específico? (2) Como representamos de maneira mais geral esta classe? Uma resposta tradicional para essas perguntas seria empregar um ponderação sobre os termos, balanceando o uso de ambos grupos de termos (e.g., ponderação através de *TFxIDF*). Entretanto, respondemos tais perguntas analisando o relacionamento entre termos, através do nosso modelo de rede. Um fato que demonstra a importância desses relacionamentos é que se um texto for embaralhado, gerando um documento randômico totalmente desprovido de semântica, diversas das propriedades associadas à freqüência de ocorrência dos termos se mantêm verdadeiras. Avaliando em detalhes os relacionamentos entre termos, observamos que a definição de um valor mínimo de *Predominância* para a seleção desses relacionamentos representa uma estratégia promissora para seleção de relacionamentos mais apropriados para CAD.

4. Algoritmos Relacionais de Classificação

Nesta seção, apresentamos uma família de algoritmos relacionais para CAD, baseada na rede definida anteriormente. Apresentamos, também, uma avaliação experimental sobre os algoritmos propostos em quatro coleções de documentos reais, a fim de mensurar a eficácia e eficiência dos nossos algoritmos, frente a algoritmos *estado-da-arte* em CAD.

4.1. Descrição dos Algoritmos

O algoritmo 1 apresenta um algoritmo relacional simples baseado em nossa rede. Para cada termo t_i de um documento a ser classificado, definimos uma vizinhança a ser analisada, que consiste do conjunto de termos que co-ocorrem com t_i em documentos do treino. Cada termo determina uma pontuação para cada classe que ocorre em sua vizinhança. Existe apenas uma classe associada a cada relacionamento, que é a classe com maior *Predominância*. A classe do documento pode ser predita de diversas maneiras, tal como por votação majoritária considerando a mais alta pontuação de cada termo ou a classe associada com a maior soma final de pontuações.

Algoritmo 1 Modelo Baseado em Análise de Vizinhança

```
function NBAnalysis(G, node)
    finalScore[] ← 0
    repeat
        relation ← GetNextNeighbor(node, G)
        class ← relation.class
        score ← DefineScore(class)
        finalScore[class] += score
    until (node.neighborhood = ∅)
    return finalScore
```

Dessa forma, instanciações deste algoritmo envolvem três importantes decisões. A primeira refere-se à vizinhança de cada nodo que será usada para análise. A segunda consiste no peso de cada voto associado a cada relacionamento da vizinhança selecionada. E, por fim, temos o critério para definir a classe de um documento a partir dos votos induzidos por cada termo para cada classe.

Avaliações sobre diversas propostas para essas três decisões permitiram definir um algoritmo relacional denominado Modelo Relacional Ponderado com Redução de Vizinhança (MRP-RV). Baseado em nossas análises, observamos que quando consideramos todos os relacionamentos de um termo, realizamos uma análise global de seus relacionamentos, a qual difere do comportamento observado quando consideramos relacionamentos com subconjuntos distintos de termos. Dessa forma, projeções dos relacionamentos de cada termo podem definir melhor a classe do termo em cada documento de teste distinto, visto que permitem focar a análise em relacionamentos mais relevantes para cada classe. Identificar a melhor projeção certamente representa um desafio, visto que uma busca exaustiva para cada termo em cada documento é impraticável. Uma estratégia promissora consiste em usar uma estratégia sob-demanda e criar uma projeção da rede para cada documento a ser classificado. Cada documento representa uma clique no grafo, definindo um restrito conjunto de relacionamentos, os quais podem ser mais informativos sobre as características do documento que todos os relacionamentos definidos no treino para cada termo. Baseado nesta observação, no MRP-RV a vizinhança de cada termo t_i é definida apenas por outros termos de teste que co-ocorrem com t_i no documento a ser classificado. Assim, restringimos a análise para uma "visão local" de relacionamentos, a qual pode ser interessante não apenas por reduzir o número de relacionamentos a ser avaliado, mas também por focar em um conjunto de vizinhos teoricamente mais informativo para a classificação de um documento específico.

