An overview of Brazilian researches in the Computer Science field in last years

Leandro Peres[1], Pablo Cecilio[1], Francielly Rodrigues[2], Nícollas Silva[3], Leonardo Rocha[1]

[1] Universidade Federal de São João del Rei (UFSJ)
leandrocamposperes@gmail.com, pablocecilio@gmail.com, lcrocha@ufsj.edu.br
[2] Laboratório Nacional de Computação Científica (LNCC)
fmunique@lncc.br
[3] Universidade Federal de Minas Gerais (UFMG)
ncsilvaa@dcc.ufmg.br

Abstract. Recently, most traditional market services have joined online service platforms. Despite the practicality achieved, such services eventually bring a large amount of data to the Web. In this sense, data analysis, data engineering, and data science activities have become extremely necessary. In general, they can extract extra information about systems and users, allowing the owners to produce insights and analyze patterns. Then, we propose an evaluation methodology to be applied in the online scenario of registration of publications and scientific productions, such as ResearchGate and Lattes Platform of CNPq. This methodology is unsupervised and divided into three main stages: (i) obtaining and representing the data; (ii) application of topic modeling; and (iii) the labeling of topics. This proposal diverges from the literature's proposes that are based on collaborative networks and supervised techniques. We applied this methodology to a Lattes database and were able to observe the evolution of Computer Science research in Brazil. Based on this analysis, it is possible to identify the most popular and least explored research lines in order to direct public investments according to a certain interest.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications

Keywords: Topic Modeling, Topic Labeling, Data Mining

1. INTRODUÇÃO

Atualmente, com a expansão e popularização da Web, cada vez mais sistemas estão migrando suas aplicações cotidianas para o cenário virtual. Em especial, sistemas que interagem diariamente com os usuários por meio de requisições de dados e informações têm se tornado ferramentas completamente *online* [Cormode and Krishnamurthy 2008]. Além de outros fatores, pode-se dizer que essa expansão também levou ao surgimento de um enorme volume de dados na Web. Grandes empresas, como a Google, por exemplo, processam mais de 24 petabytes de variados tipos de dados por dia, precisando lidar com todas as implicações desse processo [Dean and Ghemawat 2008]. Contudo, apesar dos desafios adicionais desses dados, essas empresas continuam fazendo sucesso no mercado. Um dos principais motivos é que elas passaram a explorar o conhecimento sobre esses dados. Extrair informações extras sobre sistemas e usuários têm se tornado uma peça fundamental para elas, pois são capazes de produzir *insights* e analisar padrões para aprimorar seus sistemas [Choi et al. 2016; Kim et al. 2013]. Em geral, as informações necessárias são obtidas por meio de estratégias conhecidas como processos de *data engineering*, *data analytics*, e *data science*.

Um exemplo de cenário que ilustra bem esse processo são os sistemas *online* de registro de publica-

Esse trabalho foi parcialmente financiado por CNPq, CAPES, FINEP, Fapemig, MASWeb e INWEB.
Copyright©2019 Permission to copy without fee all or part of the material printed in KDMiLe is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

ções e produções científicas, tais como ResearchGate[1] e a Plataforma Lattes do CNPq[2]. Analisar os dados publicados por pesquisadores nessas aplicações pode ser uma importante fonte de informação para diversas tarefas. Com essa análise é possível identificar as linhas de pesquisa mais populares e as menos exploradas, no intuito de direcionar os investimentos públicos de acordo com um determinado interesse. Também é possível realizar um mapeamento das *expertises* de pesquisadores em cada área de conhecimento, bem como os pesquisadores em destaque, para recomendação de profissionais para resolver problemas industriais e/ou científicos específicos e/ou complexos. Uma técnica de análise desses dados já bastante explorada na literatura é a criação de redes de colaboração científica, conectando vários pesquisadores. Basicamente, a conexão entre os pesquisadores é definida a partir da relação de coautoria de artigos publicados, no intuito de unir vários pesquisadores para facilitar e potencializar o desenvolvimento de pesquisas [Grácio 2018]. Uma outra abordagem se preocupa em identificar as áreas mais importantes de pesquisadores e instituições por meio de técnicas supervisionadas, mais especificamente, classificação hierárquica para categorizar os pesquisadores e instituições em vários níveis, visando tornar mais fácil a busca e a organização dos dados científicos [de Siqueira et al. 2018].

