CAIO CESAR TRUCOLO

Análise de tendências em redes sociais acadêmicas

São Paulo

2016

CAIO CESAR TRUCOLO

Análise de tendências em redes sociais acadêmicas

Orientador: Prof. Dr. Luciano Antonio Digiampietri

São Paulo
2016

CATALOGAÇÃO-NA-PUBLICAÇÃO
(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)

Trucolo, Caio Cesar
Análise de tendências em redes sociais acadêmicas / Caio Cesar Trucolo ; orientador, Luciano Antonio Digiampietri. – São Paulo, 2016
65 p. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, em 2015
Versão corrigida

1. Redes sociais. 2. Análise de tendências. 3. Publicações acadêmicas. I. Digiampietri, Luciano Antonio, orient. II. Título

CDD 22 ed.– 303.4833

Dissertação de autoria de Caio Cesar Trucolo, sob o título **"Análise de tendências em redes sociais acadêmicas"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Sistemas de Informação, aprovada em ___ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr. _____
Presidente
Instituição: _____

Prof. Dr. _____
Instituição: _____

Prof. Dr. _____
Instituição: _____

Dedico este trabalho aos meus pais.

Agradecimentos

Agradeço aos meus pais, Osvaldo e Maria das Graças, por todo o apoio ao estudo que me deram desde que me conheço por gente. E também à minha irmã, Fabiana, por suas palavras de carinho e amizade que sempre me incentivaram.

Agradeço à minha grande amiga e namorada, Gabriela, que me ajudou muito e me deu estratégicos empurrões nos momentos certos.

Agradeço aos meus amigos pelos gestos de incentivo.

Agradeço especialmente ao Prof. Dr. Luciano Digiampietri que, muito mais do que me orientar e me apoiar nesses últimos anos, me mostrou o quão fascinante pode ser a ciência nessa área sem limites que é a computação.

Agradeço também a todos os professores do Programa de Pós-graduação em Sistemas de Informação que me ensinaram e auxiliaram em todos os momentos que precisei.

À CAPES e à Universidade de São Paulo, agradeço pelo financiamento dessa pesquisa.

"The road and the tale have both been long, would you not say so? The trip has been long and the cost has been high... but no great thing was ever attained easily. A long tale, like a tall Tower, must be built a stone at a time."

(Stephen King, The Dark Tower)

Resumo

TRUCOLO, Caio Cesar. **Análise de tendências em redes sociais acadêmicas** 2016. 65 f. Dissertação (Mestrado em Ciências) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2016.

Conforme o volume e a diversidade de informações científicas aumentam, se torna necessário entender o que, porque e como esse aumento acontece. Estratégias e políticas públicas podem se desenvolver a partir dessas informações potencializando os serviços de educação e inovação oferecidos à sociedade. A análise de tendências é um dos passos nessa direção. Este trabalho no entanto vai além de considerar apenas o conteúdo das informações analisadas incluindo a estrutura das fontes geradoras das informações, ou seja, as redes socias, como uma dimensão adicional para modelar e predizer tendências ao longo do tempo. Os experimentos foram realizados com os títulos das publicações de todos os doutores brasileiros da área de Ciência da Computação. Os resultados mostraram que a incorporação das medidas oriundas da análise de redes sociais reduziram os erros de predição, na média, para cerca de 18% daqueles produzidos sem a utilização destas medidas. Adicionalmente, esta incorporação permitiu que previsões mais futuras fossem realizadas sem grandes aumentos no erro destas previsões.

Palavras-chaves: Análise de Tendências, Redes Sociais, Redes Sociais Acadêmicas

Abstract

TRUCOLO, Caio Cesar. **Trend Analysis in Academic Social Networks** 2016. 65 p. Dissertation (Master of Science) - School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2016.

As scientific information volume and diversity grow, the understanding of what, why and how it happens become necessary. Strategies and public politcs can be developed from these informations to power innovation and education services offered to society. Trend analysis is one of the steps in this direction. This work however goes beyond of just anlyse information content, it includes the information about the information sources structures, in other words, the social networks, as another dimension to model and predict trends through time. The experiments were made with the publication titles of all Brazilian Computer Science´s PHds. The results indicate the use o social network analysis' metrics reduced the precision errors to about 18% of the errors produced without considering these metrics.

Keywords: Trend Analysis, Social Network, Academic Social Network

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

API	Interface de programação de aplicações - *Application Programming Interface*
ART	Teoria da ressonância adaptativa - *Adaptive Resonance Theory*
EMA	Média móvel exponencial - *Exponential Moving Average*
ETD	Identificação de tendências emergentes - *Emerging Trend Detection*
HOSVD	Decomposição de valor singular de alta ordem - *High Order Singular Value Decomposition*
I-GI	Índice de Gini melhorado - *Improved Gini Index*
KDD	Descoberta de conhecimento em base de dados - *Knowledge Discovery in Databases*
KNN	K - Vizinhos mais próximos - *K - Nearest Neighbours*
LDA	Alocação latente de Dirichilet - *Latent Dirichilet Allocation*
LSH	Hashing sensitivo de localidade - *Locality Sensitive Hashing*
M5P	Implementação de *Rotation Forest* baseado em regras de M5
MAE	Erro médio absoluto - *Mean Absolut Error*
PPO	Oscilador de porcentagem de preço - *Price Percentage Oscilator*
PUK	*Kernel* universal baseado na função de Pearson VII - *Pearson VII function-based Universal Kernel*
RAE	Erro relativo absoluto - *Relative Absolut Error*
RBF	Função de base radial - *Radial Based Function*
RNA	Rede Neural Artificial
SOM	Mapas auto organizáveis - *Self Organizing Maps*
SVD	Decomposição de valor singular - *Singular Value Decomposition*

SVM	Máquina de vetores de suporte - *Support Vector Machine*
TDT	Identificação de tendências de tópicos - *Topic Trend Detection*
TF-IDF	Frequência do termo em relação ao inverso da frequência nos documentos *Term Frequency - Inverse Document Frequency*

Lista de símbolos

Γ	Letra grega Gama

Sumário

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Nos últimos anos, um grande número de aplicações de mídias sociais tem surgido, gerando uma enorme quantidade de dados na internet. Informações úteis podem ser extraídas desses dados para aumentar o poder de competitividade em negócios ou auxiliar na criação de novas políticas públicas. Análise de tendências é uma das áreas de pesquisa que podem ser exploradas para prover ideias e estratégias para organizações prevendo o comportamento de indivíduos ou grupos.

Ao se considerar o meio acadêmico, estratégias e políticas públicas têm sido inseridas no país para aumentar a produtividade e a qualidade da pesquisa. Muitas vezes essas políticas focam áreas de pesquisa já consolidadas e populares, nas quais se acredita que haverá retorno, ou ainda, identificadas como tendências globais. Entretanto, um país com dimensões continentais como o Brasil - tanto em extensão geográfica quanto em diversidade cultural - poderia investir nas áreas e temas de pesquisa com os maiores potenciais de crescimento, ampliando as chances do retorno da investigação científica. Assim, a análise de tendências se configura como uma estratégia para se encontrar temas de pesquisa com potencial de impacto.

No Brasil existe uma base de dados acadêmicos bastante rica, composta pelo currículo de mais de quatro milhões de pessoas. Esta base encontra-se na Plataforma Lattes[1], mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico[2] (CNPQ) e cujas informações são geridas pelos próprios possuidores dos currículos. Estes currículos contém informações relevantes para diferentes análises acadêmicas e bibliométricas: endereço profissional; formação acadêmica; produções bibliográficas; orientações; participação em eventos; entre outras.

Nesta dissertação, foi estudado um novo modelo de análise de tendências que agrega métricas de redes sociais à análise de séries temporais. A hipótese considerada neste trabalho é que a inserção de informações de redes sociais ajudará a explicar o comportamento temporal das produções científicas nacionais, aumentando a acurácia da predição de tendências.