Uma vez que esta vizinhança é definida, cada relacionamento desta vizinhança, com *Predominância* acima de um limiar δ, é considerado como um voto unitário. Assim, cada termo induz uma pontuação para cada classe que ocorre na vizinhança, definida como a porcentagem de vizinhos conectados por uma classe sobre todos os vizinhos analisados. Ao invés de assumir o mesmo peso para as pontuações de todos os termos, ponderamos as pontuações de cada termo considerando o *TFxIDF* [3] do termo. Além disso, de forma a balancear o poder de voto de classes maiores e menores, visto que classes maiores possuem mais relacionamentos na rede, ponderamos os votos dados a cada classe de maneira complementar à sua probabilidade de ocorrência na coleção de documentos. Ao fim, somamos os pontos assinalados a cada classe por todos os termos de cada documento de teste, e a classe com mais alta pontuação final é assinalada ao documento.

4.2. Avaliação Experimental

A fim de tornar nossa experimentação estatisticamente robusta, em todos os experimentos executamos uma validação cruzada de 10 partes, e o resultado final de cada experimento é dado como a média das dez execuções. As principais métricas usadas para avaliar os algoritmos foram Acurácia (Ac.) e MacroF1 (MacF1). Nossas análises foram geradas sobre quatro coleções de documentos reais com características bem distintas. A tabela 1 sumariza as principais características dessas coleções. É importante salientar que apenas a base NT é composta por documentos com texto completo, todas as outras coleções são compostas apenas por resumo e/ou título. Todas as coleções foram pre-processadas, de forma a remover palavras *"stopwords"* dos documentos. Além disso, cada documento em todas as coleções está associado a somente uma classe.

Coleção	Num. Documentos	Granulação	Num. Período	Descrição
ACM	25.000	Anos	22	Artigos computação
MD	861.454	Anos	16	Artigos medicina
AG	835.795	Dias	573	Notícias Web
NT	7.964	Semanas	52	Periódicos computação

Tabela 1. Informações sobre as Bases de Dados

Análise de Eficácia

[3] Acrônimo usado na literatura para *Term Frequency* $\times$ *Inverse Document Frequency*.

Algoritmo	Coleção							
	ACM		MD		AG		NT	
	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)
Rocchio	56.97	67.95	54.14	69.36	54.14	56.81	72.21	80.80
KNN	56.78	69.8	66.57	79.82	60.85	68.89	74.37	89.13
NB	56.87	74.00	65.64	79.65	61.54	68.31	74.98	84.64
SVM	**60.07**	**73.03**	**72.28**	**83.27**	**65.05**	**72.70**	**83.58**	**92.78**
MRP-RV	**66.06**	**74.37**	**71.49**	**83.35**	**61.62**	**69.01**	**84.02**	**92.24**
ganho	**+9.97**	**+1.83**	**-1.09**	**+0.09**	**-5.27**	**-5.07**	**+0.52**	**-0.58**
t-t.	▲	▲	●	●	▼	▼	●	●

Tabela 2. Resultados de Algoritmos para CAD usando informação de *TFxIDF*

De forma a avaliar a eficácia do MRP-RV, o comparamos com quatro classificadores tradicionais incluindo KNN (usando o parâmetro $K = 30$ e similaridade cosseno), Rocchio e Naïve Bayes, implementados no software de classificação Libbow [McCallum 1996]. Também incluímos o SVM-Perf, um classificador SVM que implementa um método linear. Empregamos a estratégia *one-against-all* [Manning et al. 2008] de forma a adaptar o SVM binário para coleções com mais de duas classes. Para todos os algoritmos usamos o esquema de ponderação considerando *TFxIDF*, e os melhores parâmetros foram encontrados usando validação cruzada em um conjunto de validação. Os resultados são apresentados na tabela 2. Nesta tabela as linhas com cabeçalho "$ganho$" descrevem a diferença percentual entre o SVM e o MRP-RV, as linhas com "t-t" descrevem se essa diferença é estatisticamente significante positiva (▲), negativa (▼) ou não significante (●), dado 99% de confiança em um *2-tailed t-test*. Observamos que, apesar de simples, o MRP-RV supera os primeiros três algoritmos, sendo comparável (ou mesmo superando) o SVM em nossas coleções. Um desempenho pior que o SVM foi observado apenas com relação à coleção AG. Tal comportamento pode estar relacionado à maior ocorrência do tipo de relacionamento com características mais restritivas nesta base, reduzindo a capacidade do MRP-RV gerar uma representação mais ampla das classes.