Neste trabalho realizamos uma análise global da evolução da pesquisa produzida no Brasil nos últimos dez anos, sobretudo no campo de Ciência da Computação. Para realizar essa análise, propomos uma metodologia de avaliação que difere-se bastante das propostas existentes na literatura (redes de colaboração [Grácio 2018] e técnicas supervisionadas [de Siqueira et al. 2018]). Nossa abordagem é totalmente baseada em estratégias não-supervisionadas aplicadas diretamente sobre informações relacionadas às publicações dos pesquisadores. Ela consiste em três etapas principais: (i) obtenção e representação dos dados; (ii) aplicação de modelagem de tópicos; e (iii) a rotulação dos tópicos. Para a obtenção dos dados existem diversas fontes que podem ser pesquisadas, tais como ResearchGate, DBLP ou mesmo a Plataforma Lattes. Pode-se utilizar os artigos completos publicados pelos pesquisadores, ou apenas seus resumos, ou mesmo apenas seus títulos. Feita a coleta, essas informações precisam ser devidamente representadas. Na estratégia mais comum de representação cada artigo pode ser representado por uma *bag of words* onde a importância de cada palavra em cada documento é definida pelo TF-IDF. A etapa de modelagem de tópicos pode utilizar estratégias clássicas para decompor esses dados definindo as relações latentes entre palavras e publicações para, então, capturar as palavras mais importantes de cada tópico. Existem vários algoritmos de modelagem de tópicos, dentre eles os mais utilizados são *Non-negative Matrix Factorization* (NMF) [Lee and Seung 1999] e o *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003]. Por sua vez, a rotulação de tópicos tem como objetivo descrever os tópicos gerados automaticamente utilizando palavras ou frases, dando origem a rótulos que são utilizados para melhor entendimentos dos tópicos.

A fim de avaliar a estratégia, foram coletadas as publicações dos últimos dez anos de doutores da Ciência da Computação, extraídas dos currículos cadastrados na Plataforma Lattes. Escolhemos essa plataforma pois os currículos cadastrados seguem um padrão nacional e por ela ser amplamente utilizada para avaliação individual dos pesquisadores [Mena-Chalco and Junior 2009]. Empregamos apenas os títulos das publicações e uma representação *bag of words*. Os experimentos consistem em executar a modelagem de tópicos em cada conjunto de publicações pelo ano. Para isso, utilizamos o NMF para inferir as áreas (tópicos) estudadas de cada ano. Para rotular os tópicos, aplicamos a semelhança entre as palavras geradas no tópico com as palavras do sistema de classificação da *Association for Computing Machinery* (ACM). Por fim, mostramos a mudança de relevância das áreas estudadas ao decorrer os anos. A principal contribuição deste trabalho é uma metodologia para analisar a pesquisa científica utilizando técnicas de mineração de texto no intuito de identificar a evolução das áreas estudadas. Esse trabalho está inserido em um projeto de pesquisa mais amplo, cujo objetivo final é identificar automaticamente as principais áreas de atuação e conhecimento de pesquisadores brasileiros nas mais diversas áreas de conhecimento e, a partir dessa identificação, estabelecer mecanismos que possam auxiliar autoridades a formar equipes multidisciplinares para atuar na solução de problemas

[1] https://www.researchgate.net/
[2] http://lattes.cnpq.br/

mais complexos, como por exemplo, o rompimento das barragens de Mariana e Brumadinho.

2. CONCEITOS PRELIMINARES

Nessa seção, apresentamos os conceitos preliminares relacionados ao nosso trabalho. Primeiramente, destacamos a modelagem de tópicos, apresentando as principais abordagens correlatas. Em seguida, discutimos sobre o processo de rotulação de tópicos, clássico da literatura.