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

1.1 Contextualização

Dentro do contexto de mineração de texto, tendências podem ser identificadas a partir da perspectiva do comportamento temporal do conteúdo. Para esse tipo de análise utiliza-se como fatores índices de importância baseados em frequência dos tópicos estudados e o tempo. Kontostathis et. al. (2004) descrevem uma tendência como sendo um tópico que está tendo um crescimento de interesse e utilidade ao longo do tempo. Diversos trabalhos inseridos nessa perspectiva de conteúdo identificam tendências para muitos tipos de aplicações como Twitter, mercado de ações, notícias, etc.

A análise de tendências também pode ser observada pela perspectiva de características das fontes geradoras de conteúdo. Diferentemente de se ater ao conteúdo em si, essa forma de análise considera as diferenças de características dos indivíduos que produzem a informação e as influências sociais geradas. A fim de se conseguir explicar melhor o comportamento temporal de tópicos, este mestrado utiliza a análise de redes sociais como um fator adicional à análise de tendências da primeira perspectiva discutida nesta seção.

Com essa abordagem, surgem alguns desafios e questões ligados principalmente a adição da análise de redes sociais. Primeiro, qual a melhor forma de inserir os aspectos sociais dos indivíduos geradores de informação na análise de tendências? Segundo, quais informações sociais utilizar? Além dessas, ainda existem outras questões relacionadas ao período ideal de análise da rede social e escolha das melhores técnicas de previsão.

A inserção de características sociais dos indivíduos geradores de informação na análise de tendências pode ser considerada uma área nova e que ainda carece de uma teoria formalmente definida e aceita, portanto, este mestrado apresenta uma abordagem inicial para esta área de pesquisa que pode ser bastante explorada conforme discutido ao longo da dissertação.

1.2 Objetivos

O objetivo desta dissertação é a especificação e a implementação de um modelo computacional para a análise de tendências utilizando informações de redes sociais acadêmicas. Para se atingir este objetivo geral e como objetivos específicos estão: (a) a realização de uma revisão sistemática acerca dos métodos e técnicas utilizados na identificação e análise

de tendências para a identificação do que está sendo produzido nessa área atualmente bem como os pontos passíveis de aprimoramento; (b) aplicação de algoritmos existentes de análise ou predição de tendências considerando apenas os documentos científicos (sem considerar a informação de redes sociais); (c) especificação e implementação de algoritmo que combine informações extraídas de documentos científicos com as informações da rede social acadêmica na qual os documentos foram produzidos; (d) desenvolvimento de um estudo de caso real e análise comparativa dos resultados (comparando-se a solução proposta com soluções existentes).

1.3 Metodologia

Nesta seção, a metodologia é explicada em sua forma mais geral. Primeiro, o capítulo 3 descreve a revisão sistemática utilizada para fazer o levantamento do estado da arte. Já a construção detalhada do modelo proposto desde a aquisição da base de dados até os testes finais são descritos no capítulo 4.

Neste mestrado, optou-se pela utilização da metodologia de revisão sistemática baseada em Kitchenham (2004). Mais do que apenas identificar trabalhos relevantes de determinada área de pesquisa, com esses métodos a autora propõe criar uma organização sistemática no processo de pesquisa possibilitando a avaliação da revisão e a reprodução da mesma pelos próprios leitores. Dessa forma, foi realizada uma revisão sistemática com o objetivo de aprofundamento nos conceitos de análise de tendências para alguns tipos de aplicação.

A coleta de dados foi baseada em Digiampietri *et. al.* (2012) cuja fonte é a plataforma Lattes, que mantém mais de quatro milhões de currículos de pesquisadores do Brasil.

A implementação do modelo como um todo consiste em quatro partes: a) extração automática de termos; b) análise de séries temporais; c) análise de redes sociais; d) análise de tendências com a utilização de fatores de redes sociais.

Os testes foram realizados com o arcabouço Weka utilizando a técnica de validação cruzada (*cross-validation*).

1.4 Organização da Dissertação

Este documento está organizado da seguinte forma: o capítulo 2 contém definição dos principais conceitos do trabalho, o capítulo 3 apresenta a revisão sistemática acerca das técnicas de análise de tendências, o capítulo 4 possui os experimentos e desenvolvimento do modelo e, por fim, o capítulo 5 possui uma discussão sobre os resultados obtidos.

2 Conceitos Fundamentais

Há dois principais conceitos abordados nesta dissertação. O primeiro é a análise de tendências em si, contextualizado dentro da descoberta de conhecimento em bases de dados. O segundo é a análise de redes sociais.

2.1 Descoberta de Conhecimento em Bases de Dados

Descoberta de Conhecimento em Bases de Dados (ou do inglês, *Knowledge Discovery in Databases* - KDD) é definida por Han, Kember e Pen (2011) como "*o processo de descobrir padrões interessantes e conhecimento de grandes volumes a partir de dados*". Muitas vezes, a expressão mineração de dados é utilizada como sinônimo de KDD, porém corresponde a um de seus passos (considerado por muitos o passo mais importante). O termo mineração por si indica o processo de identificar e extrair objetos preciosos a partir de grandes massas formadas por diferentes tipos de materiais.

Para Han, Kember e Pen (2011) o processo de adquirir conhecimento a partir da mineração de dados segue um sequência de passos:

1. Limpeza dos dados: remoção de ruído e dados inconsistentes;
2. Integração dos dados: integração de diferentes fontes de dados;
3. Seleção dos dados: recuperação dos dados relevantes para a base de dados;
4. Transformação dos dados: transformação e consolidação dos dados para uma mineração apropriada dos dados;
5. Mineração dos dados: aplicação de técnicas inteligentes para a extração de padrões dos dados;
6. Avaliação dos padrões: identificação de padrões interessantes;
7. Apresentação do conhecimento: aplicação de técnicas de visualização e representação do conhecimento adquirido.

A mineração de texto é uma subárea da mineração de dados. Karanikas e Theoudolidis (2002) definem o processo de mineração de texto como "*o processo não trivial de identificar padrões válidos, originais, potencialmente úteis e entendíveis em dados textuais não estruturados*".

As duas abordagens são similares tendo passos de pré-processamento, reconhecimento de padrões e apresentação de conhecimento. Na mineração de texto, os dados não estruturados (textos) são pré-processados para a apropriada análise posterior, sendo que a principal tarefa consiste nas técnicas de extração de padrões.

2.1.1 Mineração de Texto Temporal

A mineração de texto usa conjuntos de documentos como base. Esses conjuntos de documentos têm como unidade os próprios documentos. Os documentos podem ser observados como textos de diferentes aplicações como postagens de blogs, artigos científicos, notícias ou e-mails, por exemplo. Esses dados costumam ser não estruturados e algumas vezes fracamente estruturados como discutem Feldman e Sanger (2007).

Feldman e Sanger (2007) também discutem a mineração de dados temporal. Segundo esses autores, a mineração de dados temporal é um caso especial da mineração de dados, que se difere pelo fato da análise ser em documentos que possuem algum tipo de índice temporal. Especificamente, em mineração de texto temporal, o conjunto de documentos possui a característica de suas unidades terem selos de tempo, ou seja, cada documento possuir um índice temporal, podendo ser um tempo específico ou um período.

Dessa forma, o processo de mineração de texto temporal consiste em descobrir padrões temporais nos dados textuais coletados.

2.2 Análise de Tendências

Análise de tendências é um termo comum utilizado por diferentes áreas do conhecimento em diferentes pontos de vista. Não existe uma definição formal e amplamente aceita por essas diferentes áreas. Em sociologia, por exemplo, analisar tendências é o método de observação de mudanças no comportamento de pessoas ou grupos de pessoas ao longo do tempo. Por esse ponto de vista, tendências são padrões de comportamento social e estilo de vida que são identificados em determinados intervalos de tempo (Vejlgaard, 2008).