Uma análise mais detalhada sobre nossos algoritmos mostrou, ainda, que os resultados alcançados estão limitados, sobretudo, por dois fatores. O primeiro fator consiste da eventual escassez de informação, uma vez que as filtragens realizadas sobre os relacionamentos da rede podem refletir em um pequeno número de relacionamentos resultantes. Avaliações preliminares sobre formas simples de endereçar tal fator demonstraram ganhos médios acima de 2% sobre nossos resultados. O segundo fator refere-se ao fato do conjunto de treino poder ainda conter informações conflitantes. Isso decorre em virtude de evoluções temporais inerentes às próprias coleções. De forma a abordar este fator, apresentamos em seguida uma extensão do modelo e algoritmos propostos de forma a incorporar informações temporais na classificação relacional.

Análise de Eficiência

Além de apresentar resultados comparáveis aos do SVM e não necessitar um complexo processo de ajuste de parâmetros[4], podemos enfatizar duas características que tornam o MRP-RV interessante para CAD. Primeiramente, por se tratar de um algoritmo transdutivo, não requer qualquer retreinamento para classificar documentos novos, ao contrário de algoritmos *model-based* (e.g., SVM, Naïve Bayes). A segunda característica refere-se a sua natureza local. MRP-RV classifica documentos de teste analisando apenas uma pequena porção da rede de termos. Tal estratégia não é somente efetiva, como os resultados mostraram, mas também é computacionalmente eficiente. Analisando a complexidade do MRP-RV, podemos entender a eficiência alcançada por nosso algoritmo. Observamos que o MRP-RV considera cada documento como uma clique, avaliando $T-1$ vizinhos de forma a definir a classe de cada um dos T termos de um documento a ser classificado. Assim, para classificar D documentos temos uma complexidade $O(D \cdot (T^2 - T))$.

[4]O único parâmetro a ser avaliado é a *Predominância* que, em geral, não assume valores muito baixos, por permitir relacionamentos pouco discriminativos, nem demasiado altos, por eliminar relacionamentos importantes.

Como, para coleções reais, $D \gg T$, podemos considerar que nosso algoritmo é linear quanto ao número de documentos a serem classificados. Uma comparação entre os tempos de execução do SVM e do MRP-RV demonstra que nosso algoritmo é competitivo. Por exemplo, para as coleções ACM e AG os tempos de classificação dos documentos de teste foram 16% e 60% menores que os alcançados pelo SVM, respectivamente.

5. Algoritmos Temporais Relacionais

De forma a compor um conjunto de treino, modelos atuais de CAD agregarem indiscriminadamente todos os documentos disponíveis, independente do momento de publicação. Intuitivamente, quanto maior o período observado, mais informações teríamos para definir um bom modelo de classificação. Entretanto, essa informação agregada pode se tornar imprecisa ou contraditória dada à natureza dinâmica da comunicação humana, que acarreta em mudanças constantes nos relacionamentos entre termos ao longo do tempo. Assim, um grande desafio na definição de modelos de classificação consiste em selecionar a granulação temporal mais apropriada para análise.

Nesta seção, avaliamos especificamente a hipótese que considerar a evolução da linguagem pode melhorar nosso algoritmo relacional. A fim de verificar tal hipótese, a primeira modificação necessária corresponde à forma como construímos nossa rede de termos. Como descrito na seção 3, tal rede é construída considerando todos os documentos presentes em um conjunto de treino, independente do momento de publicação dos mesmos. De forma a capturar mudanças de comportamento dos termos, é necessário definir seus relacionamentos em momentos distintos. Assim, estendemos a definição da rede para um multigrafo, em que entre dois termos quaisquer passam a existir M relacionamentos distintos, um para cada momento possível definido no domínio de análise.

Posteriormente, são necessárias algumas modificações no algoritmo 1 de forma a utilizar as informações temporais contidas nos relacionamentos. Basicamente, o que fazemos é selecionar um contexto temporal relevante para análise de cada relacionamento. Este contexto representa o período de tempo no qual os relacionamentos construídos devem ser avaliados. Como discutido em [Mourão et al. 2009], esse período deve ser definido de acordo com o momento no qual queremos classificar um dado nodo da rede, uma vez que em diferentes momentos, relacionamentos são mais prováveis de pertencerem a classes distintas. Chamamos este momento de interesse de **ponto de referência**. Na dissertação, avaliamos três estratégias para definição dos contextos temporais. A seguir descrevemos a estratégia que se mostrou mais efetiva para nossas coleções.

