Análise de Redes Sociais de Pesquisadores Cadastrados na Plataforma Lattes

Rodrigo A. de Freitas Vieira, Luciano Antonio Digiampietri

Universidade de São Paulo/Escola de Artes, Ciências e Humanidades

rodrigo.vieira@usp.br, digiampietri@usp.br

Resumo

A partir da base de dados brasileira sobre pesquisadores e suas pesquisas da Plataforma Lattes, diversas informações referentes a publicações, produções técnicas e projetos de pesquisa são encontradas e podem ser utilizados para a identificação de diversos tipos de relações entre pesquisadores possibilitando assim a criação de redes sociais acadêmicas. Essas redes sociais podem ser utilizadas para um entendimento profundo da pesquisa Brasileira bem como para identificar padrões sobre os pesquisadores e suas pesquisas. O objetivo deste projeto é desenvolver ferramentas que calculem diferentes métricas individuais e estruturais de redes de pesquisadores formadas a partir de dados da Plataforma Lattes buscando modelar e prever citações de artigos presentes na base de dados, trabalhando com: especificação e desenvolvimento de ferramentas (*crawlers*) para automatizar a busca por informações para enriquecer o banco de dados; uso de modelos estatísticos para tentar modelar o comportamento das citações; estudo e uso de técnicas de inteligência artificial de agrupamento, classificação e predição para modelar o comportamento das citações de acordo com classes de artigos; verificação e validação das técnicas utilizadas para a modelagem e predição das citações.

Palavras Chaves: Redes Sociais, Plataforma Lattes, Predição, Citação

Abstract

From the Brazilian database of researchers and their researches, the Lattes Platform, several information regarding publications, technical productions and research projects are found and can be used to identify several types of relations among researchers thus enabling the creation of an academic social network. This social network can be used to deeply understand the Brazilian researches and to identify patters between researchers and their work. The objective of this project is to develop tools that calculate different individual and structural metrics of researchers networks formed from data of the Lattes Platform trying to model and predict citations of articles present in the database, working with: specification and development of tools (*crawlers*) to automate the search for information to enrich the database; use of statistical models to try to model the citations; study and use of artificial intelligence techniques for clustering, classification and prediction to model the behavior of citations according to classes of article; verification and validation of the techniques used for modeling and prediction of citations.

Key words: Social Networks, Lattes Platform, Prediction, Citation

Introdução

Atualmente, uma grande quantidade de informações referentes à produção científica está disponível na Web, como publicações científicas, informações sobre projetos de pesquisa e mesmo currículos de pesquisadores.

Ao se tratar de dados referentes à pesquisa, o Brasil apresenta uma característica peculiar: a existência de um cadastro nacional de currículos de pesquisadores, os currículos da Plataforma Lattes, que congregam informações sobre publicações, orientações, projetos de pesquisa, entre outras. Esse grande volume de informações ainda não foi amplamente usado e, tipicamente, é consultado para avaliar (ou verificar dados) de pesquisadores individualmente ou docentes de programas de pós-graduação.

Este projeto visa a estender trabalhos prévios (Alcázar et al, 2011; Digiampietri et al, 2011; Digiampietri et al 2012) sobre a extração e organização de informações de Currículos Lattes por meio do uso e implementação de diferentes técnicas para a predição de citações em artigos científicos.

Objetivos

Este projeto de iniciação científica visa a analisar o comportamento das citações de artigos científicos relacionados à Rede Social Acadêmica formada a partir de dados da Plataforma Lattes e enriquecida com informações de outras bases (como número de citações de cada artigo).

O projeto foca em duas importantes características: a utilização de diferentes técnicas para modelar as citações dos artigos de forma a verificar as mais adequadas; e com base na modelagem realizada, desenvolver e testar algoritmos de predição de citações.