Para a área de exatas, mais especificamente no ponto de vista estatístico, a análise de tendências consiste em modelar séries temporais com os dados extraídos para entender o comportamento desses dados e, então, prever valores futuros das variáveis dessas séries

modeladas. Com esses métodos, identifica-se uma direção geral que o comportamento dos dados está tomando dentro de algum intervalo de tempo (Han e Kember, p.490).

Apesar de serem pontos de vista diferentes, há em comum o fator de mudança de comportamento ao longo do tempo. Portanto, o termo tendências nessa dissertação é entendido como o comportamento futuro do objeto de estudo a partir da análise histórica de comportamento desse mesmo objeto dentro do intervalo de tempo analisado.

Dentro do contexto de mineração de texto, os objetos de análise são os documentos de texto temporais com seus devidos índices de tempo. Kontostathis et. al. (2004) definem tendência em mineração de texto como "*tópico que cresce em interesse e utilidade com o tempo*". Para a construção do conjunto de documentos para a análise de tendências, muitas vezes a identificação dos tópicos que compõem os documentos necessita de esforço humano. Um exemplo disso pode ser visto na necessidade de especialistas de determinada área do conhecimento para gerar palavras ou termos importantes para serem analisados. Procurando automatizar todo esse proesso de construção do conjunto de documentos, foi apresentada a abordagem de TDT. Essa abordagem consiste de técnicas que buscam agrupar informações similares em tópicos (Kontostathis et. al. (2004)).

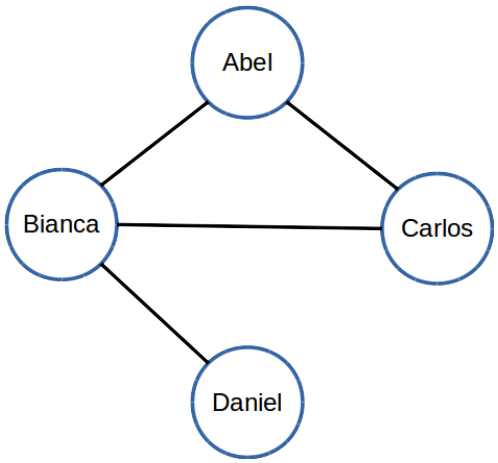
2.3 Análise de Redes Sociais

Atualmente, vive-se uma era de computação social. Quase tudo o que cerca as pessoas envolve algum tipo de interação social com o apoio da tecnologia. Todo esse ecossistema de conexões exerce influência no modo de vida das pessoas. Os grupos sociais fechados e pequenos de antigamente deram espaço aos grupos intercontinentais e multiculturais de hoje. Informações se difundem rapidamente por conta do alto volume de conexões de usuários através dos seus meios de comunicação ligados à internet.

Ao conjunto formado por entidades (pessoas, grupos ou organizações) e seus relacionamentos é dado o nome de rede social. Normalmente, essas redes são modeladas como grafos, sendo os nós do grafo as entidades, e as arestas a interação entre as entidades (LEMIEUX; OUIMET, 2008). Um exemplo de rede social representada por grafos pode ser visto na figura 1. Nessa figura, os indivíduos Abel, Bianca e Carlos são amigos em comum enquanto Daniel é apenas amigo de Bianca.

A análise de redes sociais proporciona a possibilidade de se entender algumas características importantes da rede como um todo ou mesmo dos próprios indivíduos

Figura 1 – Representação em grafo não direcionado de uma rede social formada por quatro indivíduos



Fonte: Caio Cesar Trucolo, 2015

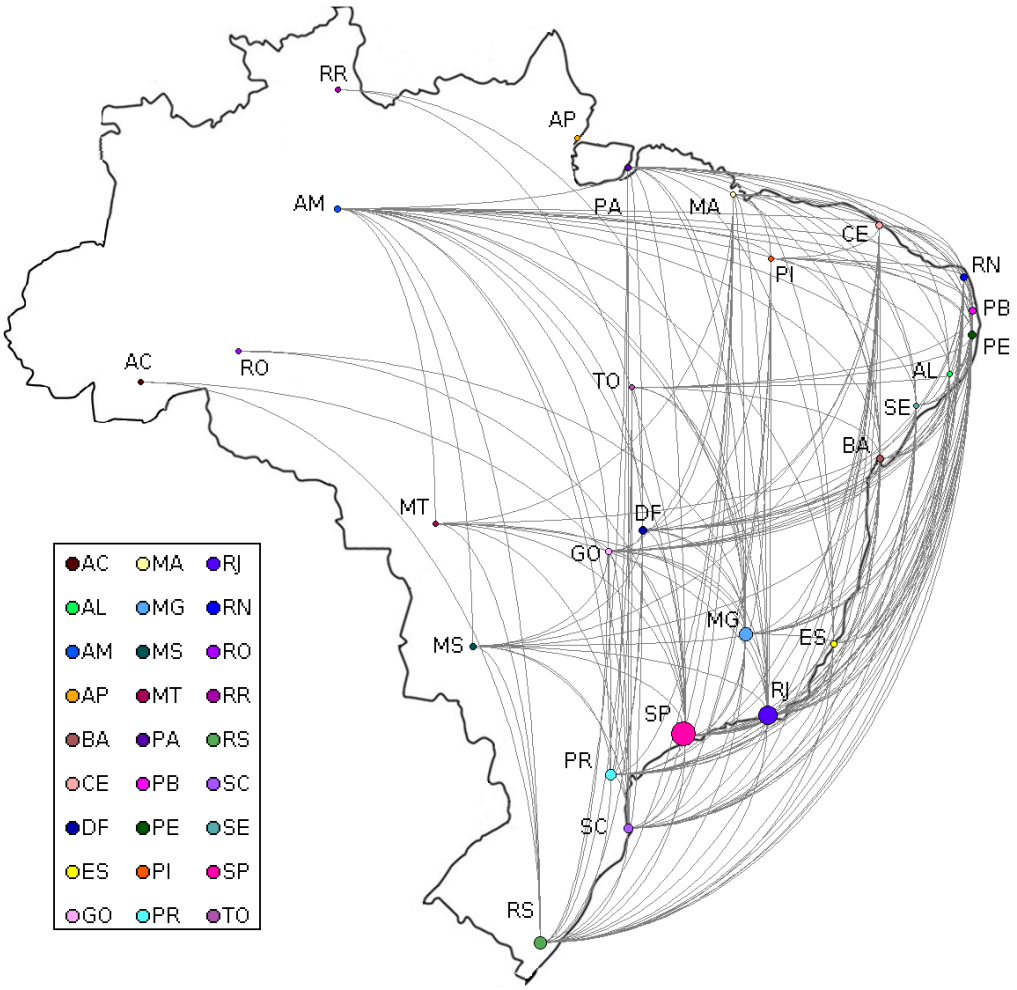
que a formam. Existem algumas medidas importantes que já vem sendo trabalhadas há algum tempo como discutidos por Zafarani, Abbasi e Liu (2014). Uma das medidas mais utilizadas é a centralidade, que indica níveis de centralização maiores para indivíduos com altos números de conexões ou com caminhos mais curtos entre dois ou mais indivíduos. Muitas vezes, quanto mais central um indivíduo é dentro da rede, mais influência ele tem sobre os demais. Singh (2013) detalha algumas técnicas para se identificar indivíduos mais influentes dentro de uma rede social. Outra medida bastante utilizada é similaridade, que indica quão parecidos dois indivíduos são, com base no formato da vizinhança de dois indivíduos ou mesmo pela similaridade dos próprios vizinhos.

