Combinando Mineração de Textos e Análise de Redes Sociais para a Identificação das Áreas de Atuação de Pesquisadores

Bruno K. O. Miyata[1], **Vitor Y. Kano**[1], **Luciano A. Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

{bruno.miyata,vitor.kano,digiampietri}@usp.br

Abstract. *The automatic identification of expertise's areas or the main subject of a scientific paper is an important activity which works as basis for different social network analysis systems for the evaluation of interdisciplinary groups, detection of new trends and search for experts. This paper presents an approach for the automatic identification of researchers' areas of expertise combining text mining and social network analysis techniques.*

Resumo. *Identificar de forma automática a área de atuação de um pesquisador ou mesmo o assunto principal de um artigo é uma atividade importante que serve de base para diversos sistemas de análise de redes acadêmicas, por exemplo, para avaliação de grupos interdisciplinares, detecção de tendências e busca por experts. Este artigo apresenta uma abordagem para a identificação de áreas de atuação de pesquisadores combinando técnicas da mineração de textos com a análise de redes sociais.*

1. Introdução

A popularização das redes sociais com os mais variados focos (por exemplo, redes acadêmicas, de amizade, redes de contatos profissionais e de compartilhamento e discussões de fotos e vídeos) traz novas oportunidades e desafios. Por um lado, existe uma riqueza muito grande na combinação entre as informações de cada usuário e seus diferentes tipos de relacionamentos com os outros indivíduos da rede. Por outro, nem sempre todas as informações de interesse estão disponíveis; algumas redes possuem milhões de usuários com dezenas de milhões de mensagens trocadas por ano o que exige o desenvolvimento de algoritmos eficientes e/ou heurísticas para tratar um volume de dados tão grande; e muitas das interações entre usuários ou mesmo da descrição de seus atributos são feitas utilizando-se textos escritos em língua natural preenchidos pelo próprio usuário.

Os desafios relacionados ao tratamento de textos escritos em língua natural podem ser tratados utilizando-se técnicas de Mineração de Textos (MT) e processamento de língua natural (PLN). Já o entendimento dos efeitos das iterações entre indivíduos conectados são estudados pela Análise de Redes Sociais (ARS).

Redes sociais acadêmicas são redes sociais compostas tipicamente por professores e alunos e as ligações entre indivíduos podem ser de diversos tipos, incluindo relações de coautoria, orientação, ser o professor de, esclarecer dúvidas de, coparticipar de projetos, entre outras. Uma questão importante na análise de redes sociais acadêmicas é a identificação das áreas de atuação e/ou expertise dos indivíduos da rede. Isto pode ser feito com base nos atributos relacionados a cada indivíduo (por exemplo, os assuntos dos

artigos que ele publicou) ou com base nos relacionamentos de cada indivíduo (através do uso da informação de seus vizinhos).

No Brasil existe uma base de dados ímpar de currículos de pesquisadores, situada na Plataforma Lattes[1]. Os mais de 2 milhões de currículos cadastrados nesta plataforma (conhecidos como Currículos Lattes) possuem diversas informações relevantes para a análise de pesquisadores e de redes acadêmicas, incluindo: dados de formação; áreas de atuação; participação em projetos de pesquisa; produções bibliográficas, técnicas e artísticas; participação em eventos; participação em bancas; e orientações. Por possuir informações sobre as áreas de atuação do pesquisador (divididas em grandes áreas, áreas, subáreas e especialidades) esta base permite que algoritmos de MT e ARS sejam treinados e testados para a identificação automática dessas áreas.

A identificação de *experts* em dadas áreas e a atribuição de categoria ou área de atuação a usuários em redes sociais são problemas que têm sido tipicamente tratados na ARS através de algoritmos de inteligência artificial, modelos de Markov ou análise de vizinhança [Wang and Krim 2012, Wang et al. 2013]. Por outro lado, o uso de textos livres para identificação de categorias ou extração de conhecimento é tipicamente tratado por técnicas de MT e PLN [Gharehchopogh and Khalifelu 2011, Gerdsri et al. 2012].

O objetivo deste artigo é combinar técnicas de Mineração de Textos e Análise de Redes Sociais para a identificação automática das áreas de atuação de pesquisadores com base nos títulos de suas publicações e de suas redes de coautoria.

O restante deste artigo está organizado da seguinte forma. A Seção 2 contém a descrição da metodologia utilizada. Na Seção 3 os resultados são apresentados e discutidos. Por fim, a Seção 4 contém as conclusões e os trabalhos futuros.

2. Materiais e Métodos

Neste artigo, todos os dados utilizados são oriundos dos currículos da Plataforma Lattes. Cinco atividades foram realizadas: seleção dos dados; organização das informações de interesse; extração de características; execução dos experimentos; e análise dos resultados.

