Análise de rede de termos em Sistemas Embarcados através de análise da rede de termos em títulos de trabalhos científicos

Jansen Souza, Moacir L. Mendonça Júnior,
Alisson V. Brito e Alexandre N. Duarte

[1]Programa de Pós-Graduação em Informática (PPGI)
Centro de Informática, Universidade Federal da Paraíba
João Pessoa, PB. Brazil

{moacir.lopes.jr,jansen.souza}@gmail.com, {alisson,alexandre}@ci.ufpb.br

Abstract. *This paper presents an analysis of research topics extracting terms of the titles of the papers in Embedded Systems field. For this, a tool has been developed for the extraction, characterization and assembly of the network using the database of IEEE Xplore. In the end, over twenty thousand articles were extracted and analyzed, and the most relevant areas and publications have been identified.*

Resumo. *Este trabalho apresenta uma análise dos temas de pesquisa abordados na área de Sistemas Embarcados com base em uma rede de palavras significativas presentes nos títulos dos artigos. Para tal, foi desenvolvida de uma ferramenta para extração, caracterização e montagem de redes utilizando a base da IEEE Xplore. Ao final, mais de vinte mil artigos foram extraídos e analisados, e as áreas e publicações mais relevantes foram levantadas.*

1. Introdução

Desde a sua origem, o conhecimento vem sendo buscado pelo ser humano de todas as suas formas. O processo social realizado através do trabalho e esforço coletivo permite a construção do conhecimento [Bourdieu 2004].

O isolamento de um pesquisador no mundo acadêmico de hoje já não é mais possível, pois para se ter uma maior integração de elementos no momento de se gerar produções cientificas mais relevantes são necessárias associações, negociações e estratégias [Silva 2002]. Então pode-se dizer que a partir do esforço mútuo dos pesquisadores foi possível impulsionar a produção de conhecimento [Balancieri 2005]. Logo, há um crescente interesse em analisar publicações compartilhadas e a relevância destas colaborações, pela possibilidade de apresentar diferentes qualidades e variadas motivações.

Diante disso, visando compreender o complexo sistema de geração de conhecimento, inúmeras abordagens de análise vêm sendo aplicadas. Entre elas estudos de produtividade científica, que geralmente se baseiam em uma análise quantitativa de citações recebidas somadas a metodologias de análise mais recentes como a Análise de Redes Sociais. Sendo elas capazes de representar a dinamicidade da produção cientifica, possibilitando assim, uma visão abrangente das interações entre as partes [Tomaél and Marteleto 2006].

Segundo [Chakrabarti 2003], Análise de Redes Sociais (ARS) é o ”mapeamento e medição das relações e fluxos entre pessoas, grupos, organizações, computadores, URLs, e outras entidades conectadas de informação/conhecimento”. Esse campo de estudo, que se encontra na interseção entre sociologia e matemática, foca na relação entre os atores ao invés dos atributos de cada ator, como tradicionalmente ocorre em outras abordagens das ciências sociais. Tais redes permitem representar a diversidade social e a complexa formação dos relacionamentos entre indivíduos e os demais componentes de um grupo. Estas são formadas pela composição de nós (representando indivíduos ou organizações) e da interligação por um ou mais tipos específicos de interdependência [Berkowitz 1982]. E, por meio desses relacionamentos, vão construindo e reconstruindo uma estrutura social.

Neste contexto, este trabalho apresenta um estudo com base em uma rede montada com palavras significativas presentes nos títulos dos artigos publicados na área de **Sistemas Embarcados** (SE) encontradas na base de publicações *IEEE Xplore*. Assim, podemos assim empregar essas análises em diversos estudos, como por exemplo, nas revisões sistemáticas de literatura.

O artigo está organizado da seguinte forma: A Seção 2 apresenta os trabalhos relacionados. A Seção 3 descreve a metodologia de pesquisa utilizada. A Seção 4 apresenta os resultados obtidos. O artigo é finalizado na Seção 5, com as conclusões e propostas de trabalhos futuros.