Outras medidas importantes provenientes da análise de redes sociais relaciona-se com o comportamento de grupos mais coesos dentro da própria rede. Um exemplo de aplicação é a recomendação baseada em comunidade. Produtos podem ser recomendados para indivíduos de acordo com o comportamento da comunidade a qual ele pertence. Para esse fim é necessário identificar essas comunidades, o que não é uma atividade trivial. Pourkazemi e Keyvanpour (2013) discutem algumas das técnicas para a identificação de comunidades.

Nesta dissertação são usadas redes sociais de colaboração científica. Sonnenwald (2007) define colaboração científica como a interação entre dois ou mais cientistas compartilhando o mesmo objetivo para facilitar o desenvolvimento de tarefas. O relacionamento utilizado nesta dissertação é de coautoria, isto é, dois pesquisadores serão ligados caso tenham uma ou mais publicações conjuntas.

As figuras 2, 3, 4 e 5 foram extraídas de Digiampietri et al. (2014) que utilizaram parte dos dados utilizados nessa dissertação com o objetivo de caracterizar doutores brasileiros da área de Ciência da Computação. Na figura 2 pode-se observar as relações inter-estaduais sendo cada nó um estado brasileiro e o tamanho do nó é proporcional à quantidade de doutores trabalhando no respectivo estado. Na figura 3 cada nó representa um doutor em ciência da computação posicionado próximo à cidade onde atual (na imagem da esquerda aparecem apenas as relações dentro de cada estado, enquanto que na da direita aparecem todas as relações, sendo as inter-estaduais representadas por arestas em cinza).

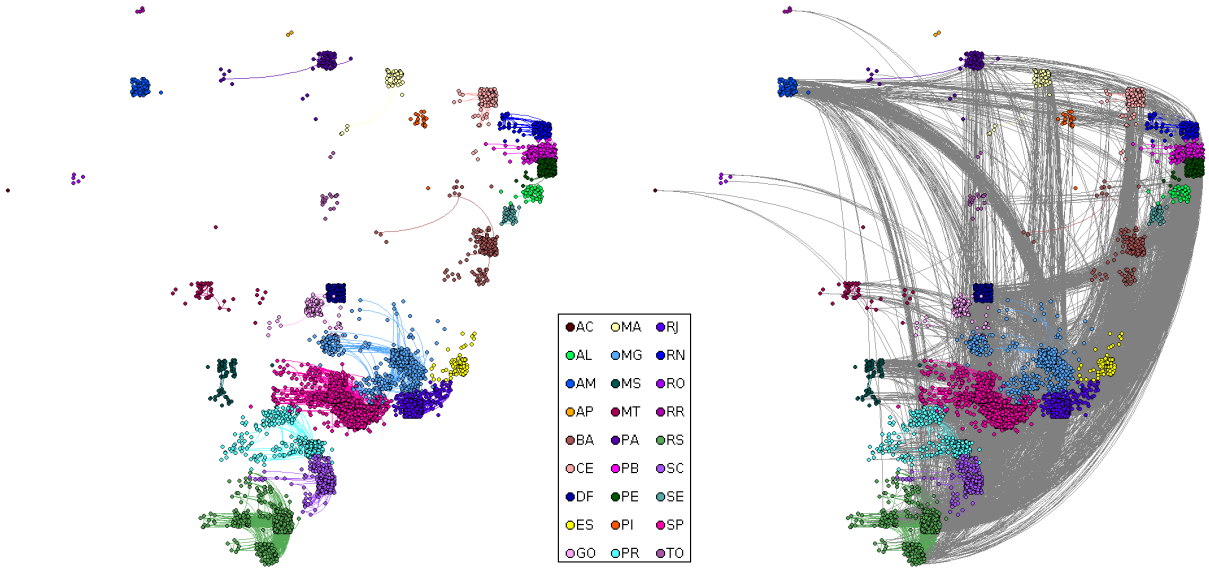
Figura 2 – Rede social de doutores brasileiros da área de Ciência da Computação por estados brasileiros



Fonte: (DIGIAMPIETRI et al., 2014)

As figuras 4 e 5 contêm os nós da figura 3 reposicionados. O algoritmo utilizado para reposicionar os nós nestas figuras usa força de repulsão para os nós que não são

Figura 3 – Rede social de doutores brasileiros da área de Ciência da Computação

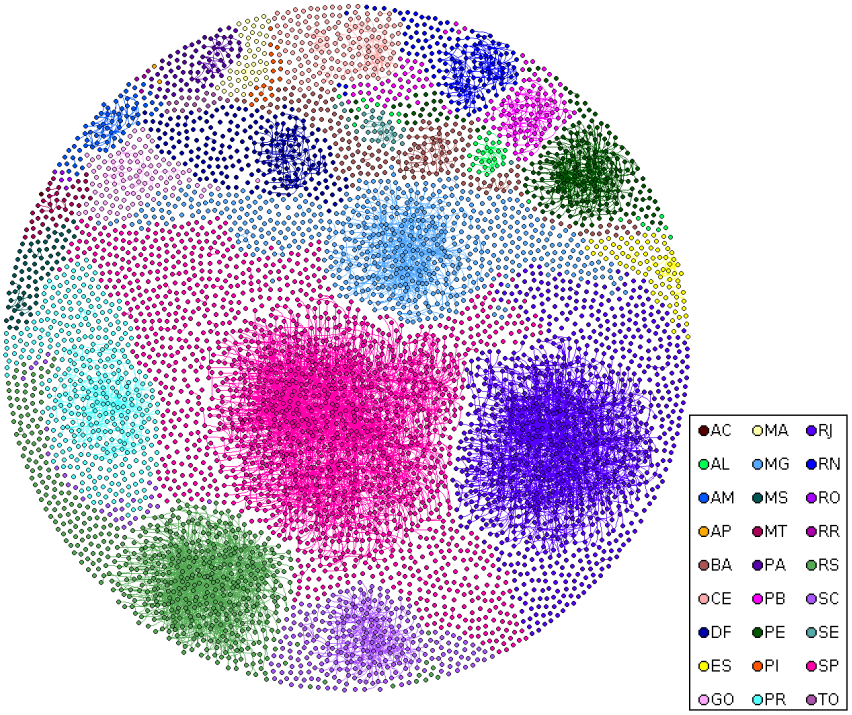


Fonte: (DIGIAMPIETRI et al., 2014)

relacionados e atração para os nós que são relacionados. A figura 4 contém apenas os relacionamentos entre nós de um mesmo estado. Nela é possível se ter noção do tamanho das redes estaduais e quantidade de relações entre seus indivíduos. Na figura 5, destaca-se o componente gigante, contendo a grande maioria dos nós da rede.