Seleção dos dados. Para este artigo a amostra selecionada contém os dados dos currículos Lattes dos detentores de bolsa produtividade do CNPq. Para isso foram consultadas as listas dos bolsistas produtividade de 2010 de todas as áreas de conhecimento e um *parser* identificou de maneira automática os currículos dos bolsistas. Ao todo, foram encontrados 13.797 currículos. Porém, muitos dos pesquisadores cadastram no campo "Áreas de Atuação" mais de uma grande área, área ou subárea. Para este trabalho optou-se por selecionar apenas os pesquisadores que haviam declarado apenas um valor diferente neste campo. Isto resultou em três conjuntos diferentes: um para pesquisadores com uma única grande área; um para pesquisadores com uma única área e um para pesquisadores com uma subárea. Isto ocorre porque um pesquisador pode ter declarado apenas uma grande área, mas com mais de uma área diferentes. Por exemplo, considere os dois conjuntos de áreas de atuação presentes no currículo de um pesquisado: *Grande área: Ciências Exatas e da Terra / Área: Ciência da Computação / Subárea: Matemática*

[1] lattes.cnpq.br/

da Computação e *Grande área: Ciências Exatas e da Terra / Área: Probabilidade e Estatística / Subárea: Estatística*. Este pesquisador possui uma única grande área (Ciências Exatas e da Terra) porém duas áreas ("Ciência da Computação" e "Probabilidade e Estatística") e também duas subáreas.

Desta forma, os conjuntos obtidos contêm diferentes números de pesquisadores. O conjunto formado por pesquisadores com uma única grande área contém 9.748 pesquisadores divididos nas 8 grandes áreas definidas pela Plataforma Lattes. O conjunto formado por pesquisadores com apenas uma área contém 7.297 divididos em 76 áreas de atuação. Por fim, foram identificados 3.427 pesquisadores com apenas uma subárea de atuação. Vale destacar que o campo subárea é de preenchimento manual (ao contrário dos campos grande área e área que obedecem a uma lista disponibilizada pelo CNPq) o que acarreta num número muito grande de subáreas, assim, os 3.427 pesquisadores estão distribuídos em 443 subáreas diferentes.

Para cada um dos 3 conjuntos, foram separados aleatoriamente 90% dos pesquisadores para treinamento e 10% para testes.

Organização das informações de interesse. Além das áreas de atuação, foram consideradas informações de interesse os dados dos artigos publicados em periódicos pelos pesquisadores selecionados. Foram consideradas as produções de 2001 a 2010 e estas produções serão utilizadas para montar redes acadêmicas de coautoria e os títulos dos artigos serão utilizados para a mineração de textos. A identificação de coautorias foi realizada utilizando-se o algoritmo proposto por Digiampietri et al [Digiampietri et al. 2012] específico para identificação de coautorias utilizando-se dados de currículos Lattes. Os dados foram organizados em arquivos csv (*Comma-Separated Values*), ao todo foram gerados 30 arquivos correspondendo à separação dos 3 conjuntos de dados em dados ano a ano (no período de 10 anos).

Extração de características. Neste artigo três características foram extraídas para a identificação das áreas de atuação dos pesquisadores: uma baseada em mineração de textos e duas baseadas na rede de coautorias. A característica baseada em **Mineração de Textos** utilizou a frequência relativa das palavras dos títulos dos artigos publicados em periódicos. Os títulos dos artigos publicados pelos 90% dos pesquisadores foram utilizados como treinamento da seguinte forma: para cada título foi atribuída a mesma grande área (ou área ou subárea) cadastrada pelo pesquisador. O conjunto de títulos de uma mesma grande área (ou área ou subárea) foi utilizado para criar um banco de dados contendo a frequência que cada uma das palavras possui neste banco. Além disso, foi criado um banco contendo todos os títulos das diferentes grandes áreas (ou áreas ou subáreas) e com a frequência de cada palavra neste banco. Para verificar a grande área (ou área ou subárea) de um novo título basta somar a frequência relativa de cada palavra deste título considerando cada uma das grandes áreas (ou áreas ou subáreas), da seguinte maneira: dividindo-se a frequência da palavra no conjunto de dados da grande área (ou área ou subárea) pela frequência da mesma palavra no conjunto de dados composto por todas as grandes áreas (ou áreas ou subáreas). Por exemplo, se a palavra "bioinformatics" tem frequência 0,001 no conjunto de dados de "Ciências Biológicas", frequência 0,0002 em "Ciências Exatas e da Terra" e frequência 0,0001 no conjunto formado por todas as

grandes áreas então esta palavra contará 10 “pontos” (0,001/0,0001) a favor de “Ciências Biológicas” e 2 “pontos” a favor de “Ciências Exatas e da Terra” (0,0002/0,0001). Para cada título foi atribuída a grande área (ou área ou subárea) com maior pontuação. Para cada pesquisador foi criado um vetor de características cujo tamanho é a quantidade de grandes áreas (ou áreas ou subáreas) e em cada posição do vetor foi colocada a quantidade de seus artigos atribuída a cada uma das áreas dividida pelo número total de artigos deste pesquisador (de forma a se obter uma medida normalizada). Vale a pena destacar que foi feito um pré-processamento nos títulos dos artigos composto por duas etapas: primeiro foram excluídas todas as *stop-words* (neste trabalho foram utilizadas *stop-words* em português e inglês por serem as duas línguas mais utilizadas nas publicações realizadas por pesquisadores cadastrados na Plataforma Lattes; em seguida foi executado um *stemmer* para se usar apenas a raiz das palavras. A implementação código livre em Java do algoritmo *Lovins Stemmer* [Lovins 1968] foi utilizada.

As duas características extraídas utilizando-se métricas da **Análise de Redes Sociais** foram a porcentagem dos vizinhos pertencentes a cada grande área (ou área ou subárea) utilizando-se o primeiro nível de vizinhos (apenas vizinhos diretos, no caso, coautores) e utilizando-se a vizinhança nível dois (vizinhos e vizinhos dos vizinhos). Tanto para a vizinhança nível um quanto vizinhança nível dois foram obtidos vetores de caraterística onde para cada grande área (ou área ou subárea) é colocada a porcentagem dos vizinhos que pertence a aquela grande área (ou área ou subárea).