2.1 Modelagem de tópicos

De maneira geral, a modelagem de tópicos tem como objetivo encontrar clusters de palavras (tópicos) em uma coleção de documentos, correlacionando semanticamente documentos e as palavras presentes nos documentos aos tópicos. Por meio dessa relação semântica entre os documentos de um mesmo tópico, é possível organizar e analisar documentos similares. Antes de modelar os tópicos, contudo, é importante pré processar os documentos removendo caracteres especiais, números, plurais e *stopwords*. Além disso, é preciso escolher como melhor representar os documentos. A maneira mais adotada para obter tal representação é utilizar um vetor que se baseia na informação de ocorrência das palavras no documento, o que pode ser feito através da relação estatística de *term-frequency* e *inverse document frequency* (TF-IDF) [Salton and Buckley 1988]. A representação por meio do TF-IDF captura a importância de uma palavra considerando não apenas o número de ocorrências no documento, mas também sua raridade em relação a uma coleção de documentos.

O processo de modelagem de tópicos pode ser classificado, basicamente, em duas abordagens: probabilística e/ou fatoração de matrizes. Os modelos probabilísticos são baseados na ideia de que documentos são, na verdade, uma mistura de múltiplos tópicos. Nesses modelos, assume-se que os k tópicos são gerados primeiro, antes dos documentos. Nesse caso, cada tópico é definido como uma distribuição de um vocabulário fixo e cada documento exibe os tópicos em diferentes proporções. Porém, todos os documentos na coleção compartilham o mesmo conjunto de tópicos. Dessa forma, cada documento é representado por uma lista de proporções dessa mistura de tópicos que é reduzida para uma distribuição probabilística em um número k de tópicos, no modelo *pLSI*. Esse modelo, porém, é muito difícil de ser estendido para dados do mundo real, pela sua dificuldade de trabalhar com extensas coleções de dados. Para solucionar esse problema foi proposto o LDA [Blei et al. 2003], que utiliza a distribuição *Dirichlet* para refletir o mundo real, onde tópicos contém apenas um pequeno conjunto de palavras e os documentos pertencem a um pequeno conjunto de tópicos. No entanto, o LDA também possui desvantagens, como a dificuldade de trabalhar com a esparsidade dos dados e a geração de tópicos incoerentes.

Para os modelos de fatoração de matrizes é preciso representar a coleção de documentos como uma matriz $Z \in \mathbb{R}^{n \times m}$, onde n é o número de documentos e m é o tamanho do vocabulário. O objetivo é decompor a matriz Z em duas outras matrizes que preservam as propriedades da matriz Z. Os dois algoritmos mais conhecidos na literatura são o SVD [Golub and Reinsch 1971] e o NMF [Lee and Seung 1999]. No SVD, a matriz Z é dividida em três matrizes $U_{n,k}$, $\Sigma_{k,k}$, $V_{k,m}$, onde $Z = U\Sigma V^t$. Enquanto a matriz U contém a relação de documentos por tópicos e V a relação de tópicos por termos, a matriz Σ contém a relação entre os tópicos. Por sua vez, o NMF divide a matriz Z em duas matrizes $H \in \mathbb{R}^{n \times k}$ e $W \in \mathbb{R}^{k \times m}$, de forma que $Z \approx H \times W$. Nesse caso, a matriz H captura a relação de documentos por tópicos e a matriz W a relação de tópicos e termos, onde k é o número de tópicos. A diferença entre os dois algoritmos é que o NMF produz apenas relações estritamente não negativas, enquanto no SVD podem haver relações positivas e negativas entre os dados.