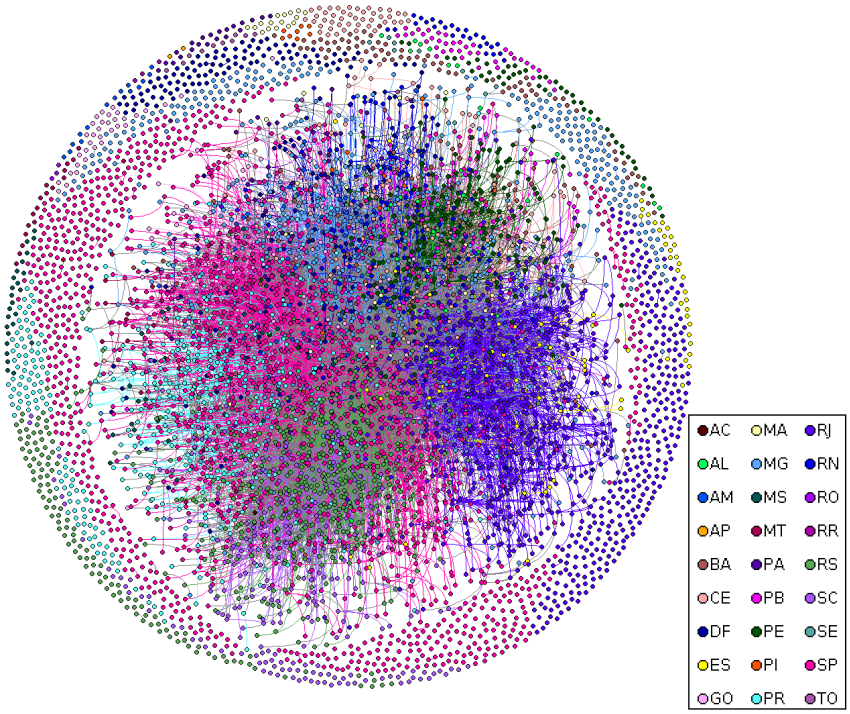
As métricas de uma rede social podem servir para muitos tipos de aplicações e pesquisas. Recomendação, viralização e predição são algumas dessas aplicações. Apesar de redes sociais serem estudadas há algumas décadas em seu sentido mais social, sua inserção nessa sociedade envolta em tecnologia e internet vem sendo explorada há pouco tempo. Diversas áreas ainda podem se beneficiar da análise de redes sociais.

Figura 4 – Rede social de doutores brasileiros da área de Ciência da Computação por estados brasileiros de forma reorganizada (contendo apenas relacionamentos entre doutores de um mesmo estado)



Fonte: (DIGIAMPIETRI et al., 2014)

Figura 5 – Rede social de doutores brasileiros da área de Ciência da Computação por estados brasileiros de forma reorganizada



Fonte: (DIGIAMPIETRI et al., 2014)

3 Revisão da Literatura

Neste trabalho, a revisão da literatura foi realizada por meio de uma revisão sistemática. A revisão sistemática é uma ferramenta bastante importante para que seja possível identificar e organizar o que está sendo produzido no meio acadêmico sobre determinado tema (KITCHENHAM, 1997). O objetivo da revisão sistemática deste projeto de mestrado foi analisar os métodos e técnicas acerca da análise de tendências em diferentes tipos de aplicação. Tendências são analisadas em inúmeros tipos de aplicação, desde aplicações de áreas biológicas até aplicações de áreas sociais. Dessa forma, julgou-se necessário diminuir a abrangência das aplicações para um universo mais próximo da aplicação final deste mestrado. Portanto, os trabalhos analisados na revisão necessitavam ter suas aplicações encaixadas em contexto sócio-técnico como análises em blogs, redes sociais digitais, fóruns, etc.

3.1 Método da Revisão

Para iniciar a revisão foi necessário identificar algumas palavras-chave para a busca dos artigos. Essas palavras-chave foram determinadas a partir de uma pesquisa exploratória inicial. As palavras-chave utilizadas foram: *trend analysis, tendency analysis, trend identification, tendency identification, trend recognition, tendency recognition, trend detection, tendency detection, trend prediction, tendency prediction* e *trending*. A busca dos artigos foi realizada entre 07/10/2013 e 11/10/2013 em duas das principais bibliotecas digitais de artigos científicos na área de computação: ACM Digital Library e IEEExplore. Essas duas bases de dados foram escolhidas por causa do foco deste trabalho em revisar apenas artigos com aplicações de interação social em âmbito tecnológico.

Duas strings de busca foram utilizadas, uma para cada biblioteca digital:

ACM Digital Library: Owner:ACM(Keywords:“trend analysis” OR Keywords: “tendency analysis” OR Keywords:“trend identification” OR Keywords:“tendency identification” OR Keywords:“trend recognition” OR Keywords: “tendency recognition” OR Keywords:“trend detection” OR Keywords:“tendency detection” OR Keywords:“trend prediction” OR Keywords:“tendency prediction” OR Keywords:“trending”)

IEEExplore: ((((((((((“Author Keywords”:“trend analysis”) OR “Author Keywords”: “tendency analysis”) OR “Author Keywords”:“trend identification”) OR “Author Keywords”:

“tendency identification”) OR “Author Keywords”:“trend recognition”) OR “Author Keywords”: “tendency recognition”) OR “Author Keywords”:“trend detection”) OR “Author Keywords”: “tendency detection”) OR “Author Keywords”:“trend prediction”) OR “Author Keywords”: “tendency prediction”) OR “Author Keywords”:“trending”)

A primeira etapa da seleção dos artigos foi baseada na leitura de seus resumos (*abstract*). Na etapa posterior, todos os artigos restantes foram lidos na íntegra. A seleção dos artigos foi realizada de acordo com critérios de inclusão e exclusão pré-estabelecidos para a condução da revisão sistemática. Foi utilizado um critério de inclusão e cinco de exclusão. Critério de inclusão:

1. Trabalhos que apresentam técnicas de análise de tendências;

Critérios de exclusão:

1. Trabalhos que não apresentem os métodos utilizados;
2. Trabalhos secundários;
3. Trabalhos duplicados;
4. Trabalhos não disponíveis integralmente;
5. Trabalhos que apresentem aplicações fora do contexto sócio-técnico.

A partir dos trabalhos selecionados após as duas etapas, foram extraídos dados para a construção da análise.

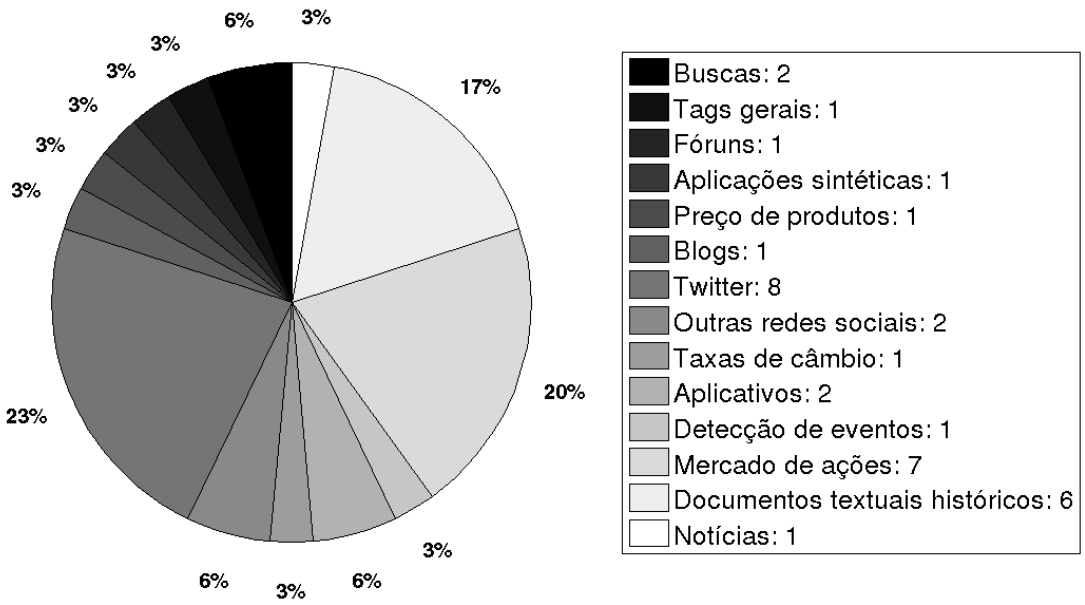
3.2 Condução

Na busca realizada nas duas bibliotecas digitais foram retornados 164 artigos. Na primeira etapa de seleção, que consistiu da leitura dos resumos, foram selecionados 62 artigos, sendo 26 da ACM Digital Library e 36 da IEEExplore. Todos esses 62 artigos foram lidos na íntegra para a segunda etapa de seleção. Nesta etapa foram selecionados 35 artigos para a condução da revisão.