Materiais e Métodos

Como amostra para este projeto, optou-se por utilizar as publicações dos pesquisadores brasileiros possuidores de bolsa produtividade em pesquisa do CNPq (cerca de 14.000 pesquisadores). A identificação destes bolsistas foi feita utilizando-se dados de seus currículos Lattes. Então, a partir do currículo de cada pesquisador, foi obtida a lista de suas publicações e as citações históricas de cada uma das publicações pelo *Google Scholar*. Apesar desta fonte de dados não ser curada ela é uma das poucas fontes de informação que disponibiliza publicamente informações sobre as citações ano a ano dos artigos indexados no sistema da Google.

O projeto foi iniciado com a leitura de uma bibliografia já produzida pelo orientador e o grupo de pesquisa em projetos anteriores, e logo após, iniciou-se um estudo prévio do comportamento das citações, o que levantou a hipótese de que seriam necessários diferentes modelos de regressão para a predição. Uma das maneiras de segregar o conjunto de artigos é por meio da clusterização (agrupamento) e, então, os artigos presentes em cada cluster poderiam ser modelados de acordo com o modelo mais adequado a este cluster. Para isso, foi feita a clusterização dos dados, utilizando-se o algoritmo *SimpleKMeans* disponíveis no arcabouço Weka que resultou na geração de 6 *clusters* (grupos).

Para a separação dos artigos atribuindo-os a cada cluster, foi desenvolvida uma classe em Java utilizando a API de desenvolvimento do Weka. Pelo fato dos testes utilizarem um conjunto de dados grande e de que seriam testados diversos modelos de classificação, optou-se por desenvolver, também, um algoritmo para automatizar a realização destes testes.

Foram testados os algoritmos de classificação disponíveis no Weka que usam algum tipo de regressão e podem ser utilizados para a predição de valor e com o objetivo de: dadas as citações dos 19 primeiros anos de publicação de cada artigo tentar estimar quantas citações este artigo teria em seu vigésimo ano de publicação. Para os testes e validação foi utilizado o método de *10-fold cross-validation*. Esta bateria de testes foi utilizada para tentar prever a quantidade de citações dos artigos em seu vigésimo ano de publicação, sem antes se ter clusterizado os artigos.

Após a clusterização dos dados e a execução de testes de classificação, iniciou-se a combinação das atividades visando a obter predições mais precisas. Os testes foram realizados sobre um mesmo conjunto de dados formatado de dois modos distintos, o conjunto possui 3.213 artigos com no mínimo 20 citações e publicados há ao menos 20 anos.

Na primeira formatação do conjunto, os anos de publicação foram numerados de 0 a 19, sendo 0 o ano em que o artigo foi publicado e, os dados foram normalizados de forma que a cada um dos 20 primeiros anos de publicação de cada artigo fosse associada a porcentagem acumulada de citações recebidas por ele no período até o ano em questão, tendo todos os artigos o valor 1 associado ao vigésimo ano, correspondendo a 100% das citações no período.

A outra formatação utilizava apenas uma parcela dos anos de publicação, sendo numerado de 0 a 11, e também era normalizado de forma que cada um dos 12 primeiros anos fosse associado à porcentagem acumulada de citações recebidas por ele no período até o ano em questão, entretanto o valor 1, referente a 100% das citações do período, era atribuído ao décimo primeiro ano tendo o décimo segundo ano valores superiores a 1.

Utilizando dos algoritmos desenvolvidos anteriormente, os conjuntos foram clusterizados e classificados utilizando 9 modelos de classificação diferentes aplicados ao conjunto total e a cada cluster. Em ambos os conjuntos de dados testados, a previsão fora realizada no décimo segundo ano. No primeiro teste a normalização dos dados compreendia dados futuros àquele que seria previsto (a predição seria do décimo segundo ano conhecendo-se até o vigésimo, por isso era de se esperar um bom resultado [este tipo de algoritmo pode ser utilizado, por exemplo, para preenchimento de informações faltantes em um banco de citações]). Após esse teste, passou-se para os testes finais com o segundo conjunto de dados, normalizados de forma a simular uma predição real, onde não há o conhecimento de dados futuros.

Resultados

A hipótese para a utilização da clusterização era de que a partir dessa segregação dos dados, houvesse uma melhora na média e/ou na mediana da classificação em relação ao total.