Execução dos experimentos. Foram realizados diferentes experimentos variando-se o período dos dados de treinamento (de um a dez anos) e dos dados de teste (também de um a dez anos). Além de se avaliar o uso de cada uma das características também foram executados experimentos com todas as combinações das três características.

Análise dos resultados. Os resultados foram analisados considerando-se as taxas de acerto de cada experimento. A próxima seção contém a apresentação e a discussão dos resultados.

3. Apresentação e Discussão dos Resultados

Nesta seção são apresentados e discutidos os resultados da identificação das áreas de atuação dos pesquisadores de acordo com o título de suas publicações e de suas redes de coautoria. Os resultados estão divididos em três: **Grandes Áreas**, **Áreas** e **Subáreas**.

3.1. Grandes Áreas

Foram identificados 9.748 pesquisadores possuidores de bolsa produtividade que haviam cadastrado apenas uma grande área de atuação em seus Currículos Lattes. A Figura 1 contém a distribuição destes pesquisadores de acordo com a grande área. Ao todo, de 2001 a 2010 estes pesquisadores publicaram 300.756 artigos em periódicos e seus títulos contém 2.660.772 palavras (após a remoção das *stop words*), sendo que foram identificadas 101.211 “palavras” diferentes (após a execução do *stemmer*).

Devido ao fato de existirem oito grandes áreas, um classificador aleatório teria 12,5% de chance de acertar a classificação de um pesquisador arbitrário, porém, pelo

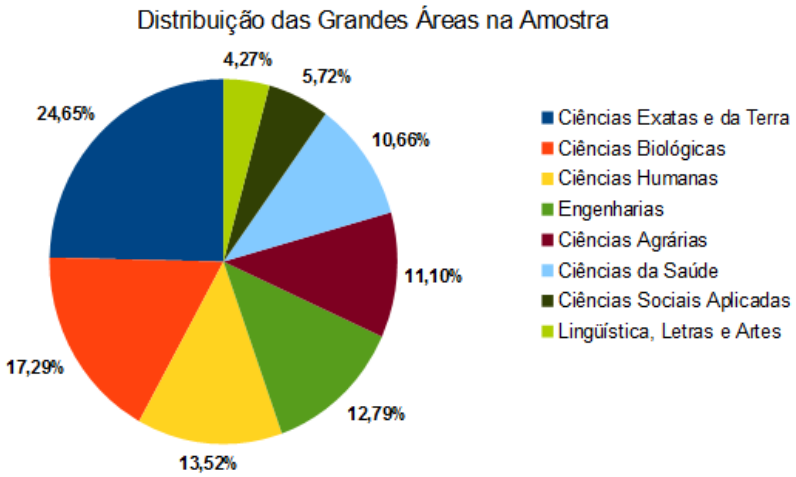

Figura 1. Distribuição das Grandes Áreas na Amostra

fato da distribuição não ser homogênea, uma classificador que classificasse todos os pesquisadores como atuantes em "Ciências Exatas e da Terra" teria 24,65% de acertar a classificação. A seguir serão apresentados os resultados utilizando-se Mineração de Textos (MT) e Análise de Redes Sociais (ARS).

A Tabela 1 contém os resultados do uso da mineração de textos para classificar os pesquisadores de acordo com a classificação do título de seus artigos. Foram utilizados diferentes períodos para a formação dos conjuntos de treinamento (composto por 90% dos pesquisadores) e de testes (composto por 10% dos pesquisadores). Cada linha da tabela corresponde a uma quantidade de anos usados no treinamento. Por exemplo, a linha iniciada por "5 anos" indica que o conjunto de treinamento contém os artigos publicados no período de 5 anos (de 2006 a 2010). A última linha utilizou para treinamento as publicações de 2001 a 2010 (10 anos). Já as colunas indicam o período utilizado nos testes, por exemplo, a coluna iniciada por "2 anos" indica que os dados de teste utilizaram os títulos dos pesquisadores de 2 anos (2009 e 2010), e assim por diante. É possível observar que os resultados melhoram progressivamente ano após ano tanto para o período dos dados de teste quanto treinamento. A exceção ocorre na última linha e em alguns resultados da última coluna, sugerindo que utilizar dados de mais de 9 anos pode atrapalhar a classificação. O melhor resultado foi de 86,67% de acerto e ocorreu utilizando-se as publicações de 9 anos (de 2002 a 2010) tanto para o conjunto de treinamento quanto para o conjunto de testes.