Em primeiro lugar, os artigos foram classificados de acordo com sua aplicação. Conforme pode ser visto pela figura 6, três tipos de aplicações se destacam: Twitter, documentos textuais históricos e mercado de ações. A alta porcentagem de utilização do Twitter como aplicação dos artigos pode ser explicada pelo alto número de usuários, pela enorme quantidade de informação gerada diariamente e a facilidade de se obter uma

parcela desses dados. Os documentos científicos também têm uma alta taxa de adoção para as pesquisas de análise de tendências por servirem como um bom exemplo de dados históricos e por existirem algumas bases de dados amplas. Já as aplicações relacionadas ao mercado de ações são bastante utilizadas por serem aplicações relevantes para técnicas que analisam séries temporais.

Figura 6 – Gráfico da distribuição das aplicações



Fonte: Caio Cesar Trucolo, 2014

Existem alguns problemas relacionados à utilização de dados de redes sociais como o Twitter como aplicação. O primeiro problema acontece na obtenção dos dados. A maioria dos pesquisadores utiliza a API do Twitter para essa ação e essa API possui alguns limites de chamadas por tempo e isso prejudica a importação desses dados[1]. O outro problema consiste em não se poder replicar os experimentos por causa do dinamismo dos dados (quando os autores não disponibilizam exatamente os dados que foram utilizados, pois estes não podem ser facilmente consultados no Twitter). Por outro lado, o Twitter apresenta o benefício de apresentar dados contemporâneos.

Com a análise dos artigos, observou-se que as técnicas de análise de tendências, em sua maioria, podem ser segmentadas em dois grupos: ETD e análise de séries temporais. Os artigos de ETD têm como aplicações análise de redes sociais em geral e documentos textuais históricos. Enquanto os artigos de análise de séries temporais têm como principal aplicação o mercado financeiro. Isso pode ser explicado pelo fato de que grande parte da

[1] https://dev.twitter.com/docs/rate-limiting/1.1

pesquisa relacionada ao mercado financeiro se utiliza apenas dos dados de volume e valores (volume de venda de ações, valor máximo e mínimo diário, por exemplo) e tempo. Já pesquisas de ETD necessitam identificar informações relevantes (como tópicos ou termos importantes) dentre uma enorme gama de informação (caso do Twitter) e identificar tendências a partir delas. ETD também difere de análise de séries temporais por focar em tópicos emergentes e muitas vezes não identificar o comportamento de tópicos cujas tendências não são positivas. Alguns artigos analisados tiveram como objetivo apenas o TDT, cujas técnicas podem ser utilizadas para um sistema completo de análise de tendências.

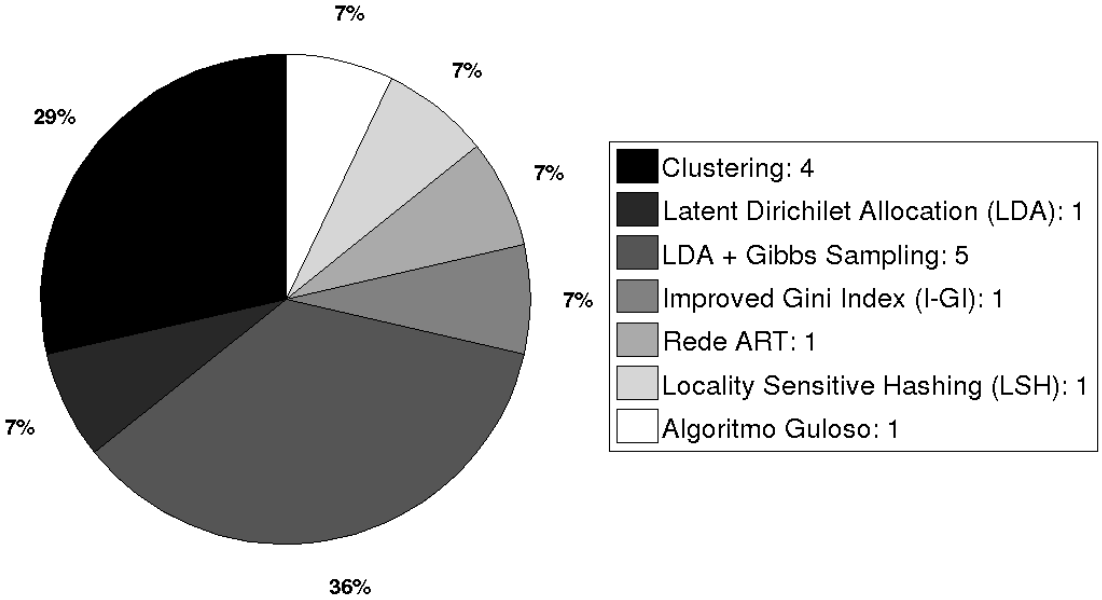
Como as técnicas analisadas nos artigos foram utilizadas de diferentes formas e algumas vezes com diferentes objetivos, foi criada uma abordagem classificatória de tais técnicas. Três classes foram determinadas: Identificação de tópicos, Detecção de tendências emergentes e Reconhecimento de padrões. As técnicas de identificação de tópicos são técnicas que buscam agrupar termos por algum tipo de similaridade ou mesmo entender a distribuição dos termos e criar tópicos a partir disso. As técnicas de detecção de tendências emergentes são baseadas em contagem e conseguem identificar aumentos súbitos no aparecimento de termos em determinados períodos. Por fim, as técnicas de reconhecimento de padrões procuram aprender o comportamento dos tópicos ou termos e auxiliar na predição de tais comportamentos. Algumas técnicas utilizadas nos artigos não se encaixam em nenhum destes três grupos principais e não foram contabilizadas porque foram utilizadas com o objetivo de auxiliar as técnicas principais. Por exemplo, as técnicas de séries simuladas que serviram como forma de suavizar séries temporais e aumentar a taxa de predição dos modelos em alguns trabalhos.

A distribuição das técnicas pode ser visualizada nas figuras 7, 8 e 9.

Nota-se, pelas figuras, que as técnicas estão bastante distribuídas pelos trabalhos. Entretanto, algumas poucas técnicas fogem desse padrão. São elas LDA com Gibbs Sampling e Clustering que são utilizadas pela maioria dos trabalhos, como mostrado na figura 2. Em relação às técnicas de reconhecimento de padrões há um ponto a se ressaltar. Alguns dos autores decidiram utilizar determinadas técnicas pela capacidade de se adaptar a fluxos de dados grandes e dinâmicos com uma complexidade computacional relativamente baixa. É a explicação, por exemplo, da utilização de KNN em alguns dos trabalhos.