Para os testes realizados com os artigos normalizados em 20 anos e o vigésimo ano associado a 100% das citações o resultado no total se mostrou muito bom com um erro absoluto médio (*mean absolute error*) de 0,0454 utilizando o algoritmo *LastMedSq* e um erro quadrático médio (*root mean squared error*) de até 0,0617 (*LinearRegression*). Apesar do bom resultado, após a clusterização dos artigos pôde-se melhorar o desempenho médio das classificações em certas métricas.

Utilizando como métrica de avaliação o erro absoluto médio resultante de cada classificador para cada cluster e comparando-os com o melhor resultado do total, pode-se notar nas tabelas 1 e 2 que diferentes algoritmos seriam mais indicados para certos clusters que outros, obtendo-se assim resultados ainda melhores. Nota-se que o melhor classificador testado no conjunto total é o *LastMedSq*. Esse mesmo classificador utilizado, separadamente, nos clusters 1 e 3, mostrou uma melhora de 33% e 11% respectivamente sobre o resultado do total. Quando testado então esse classificador no cluster 5 houve um decréscimo de 2%, enquanto a regressão *Isotonic Regression* obteve um resultado 49% melhor.

Tabela 1: Comparação do Erro Absoluto Médio dos clusters com o conjunto total do primeiro conjunto

Classificador	Total	Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
GaussianProcesses	0.0464	0.0561	-23.57%	0.0328	27.75%	0.0367	19.16%	0.0403	11.23%	0.0369	18.72%	0.0252	44.49%
IsotonicRegression	0.0458	0.0579	-27.53%	0.0309	31.94%	0.0357	21.37%	0.0439	3.30%	0.0365	19.60%	0.0233	48.68%
LastMedSq	0.0454	0.0562	-23.79%	0.0302	33.48%	0.0378	16.74%	0.0402	11.45%	0.038	16.30%	0.0461	-1.54%
LinearRegression	0.0459	0.0557	-22.69%	0.0304	33.04%	0.0384	15.42%	0.0409	9.91%	0.0377	16.96%	0.0237	47.80%
MultilayerPerception	0.0532	0.0672	-48.02%	0.0374	17.62%	0.047	-3.52%	0.047	-3.52%	0.0436	3.96%	0.0274	39.65%
PaceRegression	0.0459	0.0559	-23.13%	0.0303	33.26%	0.038	16.30%	0.0407	10.35%	0.0376	17.18%	0.0241	46.92%

Analisando-se a média ponderada (pelo número de artigos) do erro absoluto médio referente ao melhor classificador de cada cluster, percebe-se a diminuição do erro em 16,81% sobre o resultado do melhor classificador para o conjunto total, com um resultado de 0,037769.

Para os testes feitos com os artigos normalizados em 11 anos e o décimo ano associado a 100% das citações, o resultado das classificações no total se mostrou bom com um erro absoluto médio de 0,3187 utilizando o algoritmo *SMOreg* e um erro quadrático médio de até 0,6180 (*GaussianProcesses*).

Analisando novamente o erro absoluto médio resultante de cada classificador em cada cluster e comparando-os com o melhor resultado do total, temos mais uma vez diferentes modelos que seriam mais

indicados para certos clusters que outros. O erro absoluto médio desses resultados da clusterização em relação ao total chegou a ser 74,22% menor (de apenas 0,0821 contra 0,3187 para o conjunto total).

Tabela 2: Comparação do Erro Absoluto Médio dos clusters com o conjunto total do segundo conjunto