Duas medidas baseadas em ARS foram utilizadas para identificação da grande área de um pesquisador: a distribuição de seus vizinhos em cada uma das grandes áreas, utilizando-se vizinhança nível um (V1), isto é, apenas seus coautores e essa distribuição utilizando-se vizinhança nível dois (V2), isto é, seus vizinhos e os vizinhos de seus vizinhos. Para extrair essas medidas, dez grafos de coautoria foram criados: contendo as publicações apenas de 2010, de 2009 e 2010 e assim por diante. A Figura 2 contém o grafo de coautorias ocorridas entre 2001 e 2010. Os nós estão coloridos de acordo com a grande área de atuação dos docentes. Arestas entre nós da mesma cor são coloridas enquanto que arestas entre nós de diferentes cores aparecem em cinza. É possível notar os agrupamentos dos pesquisadores de uma mesma grande área, isto é, pesquisadores

Tabela 1. Taxas de Acerto utilizando Mineração de Textos - Grandes Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
1 ano	66,15%	75,79%	77,95%	80,41%	81,13%	82,36%	82,67%	82,77%	82,77%	83,08%
2 anos	66,87%	76,92%	79,59%	81,74%	81,85%	82,56%	82,97%	83,79%	84,21%	84,72%
3 anos	66,87%	76,92%	80,51%	81,54%	82,15%	82,46%	82,77%	82,87%	83,69%	83,49%
4 anos	66,56%	76,31%	80,41%	81,23%	82,87%	82,36%	82,67%	83,69%	84,31%	84,41%
5 anos	65,95%	76,72%	80,92%	81,64%	83,49%	83,18%	83,59%	84,72%	85,23%	85,03%
6 anos	66,26%	76,92%	80,72%	81,74%	83,59%	84,00%	83,49%	84,10%	84,41%	84,82%
7 anos	66,36%	77,03%	80,21%	81,03%	82,87%	84,10%	84,82%	85,23%	85,54%	85,33%
8 anos	66,77%	77,85%	80,82%	81,44%	83,38%	84,82%	85,13%	85,33%	86,15%	85,85%
9 anos	66,97%	77,85%	80,31%	81,23%	83,28%	85,03%	85,64%	85,95%	86,67%	86,46%
10 anos	67,08%	77,23%	80,62%	81,44%	83,18%	84,51%	84,92%	85,03%	86,36%	86,05%

tendem a colaborar com outros pesquisadores que atuem na mesma grande área.

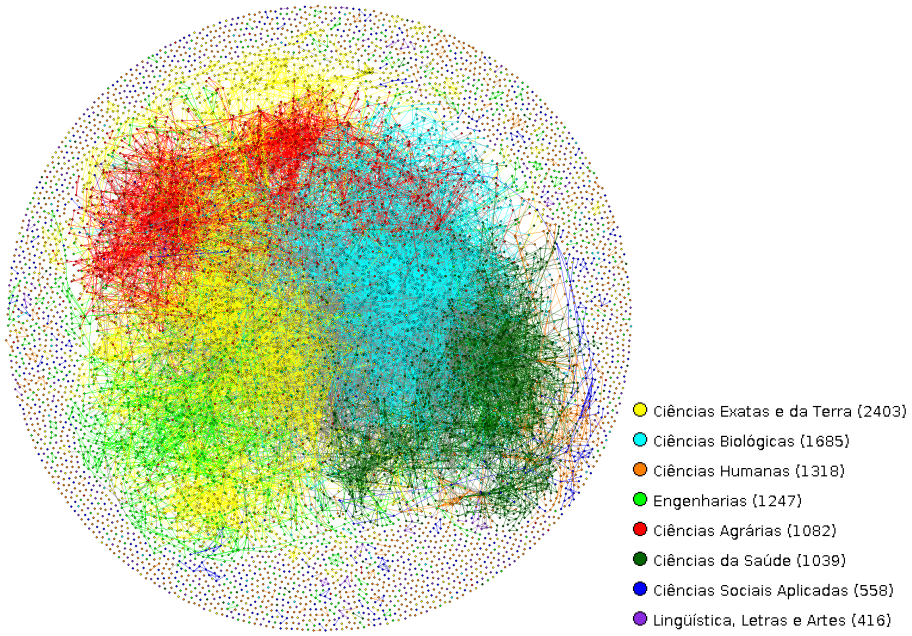

Figura 2. Rede de Coautorias - Publicações de 2001 a 2010 - Grandes Áreas

Outra informação importante da Figura 3 é que há nós que não estão ligados a nenhum outro (isto é, no período considerado eles não publicaram artigos em periódicos em coautoria com nenhum dos elementos da rede), o que inviabiliza a identificação das áreas de atuação destes pesquisadores utilizando-se a ARS. De fato, isto também ocorreu com o uso de MT, mas neste caso apenas quando o pesquisador não publicou nenhum artigo em periódico (independentemente de ser em coautoria ou não com os demais pesquisadores da rede). A Tabela 2 contém a porcentagem dos pesquisadores que não puderam ser classificados utilizando cada técnica. Nas taxas de acerto apresentadas neste artigo, os pesquisadores que não puderam ser classificados entraram na conta das classificações incorretas (já que não foi possível classificá-los corretamente).

É possível observar pela Tabela 2 que a técnica de MT é muito mais inclusiva do que as de ARS utilizadas pois a técnica de MT utiliza todos os artigos publicados no período, por outro lado, as técnicas de ARS só consideram as arestas das redes (ou seja, os artigos publicados em coautoria com outro pesquisador da rede).

Tabela 2. Pesquisadores que não puderam ser classificados - Grandes Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	13,03%	3,79%	1,44%	0,51%	0,41%	0,21%	0,21%	0,21%	0,21%	0,10%
V1 e V2	48,10%	34,46%	29,33%	26,05%	24,10%	22,77%	21,03%	20,10%	19,49%	18,77%

As taxas de acerto do uso de ARS são apresentadas na Tabela 3 juntamente com os resultados da MT e de todas as combinações das três técnicas. Para cada combinação, utilizou-se a soma simples dos valores dos vetores de características, e cada pesquisador foi classificado de acordo com a grande área cujo valor apresentava a maior soma. As colunas correspondem ao período utilizado para treinamento e teste. As linhas apresentam as técnicas (ou combinações) utilizadas.