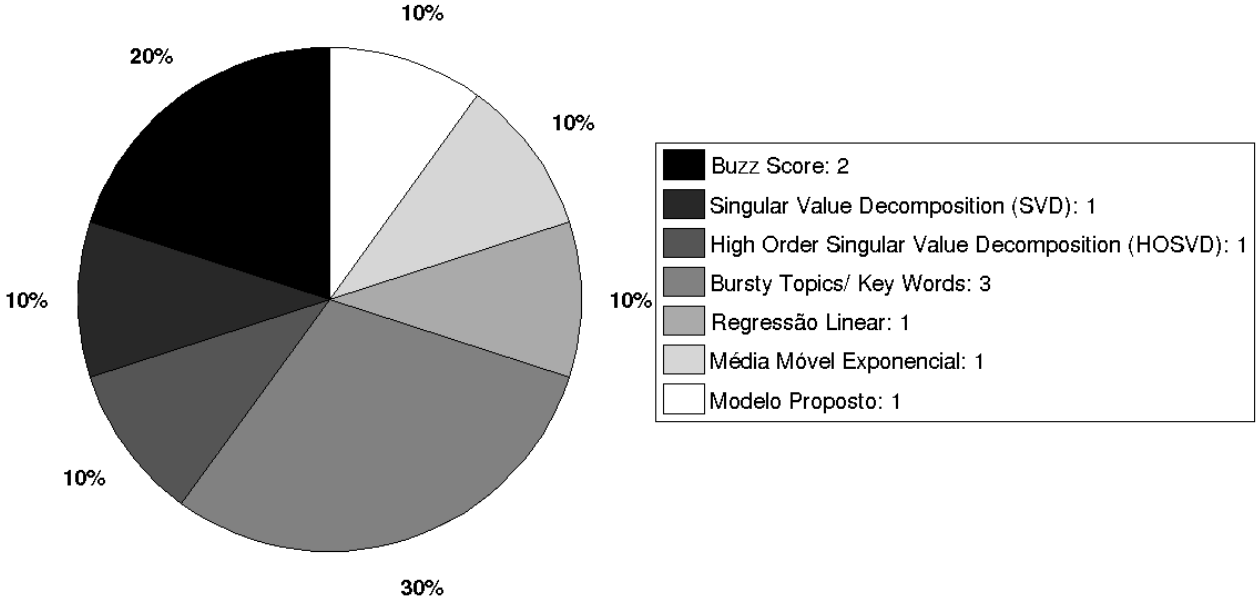
Os artigos que usam LDA com Gibbs Sampling foram os artigos de Bolelli et al. (2009), Dey et al. (2009), Kawamae (2010), Kawamae (2012) e Martie et al. (2012). No

Figura 7 – Gráfico de distribuição das técnicas de identificação de tópicos



Fonte: Caio Cesar Trucolo, 2014

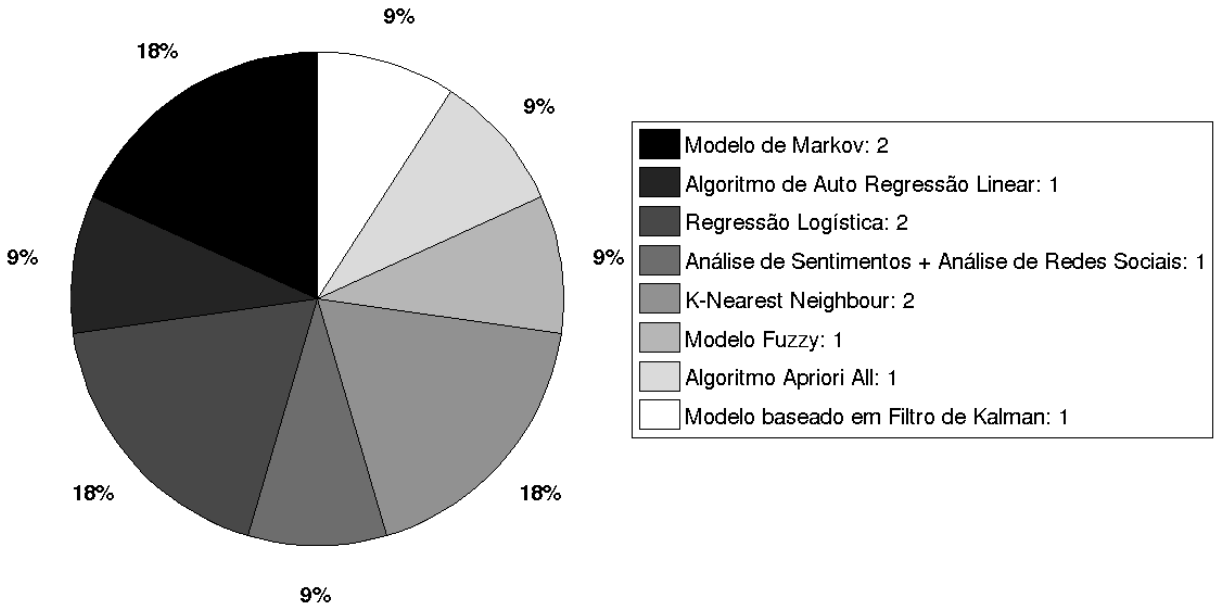
Figura 8 – Gráfico de distribuição das técnicas de tendências emergentes



Fonte: Caio Cesar Trucolo, 2014

trabalho de Bolelli et al. (2009), LDA e Gibbs Sampling foram utilizados conjuntamente com a ordem temporal dos documentos para a criação de um modelo generativo que aprende as distribuições de autor, tópico e palavra. Em uma aplicação sintética, houve uma taxa de acerto de aproximadamente 72%. Dey et al. (2009) propuseram um modelo de detecção de eventos utilizando também técnicas de agrupamento (*clustering*) para entender a correlação

Figura 9 – Gráfico de distribuição das técnicas de reconhecimento de padrões



Fonte: Caio Cesar Trucolo, 2014

dos eventos com a variação do mercado financeiro. Os autores apresentam alguns gráficos nos quais é possível perceber estas correlações. O trabalho de Kawamae (2010) consistiu em tentar predizer as distribuições dos tópicos em artigos científicos levando o tempo em consideração. Com a mesma ideia, Kawamae (2012) ainda estabeleceu uma diferença entre tópicos estáveis (não possuem variação significativa ao longo do tempo) e tópicos dinâmicos, tentando rebater outros modelos que apenas levam em consideração as explosões de tópicos (aumentos súbitos de aparecimento de tópicos em determinados períodos). Com uma aplicação diferenciada, Martie et al. (2012) apresentaram uma abordagem de identificação de tendências nas discussões sobre falhas em um fórum sobre o desenvolvimento de um dos projetos de código aberto da plataforma Android. Com as relações apresentadas entre as tendências identificadas e as atualizações das versões do Android foi possível especular sobre a correção de alguns erros durante essas atualizações. Chen et al. (2013) utilizaram apenas LDA para identificar tópicos e realizaram manualmente uma verificação para analisar quais dos tópicos eram tendências.

Técnicas de agrupamento (*clustering*) para identificação de tópicos foram utilizadas nos trabalhos de Al Bawab et al. (2012), Cvijikj et al. (2011), Pervin et al. (2013) e Voigt et al. (2013). Al Bawab et al. (2012) apresentaram um modelo que identifica os termos mais buscados pelos usuários utilizando *buzz score*; que determina por afinidade a localização espacial desses termos utilizando entropia; e agrupa esses termos em tópicos levando

em consideração a taxa de busca e a localização. O objetivo foi melhorar a ferramenta Trending Now do Yahoo!. O trabalho gerou um aumento de 5,91% de cliques nos tópicos por localização em relação apenas aos tópicos globais. No artigo de Cvijikj et al. (2011), textos de status do Facebook foram tidos como aplicação. Os termos mais importantes foram identificados por TF-IDF e posteriormente agrupados. Com a análise, foi estabelecida por eles uma classificação dos tópicos em eventos disruptivos, tópicos populares e rotinas diárias. Pervin et al. (2013) utilizaram uma abordagem invertida na qual eles primeiro agruparam as palavras hierarquicamente para então utilizar a técnica de *burst topic* e encontrar os tópicos mais citados no Twitter. Voigt et al. (2013) apresentaram um modelo que apenas identifica tópicos a partir da análise de notícias. Kaleel et al. (2013) utilizaram LSH para fazer agrupamento. Diferentemente dos anteriores, Gollapudi e Sivakumar (2004) propuseram um modelo que não necessita de reorganização dos dados em grande escala. Todo o processo de agrupamento é feito em uma etapa de pré-processamento com a utilização de métricas de atributo adequadas, selecionadas pela aproximação realizada por algoritmos de árvores métricas e algum conhecimento prévio sobre o domínio. A solução proposta utiliza o modelo de Markov para a previsão de tendências.