2.2 Rotulação dos tópicos

Após executar a modelagem de tópicos, seja ela probabilística ou não probabilística, os tópicos extraídos são constituídos de um conjunto de palavras que deve ser analisado de maneira qualitativa, no intuito de identificar o assunto de cada tópico e criar um rótulo para cada. Em outras palavras,

deve-se encontrar uma frase ou um conceito que represente cada tópico. Contudo, em um cenário onde são extraídos mais de 100 tópicos, rotular qualitativamente cada um deles é inviável. Dessa maneira, vários trabalhos propõem estratégias para rotular tópicos automaticamente. A estratégia mais comum para rotular os tópicos é utilizando palavras do vocabulário. Um exemplo é o trabalho [Mei et al. 2007], onde os autores rotulam automaticamente os tópicos gerados por modelos probabilísticos. Primeiramente, são extraídos os rótulos candidatos da coleção de documentos, podendo ser palavras, sentenças ou frases. Então é calculado um *score* desses candidatos baseados em duas abordagens. A primeira, mais simples, compara os rótulos candidatos com as palavras com maior probabilidade no tópico. A segunda representa os rótulos candidatos por um conjunto de palavras e compara esse conjunto de palavras com o tópico. Ambas estratégias de rotulação são falhas uma vez que nem sempre o melhor rótulo está definido no corpo dos documentos.

Uma maneira de melhorar essa estratégia é buscar os rótulos candidatos de uma fonte externa de dados. Alguns trabalhos na literatura utilizam essa abordagem para rotular os tópicos. Em [Hulpus et al. 2013], por exemplo, os autores utilizam uma metodologia de grafos juntamente com os dados do *DBPedia*. Para rotular um tópico, a estratégia proposta gera um grafo que conecta as palavras mais importantes de cada tópico, com um número de *hops* mínimo. A partir desse grafo de conceitos, é utilizada alguma técnica de grafos (e.g., centralidade) para encontrar o nodo central, que será considerado o rótulo desse tópico. Outros trabalhos, como [Nomoto 2011], utilizam a base do *Wikipedia* como fonte externa, considerando os títulos das seções como rótulos candidatos para nomear os tópicos gerados. Em [Coursey et al. 2009], os autores também utilizam toda base do *Wikipedia* para gerar um grafo de enciclopédia, onde os nós são os artigos e suas categorias. A partir desse grafo, são extraído os rótulos candidatos. No nosso trabalho também usaremos essa estratégia de rotular os tópicos utilizando uma base externa, descrita mais adiante.

3. METODOLOGIA DE AVALIAÇÃO

Neste trabalho, propomos uma metodologia para avaliar a evolução da pesquisa em Ciência da Computação no Brasil. De maneira geral, nossa proposta consiste em três etapas principais a serem realizadas em sequência. O primeiro passo é escolher uma fonte de dados nas quais os pesquisadores divulguem seus artigos publicados ao longo dos anos. Feita a coleta e pré-processamento desses dados, adotamos uma estratégia de Modelagem de Tópicos para identificar os principais tópicos relacionados aos artigos publicados. Em seguida, aplicamos uma estratégia de Rotulação de Tópicos para descrever cada um dos tópicos associados. Tal metodologia permite construir análises globais sobre a evolução da pesquisa ao longo dos anos. Sobretudo, essa proposta se diverge das tradicionais da literatura que se baseiam em redes de colaboração e técnicas supervisionadas.

3.1 Etapa 1: Coleta e Representação dos Dados

Atualmente, existem diversas aplicações utilizadas por pesquisadores para divulgar suas publicações, tais como ResearchGate, DBLP[3] e a Plataforma Lattes. Nessas aplicações, os pesquisadores divulgam desde títulos,veículo e ano de publicação dos artigos, até um breve resumo ou mesmo o artigo completo. Nesse trabalho, propomos a utilização apenas dos títulos dos artigos, uma vez que são capazes de sumarizar bem o assunto abordado, ao mesmo tempo que tornam todo o processo computacionalmente viável. Feita a coleta, uma etapa fundamental é pré processar esses artigos, removendo caracteres especiais, números e *stopwords*. Além disso, propomos também a remoção de adjetivos e advérbios, já que os termos mais importantes para identificar um tópico são verbos e substantivos [Luiz et al. 2018]. Por fim, restringimos os termos utilizados ao conjunto de palavras do sistema de classificação da ACM, uma vez que pretendemos encontrar tópicos que se relacionem diretamente com áreas de estudo da Ciência da Computação. Por fim, precisamos representar cada um dos artigo coletados