2. Trabalhos Relacionados

Existem consideráveis trabalhos no campo da pesquisa acadêmica. A maior parte deles está centrada na criação de perfis do pesquisador, visando à classificação dos autores. O artigo de Quinkun Zhao et al. [Zhao et al. 2008] é um destes trabalhos, ele estuda o relacionamento entre autores e comunidades usando técnicas de mineração. Outro trabalho é o ArnetMiner [Tang et al. 2008] que realiza a classificação dos autores pelo h-index. Também existem pesquisas que focam na classificação dos artigos acadêmicos mas em sua maioria usam o número de citações como métrica. Entretanto, é um tanto intuitivo que esta métrica ignora a importância da qualidade de citações, levando em consideração somente a quantidade de citações. Nós usamos o algoritmo *PageRank*, que considera os artigos que são mais citados por artigos também muito citados. Encontramos alguns artigos que usam a mesma métrica ou suas modificações para classificar artigos, entre eles podemos citar a pesquisa de Jingyu Cui [Cui et al. 2010] que, baseado em um *PageRank* modificado, realizou a classificação dos artigos e de seus autores a partir de um grafo multi-camada.

As redes semânticas criadas neste trabalho foram construídas baseadas no método desenvolvido por Fadigas et al. [Fadigas et al. 2009] e Cunha et al. [Cunha et al. 2003], onde os autores definiram essas redes a partir de títulos de artigos científicos. O trabalho apresenta todo o processo de tratamento manual das palavras e também o processo de tratamento a partir de ferramentas computacionais, semelhante ao que foi construído no nosso trabalho. A partir desse tratamento, foi construída a rede semântica, analisando, segundo os conceitos de análise de redes sociais, a importância e frequência das palavras nos títulos.

O principal foco desta pesquisa é realizar a classificação de artigos acadêmicos

e suas comunidades, oferecendo um suporte adicional ao pesquisador para criação de definições estratégicas que caracterizem determinadas áreas de pesquisa, ajudando no processo de difusão do conhecimento em campos específicos.

3. Metodologia

Neste estudo fizemos uma análise das publicações sobre sistemas embarcados utilizando métricas de análise de redes sociais com o intuito de identificar quais publicações e meios de publicação são relevantes para a área, como também realizamos a classificação das comunidades identificadas na rede, através das palavras significativas encontradas nos títulos de cada publicação. A Figura 1 ilustra o processo metodológico seguido no experimento, desde o envio do termo de busca, pela ferramenta, até o processo final de classificação das comunidades.

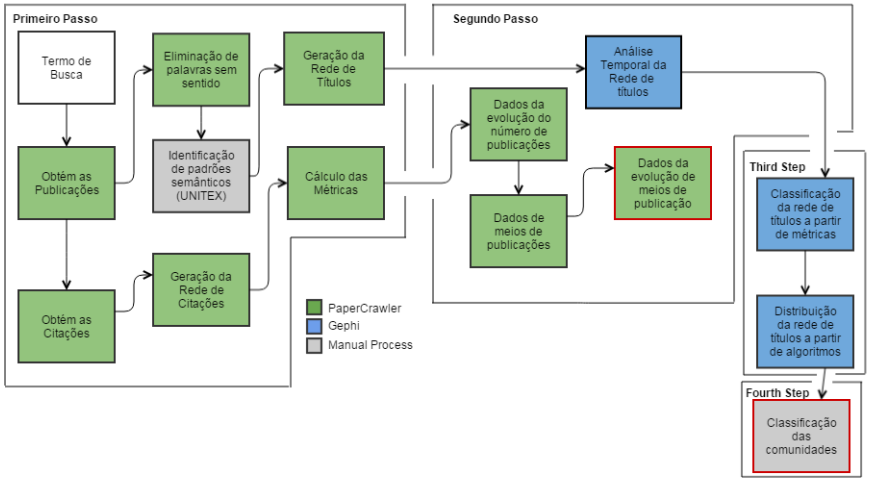

Figura 1. Processo Metodológico

Passo 1: Modelagem da rede - O primeiro passo da metodologia é obter as publicações e suas citações a partir de uma *query* (esta podendo representar uma área de conhecimento, como por exemplo, "Análise de Redes Sociais" ou "Sistemas Embarcados"), com isso pode-se gerar a rede de publicações, de títulos e realizar cálculo das métricas de ARS, como grau, grau de entrada, grau de saída, centralidade de intermediação, modularidade e *Page Rank*, utilizadas para diversos fins neste estudo, entre eles a identificação da relevância e a classificação das comunidades.