Tabela 3. Resultados da combinação das técnicas para Grandes Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	66,15%	76,92%	80,51%	81,23%	83,49%	84,00%	84,82%	85,33%	86,67%	86,05%
V1	45,44%	57,44%	62,46%	66,77%	68,31%	69,23%	70,67%	71,69%	72,31%	73,03%
V2	48,10%	59,49%	64,21%	67,59%	69,23%	69,64%	70,67%	71,18%	71,59%	71,90%
MT+V1+V2	70,67%	80,41%	85,44%	85,85%	87,69%	88,41%	88,82%	89,03%	89,85%	89,54%
MT+V1	70,26%	80,21%	85,85%	86,05%	88,00%	88,92%	88,72%	89,23%	89,85%	89,44%
MT+V2	72,00%	81,85%	86,36%	87,08%	88,31%	89,03%	89,33%	89,95%	90,56%	90,26%
V1+V2	46,77%	58,56%	63,59%	67,49%	68,92%	69,44%	70,97%	71,79%	72,41%	72,92%

É possível observar que, individualmente, a técnica baseada em MT obteve melhores resultados do que as técnicas baseadas em ARS considerado os dez períodos de treinamento. Por outro lado, a combinação de MT com V2 (vizinhança de nível dois) obteve melhores resultados do que todas as demais combinações (incluindo MT sozinha e a combinação das três técnicas), atingindo uma taxa de acerto de 90,56% para o período de treinamento e testes de 9 anos. Apenas para ilustrar, a Tabela 4 contém a matriz de confusão usando MT + V2 para o período de 10 anos (cuja taxa de acerto foi de 90,26%).

Tabela 4. Matriz de Confusão - Resultados utilizando MT combinada com V2

	Ciências Agrárias	Ciências Biológicas	Ciências Exatas e da Terra	Ciências Humanas	Ciências Sociais Aplicadas	Ciências da Saúde	Engenharias	Lingüística, Letras e Artes	Não Classificado	Total
Ciências Agrárias	96	6	0	0	0	0	0	0	0	**102**
Ciências Biológicas	3	172	1	0	0	4	1	0	0	**181**
Ciências Exatas e da Terra	1	2	216	0	0	2	6	0	0	**227**
Ciências Humanas	0	2	1	110	9	4	0	10	0	**136**
Ciências Sociais Aplicadas	0	0	0	5	41	0	1	2	0	**49**
Ciências da Saúde	0	10	0	5	0	88	0	0	0	**103**
Engenharias	3	1	7	1	6	0	115	0	0	**133**
Lingüística, Letras e Artes	0	0	0	0	1	0	0	42	1	**44**
Total	**103**	**193**	**225**	**121**	**57**	**98**	**123**	**54**	**1**	**975**

Alternativamente a combinação simples das técnicas (através da soma dos vetores de característica) também foi utilizado o algoritmo de inteligência artificial *Rotation*

Forest (selecionado após a realização de alguns testes com diversos algoritmos). Este algoritmo foi capaz de obter resultados levemente melhores do que a combinação simples, sendo que o melhor resultado ocorreu para o período de 10 anos e a taxa de acerto foi de 90,87%. A Tabela 5 contém a matriz de confusão produzida.

Tabela 5. Matriz de Confusão - Resultados utilizando Rotation Forest

	Ciências Agrárias	Ciências Biológicas	Ciências Exatas e da Terra	Ciências Humanas	Ciências Sociais Aplicadas	Ciências da Saúde	Engenharias	Lingüística, Letras e Artes	Total
Ciências Agrárias	95	5	0	0	0	0	2	0	102
Ciências Biológicas	6	167	1	0	0	6	1	0	181
Ciências Exatas e da Terra	0	2	216	0	0	1	8	0	227
Ciências Humanas	0	2	1	121	1	4	0	7	136
Ciências Sociais Aplicadas	0	1	0	9	34	0	2	3	49
Ciências da Saúde	1	8	0	6	0	88	0	0	103
Engenharias	2	0	4	1	1	0	125	0	133
Lingüística, Letras e Artes	0	0	0	3	2	0	0	39	44
Total	104	185	222	140	38	99	138	49	975

3.2. Áreas

Da amostra utilizada neste artigo foram identificados 7.297 pesquisadores que declaram atuar em apenas uma área, dentre as 76 áreas selecionadas por eles. A Figura 3.2 contém a distribuição dos pesquisadores por área. Nesta figura são apresentadas apenas as 25 áreas com maior número de pesquisadores.