[3]https://dblp.uni-trier.de/

utilizando alguma modelagem matemática interpretável por algoritmos de mineração de dados, em especial, algoritmos de modelagem de tópicos. Para isso, propomos a representação desses dados como uma matriz, a qual denominamos de Z, onde cada linha representa um artigo (i.e., documento) e cada coluna um termo (i.e., palavras) que compõem os artigos. Em nossa proposta, com intuito de prover uma melhor desambiguação semântica dos termos, os mesmos são representados como bigramas e/ou trigramas, de acordo com a incidência dos mesmos nos artigos [Figueiredo et al. 2011]. Além disso, propomos que cada posição dessa matriz armazene a relação TF-IDF para abstrair a relevância dos termos para o documento. Essa matriz será então utilizada na próxima etapa da metodologia.

3.2 Etapa 2: Modelagem de Tópicos

Para identificar áreas de estudo a partir de uma coleção de artigos acadêmicos, nossa proposta é utilizar estratégias de modelagem de tópicos. Em meio as principais estratégias destacadas na Seção 2, nossa metodologia é baseada no modelo *Non-negative matrix factorization* (NMF), um algoritmo não supervisionado que visa reduzir a dimensionalidade dos dados e agrupá-los em grupos com características similares (*clustering*) [Cichocki et al. 2009; Lee and Seung 1999]. De maneira geral, esse algoritmo decompõe a matriz $Z \in \mathbb{R}^{n \times m}$, onde n é o número de documentos e m é quantidade de termos, em duas matrizes $H \in \mathbb{R}^{n \times k}$ e $W \in \mathbb{R}^{k \times m}$, sendo k o número de dimensões para se representar a matriz Z, tal que $k \ll m$. A partir destas duas matrizes resultantes é possível inferir relações acerca dos k-tópicos gerados. No nosso caso, propomos utilizar a matriz W para abstrair a relação entre termos e tópicos e a matriz H a relação documentos e tópicos. A ideia central por trás do NMF é aproximar as colunas de Z por combinações lineares não negativas dos vetores base (colunas em H) [Luiz et al. 2018]. Assim como em vários algoritmos de *clustering*, a escolha do k (i.e., do número de tópicos) é um problema, pois não existem garantias de encontrar a solução ótima global [Lin 2007].

3.3 Rotulação de tópicos

Para uma análise mais consistente dos tópicos, propomos uma maneira simples de rotulá-los a fim de que seja possível relacionar os tópicos com áreas de estudo de fontes externas. Para isso, como a ideia é avaliar a área de Ciência da Computação, propomos a utilização do sistema de classificação da ACM[4]. Tal sistema de classificação organiza os artigos publicados na ACM por área de estudo sendo composto por 13 grandes áreas, cada uma formada por várias subáreas, todas relacionadas à Ciência da Computação. Dessa forma, para cada tópico gerado, temos possíveis bigramas que representam diretamente uma área da ACM. Como cada documento contém apenas palavras do conjunto de palavras das áreas da ACM, a maneira mais simples de se rotular nossos tópicos foi calcular a similaridade entre cada termo pertencente ao conjunto dos top-10 termos do tópico com as áreas da ACM. Essa similaridade foi calculada a partir da distância de *Leveinstein* de forma que, para cada tópico, se algum termo do conjunto dos top-10 tiver uma similaridade acima de 90% com alguma área da ACM, o tópico está então relacionado a essa área.

Devido a necessidade de estudar os tópicos ao longo do tempo, realizamos uma análise qualitativa para identificar a coocorrência de tópicos. Basicamente, comparamos dois tópicos de tempos distintos por meio da similaridade de cosseno entre os vetores dos mesmos (retirados da matriz H). Para tal, normalizamos a matriz H para que os tópicos de cada ano sejam representados pelas mesmas dimensões, nesse caso os termos do TF-IDF. Com base nisso, geramos uma matriz $L \in \mathbb{R}^{N \times k}$ para cada um dos anos, onde N é a união de todos termos gerados em cada ano pelo TF-IDF. Com isso, é possível realizar a similaridade entre os tópicos do tempo t_0 com os tópicos do tempo t_1. Em seguida, podemos afirmar que o tópico em t_1 que obtiver o maior valor de similaridade acima de um *threshold* será análogo ao tópico comparado em t_0. Por fim, podemos analisar qualitativamente quais as mudanças estruturais do tópico ao decorrer dos anos. Em complemento, se não houverem tópicos com a similaridade acima do

[4] `https://dl.acm.org/ccs/ccs_flat.cfm`

threshold descrito, podemos inferir que aquele tópico não está entre os tópicos mais importantes em t_1.