Passo 2: Análise temporal da rede de títulos e Análise da rede de citações - O segundo passo é analisar os dados obtidos referentes à área de conhecimento como, os meios de publicações, a evolução das publicações ao longo dos anos, a evolução do número de publicações para os meios de publicações ao longo dos anos, a evolução das comunidades ao longo dos anos, análise temporal dos termos utilizados nos títulos e por último obter as *keywords* de cada comunidade.

Passo 3: Classificação e distribuição da rede - O terceiro passo é obter uma visualização amigável da rede para então analisar a área de conhecimento como um todo.

Passo 4: Definição e classificação de comunidades - E finalmente, identificar as comunidades mais relevantes.

4. Resultados

Esta seção é dedicada à apresentação dos resultados obtidos pela ferramenta *Paper Crawler* para a cadeia de busca *"Embedded Systems"* na base de publicações *IEEE Xplore*, realizado no final do ano de 2014.

4.1. Análise da Rede de Termos

A partir dos resultados obtidos nos últimos trabalhos, resolvemos consolidar esta pesquisa aplicando nossa ferramenta em outro contexto como estudo de caso. Portanto, nosso passo inicial foi obter as publicações para que com a rede gerada pudéssemos realizar a análise da área de sistemas embarcados. Portanto, a ferramenta *Paper Crawler*, através do termo de busca *Embedded Systems*, solicitou os resultados para o repositório acadêmico *IEEE Xplorer*, que nos retornou mais de dez mil títulos de artigos.

No entanto, para alcançarmos um melhor resultado em nossa análise, foram retirados dos títulos encontrados, todas as palavras-chave *embedded* e *system*, por fazerem parte do termo de busca. Destarte, através dos títulos e seus respectivos anos de publicação encontrados pela ferramenta, montamos a rede semântica variável ao longo do tempo, conforme ilustrado na Figura 2, onde a cor representa as comunidades e o tamanho das circunferências o grau.

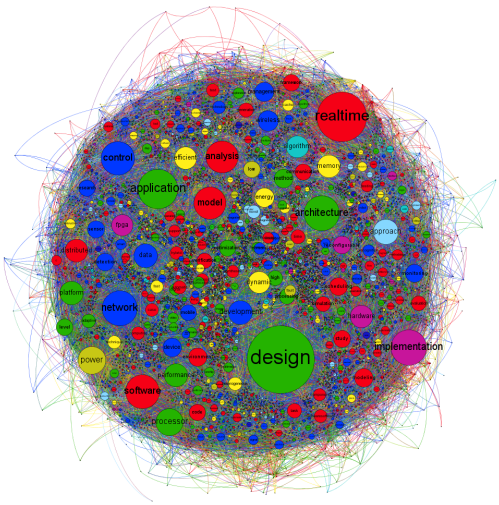

Figura 2. Rede Semântica

A Tabela 1 lista os cinco vértices de maior grau, representando palavras-chave em sua forma normalizada, presentes na rede de títulos dos artigos científicos publicados na área de sistemas embarcados, além do seu coeficiente angular, obtido através de regressão angular, que denota a tendência de crescimento.

Para geração das retas de tendência de cada palavra, foi aplicado a regressão linear e gerada a equação para cada com previsão linear para os próximos cinco anos. Quanto maior o coeficiente angular, maior a tendência de crescimento. É possível observar que os

termos *design* e *realtime* possuem as maiores tendências de crescimento dentre os termos mais relevantes.

Tabela 1. Estatísticas referentes ao grau dos vértices da rede semântica dos títulos

Palavra-chave	Grau	Coef. angular
Design	2661	5,95
Realtime	1950	2,99
Application	1548	2,17
Implementation	1391	1,8
Architecture	1380	1,79

A Figura 3 apresenta um gráfico com a evolução e tendência do termo ”Application”, com coeficiente angular de 2,17.