Análise de Ponderação Temporal

Uma estratégia para considerar mudanças temporais nos relacionamentos seria definir uma função de ponderação temporal. A idéia consiste em valorizar relacionamentos cujos comportamentos mais se aproximam dos observados em um dado ponto de referência. Dessa forma, é possível considerar todos os relacionamentos da rede, ao mesmo tempo em que valorizamos relacionamentos potencialmente mais relevantes. Todos os relacionamentos de cada momento M_i podem ser ponderados através do peso p_i, definido por uma função de decaimento temporal $p_i \propto {D_{ir}}^{-a}$, que penaliza relacionamentos temporalmente mais distantes do ponto de referência. Nesta função, D_{ir} representa a distância temporal, medida em unidades temporais definidas para o domínio, entre o momento M_i e o ponto de referência r, e a representa o fator de decaimento da função.

Definindo os parâmetros da função temporal de forma empírica sobre o comportamento da classe predominante de cada relacionamento ao longo do tempo, temos o algoritmo denominado MRP-RVT, cujos resultados são apresentados na tabela 3. Comparando

Algoritmo	Coleção							
	ACM		MD		AG		NT	
	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)	$MacF_1$(%)	Ac(%)
MRP-RV	66.06	74.37	71.49	83.35	61.62	69.01	84.02	92.24
MRP-RVT	67.87	75.94	66.89	84.89	62.98	70.66	85.15	92.86
ganho	+2.73	+2.11	-6.43	+1.85	+2.21	+3.84	+1.34	+0.67
t-t.	▲	▲	▼	•	▲	▲	•	•

Tabela 3. Resultados do MRP-RV *versus* Resultados do MRP-RVT3

tais resultados com os obtidos para o MRP-RV, observamos uma melhora para duas bases e resultados estatisticamente equivalentes para as outras duas, quanto a Acurácia. Considerando a MacroF1, observamos um deterioramento para a coleção MD, explicado pela priorização das classes maiores quando aplicamos o MRP-RVT nesta base. Assim, apesar destes resultados suportarem a hipótese inicial que abordar a evolução da linguagem poderia melhorar nosso algoritmo relacional, os ganhos obtidos ainda se mostraram tímidos, evidenciando a necessidade de formas mais elaboradas de tratar a evolução temporal.

6. Conclusões e Trabalhos Futuros

Neste trabalho apresentamos uma proposta de modelagem de documentos através de uma rede de relacionamento entre os termos que co-ocorrem em cada documento. A partir desta rede, propomos algoritmos relacionais de classificação para CAD. Os resultados obtidos demonstraram que a utilização das informações contidas nos relacionamentos entre termos é bastante útil para CAD. Além disso, embora simples, o algoritmo apresentado mostrou-se competitivo com algoritmos *estado-da-arte* para a área, dada a eficiência e eficácia observados. Extensões simples que visam endereçar limitações do algoritmo, bem como a incorporação de informações temporais no modelo foram avaliadas, demonstrando que há espaço para melhorias significativas em nossas propostas.

Como trabalhos futuros destacamos a análise de outros algoritmos relacionais, a definição de funções temporais mais robustas, além de formas mais elaboradas de abordar a escassez de informação no nosso algoritmo. A aplicação do nosso modelo a outros contextos também representa uma promissora direção de pesquisa, visto que a generalidade das propriedades lingüísticas utilizadas, permite-nos acreditar que nossa proposta é potencialmente útil em vários cenários de aplicação de CAD.

Referências

Macskassy, S. A. and Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8:935–983.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. `http://www.cs.cmu.edu/~mccallum/bow`.

Montejo-Raez, A., Urena-Lopez, L. A., Garcia-Cumbreras, M. A., and Perea-Ortega, J. M. (2008). Using linguistic information as features for text categorization. In *Proc. of the MMDSS*, Varese, Italy. Ios Press Inc.

Mourão, F. (2009). Um modelo temporal-relacional para classificação de documentos. Master's thesis, UFMG. Disponível em http://www.dcc.ufmg.br/ fhmourao/master.pdf.

Mourão, F., Rocha, L., Miranda, L., A., V., and Meira Jr., W. (2009). Quantifying the impact of information aggregation on complex networks: A temporal perspective. In *Proc. of the 6th WAW*, Barcelona, Spain.

Rocha, L., Mourão, F., Pereira, A., Gonçalves, M., and Meira Jr, W. (2008). Exploiting temporal contexts in text classification. In *Proc. of the 17th CIKM*, CA, USA. ACM.