4. ANÁLISE EXPERIMENTAL

Nessa seção apresentamos uma avaliação de nossa metodologia apresentando a evolução da pesquisa em Ciência da Computação nos últimos 10 anos. Conforme detalhamos a seguir, nossa avaliação foi feita com base na extração das publicações de pesquisadores doutores publicadas na Plataforma Lattes.

4.1 Base de dados

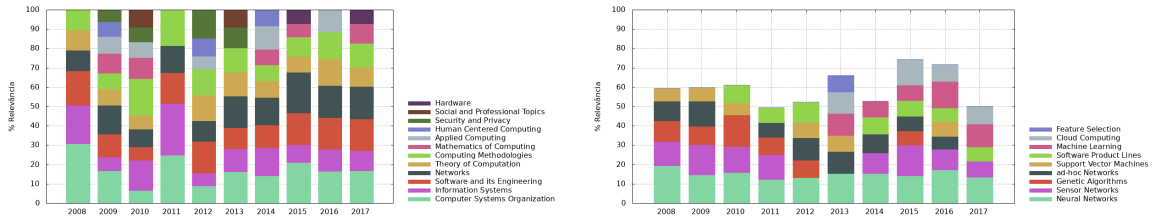
A escolha pela Plataforma Lattes como nossa fonte de dado se deu pela sua grande importância na avaliação dos pesquisadores brasileiros, bem como dos próprios programas de pós-graduação do Brasil. Atualmente, essa plataforma integra dados de vários pesquisadores e instituições das mais diversas áreas, dentre elas a área de Ciência e Tecnologia do Brasil. Para isso, desenvolvemos um coletor que primeiramente fez um levantamento de todos os pesquisadores brasileiros com doutorado. A partir dessa lista, foi feita uma coleta de todas as informações referentes às publicações em periódicos em inglês desses pesquisadores nos últimos 10 anos: título e ano da publicação e lista de autores. Em seguida, realizamos uma eliminação das duplicatas de artigos, uma vez que um mesmo artigo pode ter como autores vários pesquisadores. Em seguida, separamos essas informações por ano, criando diversas sub coleções. Por fim, aplicamos todas as estratégias de representação de dados e pré-processamento descritas na Seção 4. Na Tabela I mostra a quantidade de publicações coletadas em cada ano. É possível observar um crescimento gradual do número de publicações por ano, sugerindo um crescimento da pesquisa brasileira na área da Ciência da Computação. No total, 27.816 artigos foram coletados no intervalo de 10 anos.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
Quantidade de Artigos	1.735	1.868	2.101	2.421	2.782	2.988	3.190	3.521	3.538	3.672	27.816

Table I: Quantidade de artigos em periódicos internacionais publicados por pesquisadores brasileiros por ano, no intervalo de tempo de 2008 a 2017.

4.2 Análise do Perfil da Pesquisa Brasileira

Feita a coleta e o tratamento dos dados, aplicamos a etapa 2 da metodologia descrita na Seção 2 visando extrair os principais tópicos de pesquisa estudados entre os anos de 2008 a 2017. Além disso, aplicamos a etapa 3 da metodologia visando rotular esses tópicos mapeando as palavras mais importantes de cada um deles, definidas pela matriz de fatoração, nas grandes áreas de Ciência da Computação definidas pela ACM (nível mais alto da hierarquia da ACM). Por fim, analisamos também a importância de cada uma dessas áreas de estudo em cada ano. Basicamente, calculamos a relevância de cada área para cada ano por meio dos valores obtidos pelo modelo de extração de tópicos latentes. Todos os tópicos extraídos estão associados a valores semânticos da matriz H de tópicos latentes. Para identificar a relevância do tópico, somamos os valores latentes de cada documento pertencente à essa área.