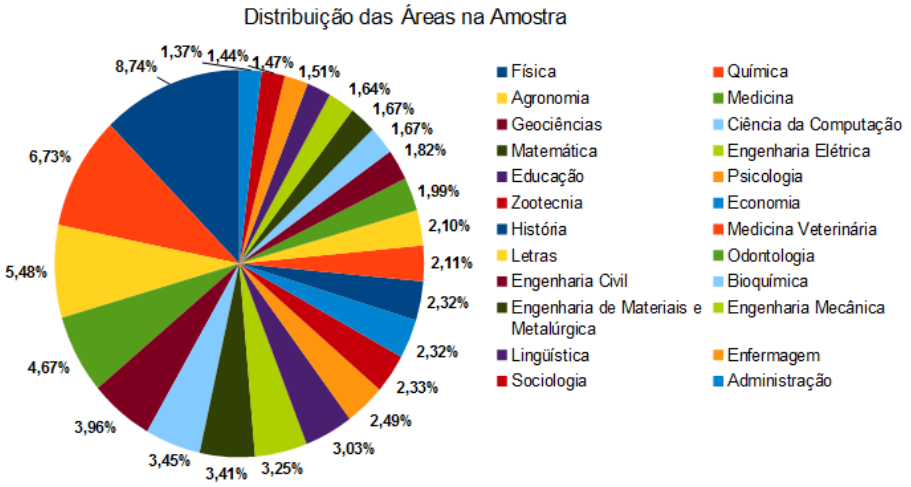

Figura 3. Distribuição das Principais Áreas na Amostra

A Figura 4 contém os grafos das coautorias ocorridas entre 2001 e 2010. A legenda apresenta apenas o nome das áreas que possuem 100 ou mais pesquisadores. Neste grafo é possível notar o agrupamento de pesquisadores de mesma área de atuação, porém, devido à existência de 76 áreas, há diversas arestas cinza (que são aquelas entre pesquisadores de diferentes áreas).

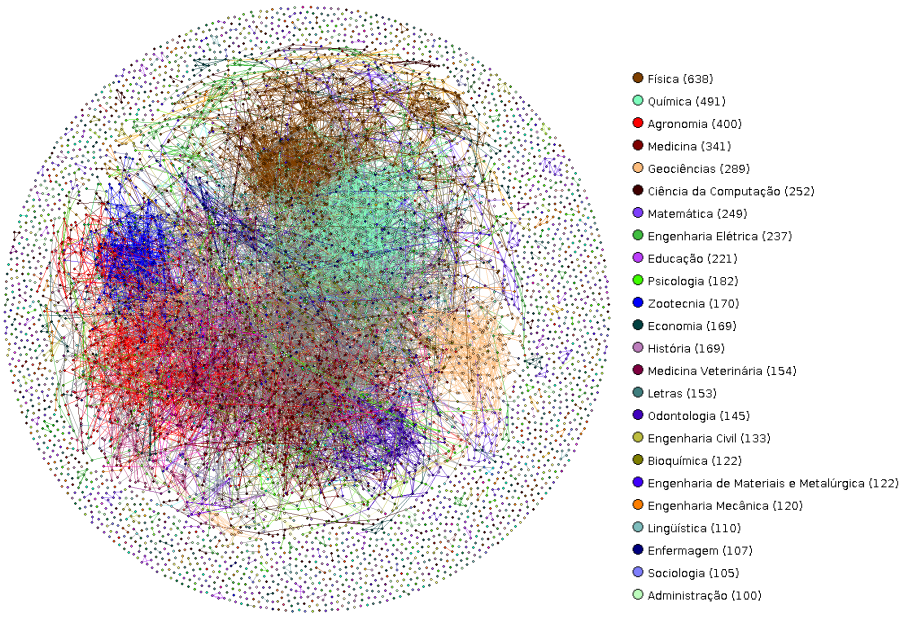

Figura 4. Rede de Coautorias - Publicações de 2001 a 2010 - Áreas

A Tabela 6 contém a porcentagem de pesquisadores que não puderam ser classificados (por não terem nenhuma publicação no período ou por não terem nenhuma publicação em coautoria com outro pesquisador da rede no período).

Tabela 6. Pesquisadores que não puderam ser classificados - Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	13,42%	3,42%	1,37%	0,55%	0,41%	0,27%	0,27%	0,27%	0,27%	0,14%
V1 e V2	54,79%	41,51%	35,75%	32,05%	29,73%	28,49%	26,03%	24,93%	24,38%	23,84%

Assim como feito para as grandes áreas, na análise das áreas foram executados testes variando o período (com dados de 1 a 10 anos). A Tabela 7 contém as taxas de acerto das três técnicas utilizadas, bem como das combinações das mesmas.

Ao contrário do que ocorreu com as grandes áreas, a técnica baseada em MT não obteve os melhores resultados em todos os períodos. Nos três primeiros períodos a técnica V2 obteve melhores resultados e só então a técnica MT superou esses resultados (quando consideradas as técnicas individualmente). Isto provavelmente ocorreu porque com poucos anos o conjunto de dados de treinamento era relativamente pequeno (considerando a existência de 76 áreas), mostrando que esta técnica é mais sensível a quantidade de dados do que as técnicas utilizadas baseadas em ARS.

Assim como ocorreu com as grandes áreas, os melhores resultados foram obtidos combinando-se MT e V2. A maior taxa de acerto foi de 84,11% e ocorreu utilizando-se o período de 10 anos. Para a identificação das áreas também foi utilizado o algoritmo *Rotation Forest* porém seus resultados foram piores do que a combinação das técnicas, sendo que sua maior taxa de acerto foi de 70% para o período de 10 anos.