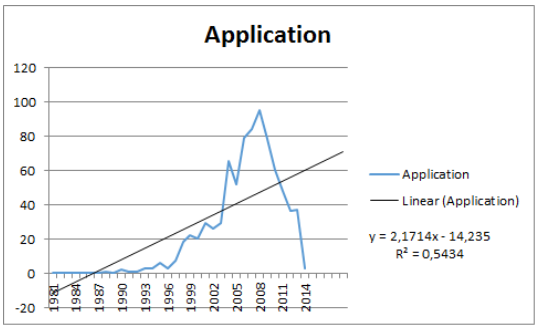

Figura 3. Tendência do termo Application

De acordo com [Pereira et al. 2011], as palavras mais utilizadas na elaboração dos títulos dos artigos científicos relacionados com a área de sistemas embarcados, formam uma rede de pequeno-mundo (*small-word*), pois ao longo do tempo o coeficiente de agrupamento médio (C) mostra um grande valor ($0{,}767 \leq C \leq 0{,}899$) e pequenos valores para o caminho mínimo médio (L) ($2{,}726 \leq L \leq 3{,}194$), quando comparado com o C de uma rede aleatória com a mesma quantidade de vértices e valores semelhantes para o Caminho Mínimo e Densidade.

A densidade reflete a quantidade de ligações entre as palavras encontradas nos títulos e indica a tendência de palavras que aparecerem juntas em um grande conjunto de palavras. Em [Pereira et al. 2011] foi demonstrado que dos anos oitenta até metade da década de noventa, pegando como referência a métrica Densidade, os meios de publicação eram voltados a um tema mais específico, enquanto que ao final do século vinte as publicações foram se tornando de caráter extremamente multidisciplinar, o que condiz a densidade de 2,7 encontrada em nossos experimentos.

5. Considerações finais

Este estudo processou os metadados dos 11303 artigos publicados em diversos meios de publicação realizados entre os anos de 1981 e 2014 com o intuito de analisar os termos mais relevantes na área de sistemas embarcados.

A partir dos dados analisados é possível afirmar que a área de sistema embarcados está com um crescimento menor nos últimos anos, mas apesar disso os meios de publicações relevantes identificados em sua maioria tem uma tendência de crescimento.

Verificamos que em sua maior parte as publicações pertencentes a uma comunidade realmente abordam algum assunto de proximidade a classificação dada. Isso futuramente permitirá descobrir, a partir dos metadados de uma publicação fora da rede, em qual comunidade um determinado assunto se encaixa.

Referências

Balancieri, R. e. a. (2005). A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na plataforma lattes. 34:64–77.

Berkowitz, S. D. (1982). *An introduction to structural analysis: The network approach to social research*. Butterworth-Heinemann.

Bourdieu, P. (2004). *Os usos sociais da ciência: por uma sociologia clínica do campo científico*. UNSEP.

Chakrabarti, S. (2003). *Social Network Analysis, Mining the Web*.

Cui, J., Wang, F., and Zhai, J. (2010). Citation networks as a multi-layer graph: Link prediction and importance ranking.

Cunha, M. V., Rosa, M. G., Fadigas, I., Miranda, J. G. V., and Pereira, H. B. B. (2003). Redes de títulos de artigos científicos variáveis no tempo. *XXXIII Congresso da Sociedade Brasileira de Computação*, 1:1744–1755.

Fadigas, I., Henrique, T., Pereira, H., Senna, V., and Moret, M. (2009). Analise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. *Educação Matemática, Pesquisa*, 11:167–193.

Pereira, H., Fadigas, S., Senna, V., and Moret, M. (2011). Semantic networks based on titles of scientific papers. *Physica A: Statistical Mechanics and its Applications*, 390(6):1192–1197.

Silva, E. L. (2002). Rede científica e a construção do conhecimento. *Informação e Sociedade: Estudos*.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998. ACM.

Tomaél, M. I. and Marteleto, R. M. (2006). Redes sociais: posições dos atores no fluxo da informação. *IV ENANCIB*.

Zhao, Q., Bhowmick, S. S., Zheng, X., and Yi, K. (2008). Characterizing and predicting community members from evolutionary and heterogeneous networks. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 309–318. ACM.