Classificador	Total	Cluster 0		Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
		Valor	%	Valor	%	Valor	%	Valor	%	Valor	%	Valor	%
GaussianProcesses	0.3228	0.0849	73.35%	0.1435	54.95%	0.3788	-18.88%	0.0947	70.27%	0.2867	10.03%	0.7376	-131.49%
IsotonicRegression	0.3488	0.0856	73.12%	0.1458	54.25%	0.3775	-18.46%	0.1009	68.33%	0.2925	8.21%	0.7424	-132.97%
LastMedSq	0.3519	0.0821	74.22%	0.1387	56.48%	0.3741	-17.42%	0.0865	72.84%	0.2775	12.93%	0.7335	-130.20%
LinearRegression	0.3471	0.0901	71.72%	0.1488	53.32%	0.3794	-19.07%	0.1000	68.62%	0.3020	5.23%	0.7472	-134.49%
MultilayerPerception	0.3971	0.1264	60.33%	0.1978	37.92%	0.4621	-45.03%	0.1294	59.39%	0.3461	-8.60%	0.7693	-141.44%
PaceRegression	0.3466	0.0892	72.02%	0.1511	52.59%	0.3778	-18.58%	0.1008	68.37%	0.3008	5.61%	0.7430	-133.18%
RBFNetwork	0.3734	0.0875	72.55%	0.1476	53.67%	0.3811	-19.60%	0.0968	69.62%	0.2968	6.85%	0.7407	-132.46%
SimpleLinearRegressi	0.3551	0.0882	72.33%	0.1537	51.75%	0.3778	-18.58%	0.0975	69.40%	0.3095	2.87%	0.7441	-133.52%
SMOreg	0.3187	0.0826	74.09%	0.1380	56.69%	0.3676	-15.37%	0.0856	73.15%	0.2747	13.80%	0.7085	-122.34%

Com a média ponderada (pelo número de artigos) do erro absoluto médio referente ao melhor classificador de cada cluster, temos novamente uma diminuição do erro utilizando os melhores classificadores para cada cluster, apresentando uma melhora de 2,63% sobre o resultado do melhor classificador para o conjunto total.

Conclusões

O tratamento dos dados e organização dos testes se mostraram adequados para os testes e análises, e foi possível testar, desenvolver e combinar diferentes estratégias de forma a modelar e prever citações (minimizando-se os erros associados à modelagem ou predição).

Os resultados apresentaram boas previsões iniciais, com um erro absoluto médio de até 0,0454 no primeiro conjunto de dados. Adicionalmente, este resultado foi melhorado após a clusterização dos artigos, com uma otimização do resultado em até 74% sobre o teste original do segundo conjunto de dados, garantido, pela possibilidade de utilização de modelos de regressão diversos. Temos então, resultados que se mostraram eficientes, alcançando as expectativas na qualidade de previsões e confirmando as hipóteses e a metodologia de análise.

Referências Bibliográficas

ALCAZAR, J. J. P. ; ESCALANTE, M. ; KURECK, E. ; DIGIAMPIETRI, L. A. ; TUESTA, E. F. . Avaliação de Redes de Inovação usando uma ferramenta baseada em redes sociais - caso Brasileiro de Nanotecnologia. In: XIV Congresso Latino-Iberoamericano de Gestión Tecnológica (ALTEC 2011), 2011, Lima, Peru. Anais do XIV Congresso Latino-Iberoamericano de Gestión Tecnológica, 2011.

BERKOWITZ, S.D. An Introduction to Structural Analysis: The Network Approach to Social Research. Butterworths, Toronto. 1982.

BREIGER, R.L. The Analysis of Social Networks. Handbook of Data Analysis. Londres, Sage Publications. pp 505–526, 2004.

DIGIAMPIETRI, L. A., DA SILVA, E. E. A Framework for Social Network of Researchers Analysis. Iberoamerican Journal of Applied Computing, v. 1, n. 1, p. 1-24, 2011.

DIGIAMPIETRI, L. A. ; MENA-CHALCO, J. ; ALCAZAR, J. J. P. ; TUESTA, E. F. ; DELGADO, K. V. ; MUGNAINI, R. ; SILVA, G. S. . Minerando e Caracterizando Dados de Currículos Lattes. In: Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012, Curitiba, PR, Brasil. Proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2012.

MENA-CHALCO, J. P.; CESAR JUNIOR, R. M. scriptLattes: an open-source knowledge extraction system from the Lattes platform. Journal of the Brazilian Computer Society, 2009; 15(4):31-39.

ULRIK, B.; ERLEBACH, T. Network Analysis: Methodological Foundations Berlin, Heidelberg: Springer-Verlag. 2005.