3.3. Subáreas

Da amostra utilizada neste artigo foram identificados 3.427 pesquisadores que declaram atuar em apenas uma subárea, dentre as 443 subáreas selecionadas por eles. A Figura 5

Tabela 7. Resultados da combinação das técnicas para Áreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	36,58%	48,49%	53,56%	59,45%	63,15%	64,11%	65,62%	66,99%	67,26%	68,22%
V1	37,12%	46,30%	51,78%	55,34%	57,67%	58,77%	61,10%	61,51%	61,51%	61,64%
V2	40,00%	49,73%	55,34%	58,49%	60,55%	61,10%	62,88%	63,42%	63,84%	63,70%
MT+V1+V2	51,64%	62,05%	66,85%	73,70%	76,85%	78,22%	79,45%	79,18%	80,14%	80,96%
MT+V1	50,00%	61,23%	66,58%	72,47%	76,03%	76,58%	78,49%	77,81%	78,49%	79,18%
MT+V2	52,33%	64,38%	68,63%	75,07%	78,77%	80,27%	81,92%	81,78%	83,29%	84,11%
V1+V2	37,95%	46,71%	52,60%	56,99%	60,00%	60,68%	62,05%	62,88%	63,01%	63,01%

contém a distribuição dos pesquisadores por subárea. Nesta figura são apresentadas apenas as subáreas com maior número de pesquisadores.

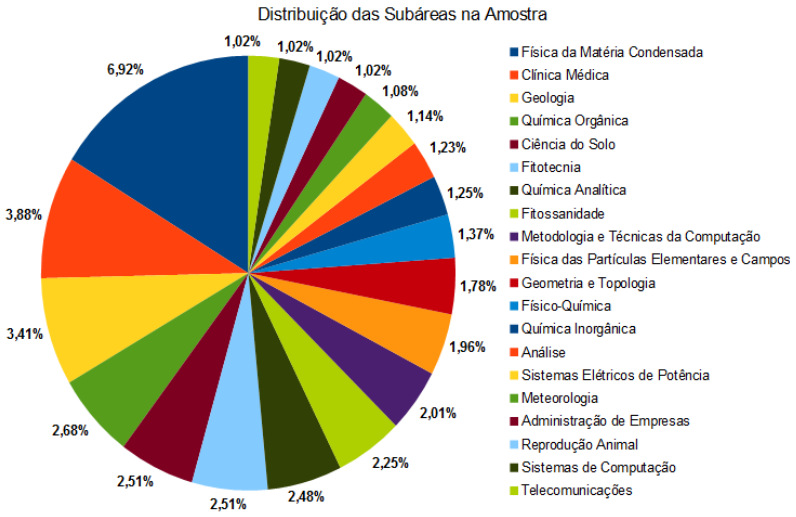

Figura 5. Distribuição das Subáreas na Amostra

A Figura 6 contém o grafo das coautorias ocorridas entre 2001 e 2010. A legenda apresenta apenas o nome das áreas que possuem mais de 35 mais pesquisadores. Assim como nas outras redes de coautoria, é possível notar certo agrupamento entre os pesquisadores da mesma subárea (nós com a mesma cor).

A Tabela 8 contém a porcentagem de pesquisadores que não puderam ser classificados (por não terem nenhuma publicação no período ou por não terem nenhuma publicação em coautoria com outro pesquisador da rede). Para as métricas de ARS, mais de metade dos pesquisadores não poderiam ser classificados com os dados dos dois primeiros anos e, mesmo considerando o período de 10 anos, 30% dos pesquisadores não podem ser classificados por não possuir uma coautoria com outro pesquisador da rede.

Tabela 8. Pesquisadores que não puderam ser classificados - Subáreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	11,66%	3,21%	1,17%	0,87%	0,58%	0,58%	5,25%	0,58%	0,58%	0,29%
V1 e V2	64,14%	53,35%	45,77%	39,94%	38,48%	35,86%	33,24%	31,49%	30,61%	30,03%

A Tabela 9 contém as taxas de acerto das três técnicas utilizadas, bem como das combinações das técnicas variando o período de 1 a 10 anos.

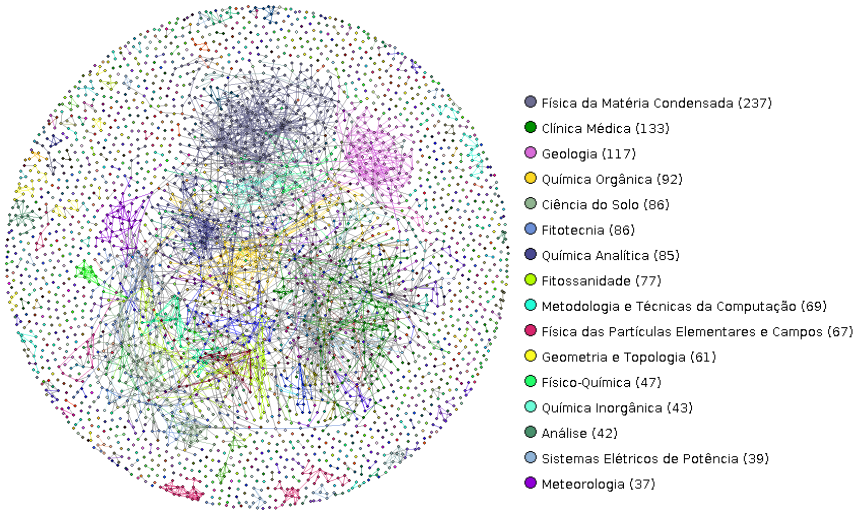

Figura 6. Rede de Coautorias - Publicações de 2001 a 2010 - Subáreas

A técnica baseada em MT teve os piores resultados em todos os períodos. Conforme constatado para as áreas, está técnica é mais sensível ao volume de dados então sua piora de desempenho foi causada pela diminuição no número de pesquisadores para o treinamento e aumento no número de classes (no caso, as 443 subáreas). A técnica V2 foi novamente melhor do que a V1para todos os períodos, além disso, a combinação entre MT e V2 apresentou os melhores resultados, destacando-se as taxas de acerto de 59,77% ocorridas para os períodos de 9 e 10 anos.