(a) Classificação pelas grandes áreas de estudo da ACM (b) Classificação pelas sub-áreas de estudo da ACM

Fig. 1: Relevância das áreas de pesquisa na Ciência da Computação no Brasil nos últimos dez anos.

O resultado dessa primeira avaliação pode ser vista no gráfico da Figura 1a. Nela, é possível perceber uma constância de estudos nas áreas de *Computer System Organization* que engloba vários tipos de arquiteturas, sistemas embarcados, sistemas de tempo real, e *networks*. Em especial *Networks* agrupam os tópicos relacionados a redes em geral, como *Neural Networks* e *Sensor networks*. A área de *Hardware*, por sua vez, é pouco estudada no Brasil. Esse tópico aparece em apenas 2 anos (2017 e 2015) com muita pouca importância frente às demais áreas. Avaliando-se a utilidade prática dessas informações, temos por exemplo o direcionamento de recursos de pesquisa para áreas específicas. Por exemplo, se a política nacional estabelece que uma área estratégica de pesquisa seria a área de Hardware e contrapondo esse objetivo as informações da Figura 1, fica claro que haveria grande necessidade de direcionamento de editais de fomento para essa área, bem como políticas de incentivo para abertura de programas de pós-graduação nessa grande área.

Uma avaliação semelhante a descrita acima também foi feita, porém realizando a rotulação dos tópicos por meio de uma mapeamento no nível mais baixo da classificação hierárquica da ACM. Nesse caso, existem um número muito grande de áreas inviabilizando uma apresentação gráfica dessas informações. Dessa forma, adicionamos um novo passo nessa análise. Avaliamos a relevância das diversas áreas de uma forma global, considerando toda a coleção de dados. A partir disso, selecionamos aquelas que, globalmente, consistiam das nove mais relevantes e plotamos essas áreas pela sua relevância em cada um dos anos, separadamente. Os resultados dessa avaliação podem ser observados no gráfico da Figura 1b.

Conforme podemos observar, apesar do grande número de áreas no menor nível da hierarquia ACM, apenas essas nove áreas correspondem a mais de 50% de relevância em todos os anos, demonstrando que há uma concentração muito alta da pesquisa brasileira nessas poucas áreas. Novamente, se o objetivo estratégico nacional fosse diversificar a linhas de pesquisa de computação, incentivos específicos deveriam ser realizados. Uma dessas áreas que está presente em todos os anos é a *Neural Networks*. Outro detalhe importante desses dados está relacionado ao tópico de *cloud computing* que começou a expandir sua relevância nos últimos 5 anos, com o crescimento de vários serviços Web, que utilizam essa arquitetura, como a Amazon e a Google. Outra área também muito relevante nesses últimos anos é a de *Sensor Networks*, presente em praticamente todos os anos.

Para demonstrar o potencial da metodologia aqui proposta, avaliamos em detalhes o tópico de *Sensor Networks* destacando as principais palavras que ocorreram nesse tópico ao logo dos anos, conforme apresentado na Tabela II. Nessa tabela é possível verificar alguns termos que indicam qual o foco de estudo em *Sensors Networks* em cada ano. Em 2008, por exemplo, o termo *'protocol wireless'* indica estudos sobre protocolos em redes de sensores, assim como o ano de 2010. Por sua vez, em 2013, os trabalhos de *Sensor Networks* estavam mais focados em sincronização de tempo, pois várias aplicações de redes de sensores exigem a sincronia temporal [Yildirim and Kantarci 2013].

	termos 1, 2	termos 3, 4	termos 5, 6	termos 7, 8	termos 9, 10
2008	protocol wireless	integer programming	system support	protocols wireless	query processing
2009	time applications	programming model	real time	routing wireless	intrusion detection
2010	protocol wireless	density data	evolutionary design	wireless network	sensor network
2011	mobile wireless	network integrated	system chip	hybrid evolutionary	networks internet
2012	algorithm wireless	clustering algorithm	applications wireless	control algorithms	heterogeneous wireless
2013	time synchronization	fuzzy logic	protocol networks	protocol sensor	system wireless
2014	heterogeneous wireless	vector space	genetic algorithms	connectivity wireless	petri nets
2015	mobile wireless	evaluation energy	heuristics routing	social driven	sensor network
2016	network coding	routing sensor	routing networks	time sensor	theory applications
2017	networks nodes	service wireless	planning algorithm	model mobile	monitoring systems

Table II: Dez termos mais relevantes do tópico de *Sensor Networks* ao longo dos anos.

5. CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho propomos uma metodologia que visa combinar diferentes técnicas de mineração de dados para se realizar uma análise global da evolução da pesquisa produzida no Brasil. As etapas da metodologia foram: (i) obtenção de dados referentes à publicação científica de pesquisadores e a representação desses dados; (ii) aplicação de modelagem de tópicos; e (iii) a rotulação dos tópicos a partir de uma base externa. Instanciamos nossa metodologia avaliando a evolução da área de Ciência da Computação nos últimos 10 anos por meio de informações disponibilizadas na Plataforma Lattes. Em nossas análises observamos que há uma concentração de esforço em um número muito restrito de áreas. Nosso objetivo imediato é aplicar esse mesmo processo em diversas áreas e, posteriormente realizar uma análise semelhante com os pesquisadores. A médio e longo prazo, o objetivo é identificar automaticamente as principais áreas de atuação e conhecimento de pesquisadores brasileiros nas mais diversas áreas de conhecimento e, a partir disso, estabelecer mecanismos que possam auxiliar a estabelecer equipes multidisciplinares para atuar em problemas mais complexos.

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* 3 (Jan): 993–1022, 2003.

Choi, T.-M., Chan, H. K., and Yue, X. Recent development in big data analytics for business operations and risk management. *IEEE transactions on cybernetics* 47 (1): 81–92, 2016.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S.-i. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

Cormode, G. and Krishnamurthy, B. Key differences between web 1.0 and web 2.0. *First Monday* 13 (6), 2008.

Coursey, K., Mihalcea, R., and Moen, W. Using encyclopedic knowledge for automatic topic identification. In *Proceedings of CoNLL*. pp. 210–218, 2009.

de Siqueira, G. O., Canuto, S., Gonçalves, M. A., and Laender, A. H. A pragmatic approach to hierarchical categorization of research expertise in the presence of scarce information. *IJDL*, 2018.

Dean, J. and Ghemawat, S. Mapreduce: simplified data processing on large clusters. *Communications of the ACM* 51 (1): 107–113, 2008.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., and Jr., W. M. Word co-occurrence features for text classification. *Information Systems* 36 (5): 843 – 858, 2011.

Golub, G. H. and Reinsch, C. Singular value decomposition and least squares solutions. In *Linear Algebra*. Springer, pp. 134–151, 1971.

Grácio, M. C. C. Colaboração científica: indicadores relacionais de coautoria. *Brazilian Journal of Information Science: research trends* 12 (2), 2018.

Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, pp. 465–474, 2013.

Kim, K., Chung, B.-S., Jung, J.-Y., and Park, J. Revenue maximizing itemset construction for online shopping services. *Industrial Management & Data Systems* 113 (1): 96–116, 2013.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (6755): 788, 1999.

Lin, C.-J. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19 (10): 2756–2779, 2007.

Luiz, W., Viegas, F., Alencar, R., Mourão, F., Salles, T., Carvalho, D., Gonçalves, M. A., and Rocha, L. A feature-oriented sentiment rating for mobile app reviews. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1909–1918, 2018.

Mei, Q., Shen, X., and Zhai, C. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 490–499, 2007.

Mena-Chalco, J. P. and Junior, R. M. C. Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society* 15 (4): 31–39, 2009.

Nomoto, T. Wikilabel: an encyclopedic approach to labeling documents en masse. In *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pp. 2341–2344, 2011.

Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24 (5): 513–523, 1988.

Yildirim, K. S. and Kantarci, A. Time synchronization based on slow-flooding in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems* 25 (1): 244–253, 2013.