Tabela 9. Resultados da combinação das técnicas para Subáreas

	1 ano	2 anos	3 anos	4 anos	5 anos	6 anos	7 anos	8 anos	9 anos	10 anos
MT	10,79%	14,87%	17,20%	20,41%	20,70%	23,03%	23,62%	26,53%	26,53%	26,24%
V1	26,53%	32,36%	38,78%	43,44%	44,31%	46,94%	48,69%	48,10%	48,10%	47,81%
V2	30,32%	37,61%	45,48%	48,98%	50,44%	52,19%	53,64%	54,81%	55,98%	56,27%
MT+V1+V2	31,20%	38,78%	43,44%	48,69%	50,44%	53,35%	55,39%	57,14%	55,98%	55,69%
MT+V1	29,15%	37,32%	41,98%	46,06%	47,52%	50,15%	52,48%	53,64%	52,19%	51,90%
MT+V2	32,07%	41,11%	46,65%	51,31%	53,94%	55,39%	58,02%	59,48%	59,77%	59,77%
V1+V2	27,11%	33,82%	39,36%	44,61%	46,06%	48,10%	49,85%	50,44%	50,15%	50,44%

Vale ressaltar que um classificador aleatório acertaria 0,226% (considerando as 443 subáreas) e um classificador que classificasse todos os pesquisadores como pertencentes à subárea mais frequente (no caso, “Física da Matéria Condensada”) acertaria 6,92%. Assim, os resultados obtidos de cerca de 60% de acertos são bastante satisfatórios. O algoritmo *Rotation Forest* aplicado sobre o conjunto de dados das subáreas obteve uma taxa máxima de acerto de 36,15% e esta taxa foi obtida usando-se os dados dos 10 anos.

4. Conclusões e Trabalhos Futuros

Este artigo apresentou uma técnica que combina a mineração de textos com a análise de redes sociais a fim de se identificar automaticamente a área de atuação de um pesquisador. A técnica baseada em mineração de textos utilizou os títulos dos artigos publicados pelos pesquisadores enquanto que as técnicas baseadas em redes sociais utilizaram as áreas de atuação dos vizinhos e dos vizinhos dos vizinhos.

Apesar das técnicas empregadas serem relativamente simples, os resultados obtidos foram bastante satisfatórios, atingindo taxas de acerto superiores a 90% para a identificação das grandes áreas (dentre as 8 disponíveis); superiores a 84% para a identificação de áreas (dentre 76); e de 59,77% para a identificação de subáreas (dentre 443). Verificou-se que a combinação simples da técnica de mineração de textos utilizada com a análise da vizinhança de nível dois obteve melhores resultados do que aqueles produzidos pelas técnicas usadas individualmente. Para o uso de mineração de texto, observou-se que usar dados com mais de 9 anos para a identificação das grandes áreas resultou, em muitos casos, na diminuição da taxa de acerto. Este fenômeno precisa ser melhor investigado ampliando-se o tamanho da amostra e o período das publicações.

Os trabalhos futuros seguirão diferentes linhas. Na primeira, pretende-se avaliar outras métricas de redes sociais bem como outras características da mineração de texto de forma a se obter atributos adicionais para a identificação de áreas. Uma segunda linha de pesquisa é utilizar outras técnicas estatísticas ou de inteligência artificial para combinar os atributos obtidos de forma a melhorar as taxas de acertos na identificação de áreas. Na terceira, pretende-se explorar a janela de tempo utilizada na identificação de áreas de forma a se estabelecer qual janela de tempo é mais recomendada para o conjunto de dados e de treinamento. Além disso, pretende-se explorar a característica hierárquica das áreas de atuação, desenvolvendo um classificador em níveis. Por fim, pretende-se utilizar a identificação de áreas desenvolvidas como base para resolver outros problemas como a identificação de *experts*, detecção de tendências e avaliação de grupos interdisciplinares.

Agradecimentos

Este trabalho foi parcialmente financiado pela FAPESP (Projeto Jovem Pesquisador processo 2009/10413-5) e pelo CNPq (Produtividade em Pesquisa processo 304937/2010-0).

Referências

Digiampietri, L., Mena-Chalco, J., Silva, G. S., Oliveira, L., Malheiro, A., and Meira, D. (2012). Dinâmica das Relações de Coautoria nos Programas de Pós-Graduação em Computação no Brasil. In *CSBC 2012 - BraSNAM*.

Gerdsri, N., Kongthon, A., and Puengrusme, S. (2012). Discovering the professional communities and social networks of emerging research areas: Use of technology intelligence from bibliometric and text mining analysis. In *Technology Management for Emerging Technologies (PICMET 2012)*, pages 114–121.

Gharehchopogh, F. S. and Khalifelu, Z. A. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *5th Int. Conf. on Application of Information and Communication Technologies (AICT'2011)*, pages 1–4.

Lovins, J. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Wang, T. and Krim, H. (2012). Statistical classification of social networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3977–3980.

Wang, T., Krim, H., and Viniotis, Y. (2013). A generalized markov graph model: Application to social network analysis. *Selected Topics in Signal Processing, IEEE Journal of*, 7(2):318–332.