Barbagallo et al. (2011) e Mathioudaki e Koudas (2010) utilizaram técnicas de *bursty topics / key words* para identificar os tópicos mais citados no Twitter em períodos determinados. O segundo integrou um algoritmo guloso para a identificação de tópicos para que a complexidade do modelo fosse baixa, se adaptando com mais facilidade a um fluxo alto de dados.

Abe e Tsumoto (2009) selecionaram termos de importância por TF-IDF e coeficiente de Jaccard, que consiste em uma medida de comparação de similaridade de conjuntos de amostras, utilizando regressão linear posteriormente para detectar tendências emergentes. Todas as tendências detectadas foram confirmadas como tendências reais por especialistas do domínio.

Golbandi et al. (2013) e Kamath e Caverlee (2011) também utilizaram técnicas de detecção de tendências emergentes. Golbandi et al. (2013) apresentaram um modelo com um algoritmo de auto regressão linear para prever o comportamento das buscas dos usuários que serviu como entrada para o algoritmo de *buzz score*, que identifica as tendências emergentes. Já Kamath e Caverlee (2011) propuseram um modelo em que "frases" (conjunto de um ou mais termos) recebem pontuações para cada tempo e então se determina um limiar para indicar o que é e o que não é tendência.

Zhu et al. (2011) propuseram um algoritmo para a interpretação do comportamento da variação de preços de telefones celulares em algumas lojas virtuais. Park et al. (2011) apresentaram uma abordagem baseada em I-GI para seleção de atributos de tópicos escolhidos e utilizaram SVM para avaliação. Utilizando *F1-measure*, o resultado chegou a 98%.

No trabalho de Jayashri e Chitra (2012) foi proposto um modelo com rede ART para identificar tópicos e detectar tendências através dos picos dos tópicos extraídos de documentos científicos.

Os trabalhos que utilizaram técnicas para análise de séries temporais foram os de Christiansen et al. (2012), Heshan e Qingshan (2008), Huang et al. (2011), Kato et al. (2010), Teixeira e Oliveira (2009), Wu et al. (2012), Yonghong e Wenyang (2009) e Gong e Sun (2009). Christiansen et al. (2012) apresentaram uma abordagem que consiste em segmentar séries temporais de acordo com variações bruscas, agrupar esses segmentos por similaridade e descrever cada tópico como um conjunto desses segmentos. Por fim, um modelo de Markov foi utilizado para fazer a previsão de tendência dos tópicos no mês seguinte e foi obtida uma média de acerto de 92,9% para uma parte da base de dados do Google Trends[2].

Heshan e Qingshan (2008) utilizaram uma técnica de séries simuladas chamada *Ascending Triangle* para a entrada do algoritmo KNN, enquanto Huang et al. (2011) utilizaram PPO como uma técnica de suavização de séries temporais PPO com um modelo Fuzzy para predição. Kato et al. (2010) utilizaram EMA para detectar os altos e baixos das taxas de câmbio e alimentaram um algoritmo genético com essa informação para otimizar o ganho de compra e venda de moedas. Teixeira e Oliveira (2009) propuseram um modelo com um algoritmo de KNN para prever os preços de ações cujas informações de entrada foram obtidas de ferramentas técnicas de mercado financeiro. No final, foi comparado o resultado do modelo proposto com a estratégia de mercado chamada *Buy and Hold*. Simulando transações financeiras, observou-se que a estratégia proposta obteve um melhor resultado para compra e venda em 12 das 15 ações do experimento.

Wu et al. (2012) utilizaram K-médias (*K-means*) para agrupar segmentos de séries temporais e o algoritmo AprioriAll para buscar por padrões. Yonghong e Wenyang (2009) propuseram um modelo que adota a matriz de ganho de Kalman para calcular automaticamente a estimativa de probabilidade máxima de um fluxo de dados. Por fim,

[2] http://www.google.com.br/trends/

Gong e Sun (2009) utilizaram regressão logística para a previsão de tendências em mercado de ações.

Além de todos os modelos apresentados, existem ainda abordagens que levam em consideração a estrutura das conexões entre as fontes das informações analisadas e a importância individual delas nessa estrutura. Chi et al. (2006) compuseram três vetores de informação para a análise de tendências: vetor de tendências (aspecto temporal), vetor de autoridade (importância das fontes de informação) e vetor de hub (modelagem das fontes de informação como um estrutura de grafo). A técnica HOSVD foi utilizada e os resultados mostraram que o acréscimo da informação de localização das fontes em relação à estrutura de conexões aumenta a precisão do modelo. Sha et al. (2012) realizaram um estudo sobre tendências em um aplicativo específico de recomendação de fotos. Os usuários do aplicativo foram classificados entre as classes *trend makers* e *trend spotters* e a importância e a relação entre eles foram consideradas para a detecção de tendências. Por fim, regressão linear foi utilizada para prever se uma determinada foto será ou não tendência baseada nos usuários relacionados à foto.

Gloor et al. (2009) desenvolveram um conceito chamado *Web buzz index* para tópicos minerados por diversas fontes na internet. Esse índice é baseado inteiramente em técnicas de análise de redes sociais incluindo análise de sentimentos. O gráfico do Web buzz index foi plotado juntamente a uma série temporal financeira e pôde ser observado que as duas têm um comportamento bastante parecido.

Alguns artigos analisados tinham como foco apenas classificar de alguma forma as informações de tendências em redes sociais digitais. Cheong e Lee (2009) utilizaram rede SOM para verificar a distribuição dos *trending topics* do Twitter em relação à localização dos usuários. Khan et al. (2012a) propuseram classificar *tweets* que realmente sinalizavam um surto de doenças em alguma localidade utilizando *Naive Bayes* para classificação. Khan et al. (2012b) também utilizaram Naive Bayes, mas com o objetivo de colocar *tags* automáticas em tweets. Zubiaga et al. (2011) utilizaram SVM para classificar os tweets em notícias, eventos correntes, *memes* ou comemorações.

3.3 Considerações Finais da Revisão

Nesta revisão os métodos foram categorizados de acordo com sua função exercida com o objetivo comum de identificação de tendências. Ficou claro que grande parte dos

trabalhos consiste da experimentação de intercalação de técnicas em diferentes etapas do processo. Além disso, é importante considerar que o domínio das aplicações também influi na utilização de técnicas diferentes. Os métodos utilizados em séries temporais, por exemplo, se diferenciam bastante das técnicas utilizadas em aplicações de documentos históricos.

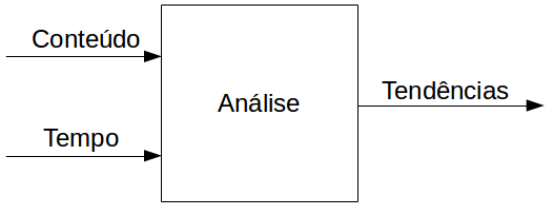
Uma dificuldade ainda existente nessa área é relacionada a forma de avaliação das técnicas. Ainda não existe um padrão de avaliação que facilite a comparação de resultados entre os trabalhos e isso dificulta a determinação se um método leva vantagem sobre outro em relação a alguma variável global. Outro problema, como já discutido, é a replicação dos experimentos. Grande parte dos trabalhos utiliza dados dinâmicos para os experimentos, como dados do Twitter, e não é possível obter os mesmos dados para replicar os experimentos. Apenas em 23% dos artigos analisados na revisão é possível fazer a replicação dos experimentos. Uma alternativa para solucionar esse problema seria os autores armazenarem e disponibilizarem esses dados de alguma forma para que outras pessoas conseguissem utilizar os mesmo dados.

Por fim, de todos os trabalhos analisados, ainda são poucos os que utilizam a estrutura das fontes de informação como variável nos experimentos. Há, ainda, um grande campo a ser explorado sobre a utilização da teoria de redes sociais na detecção e análise de tendências.

4 Especificação, Desenvolvimento e Resultados

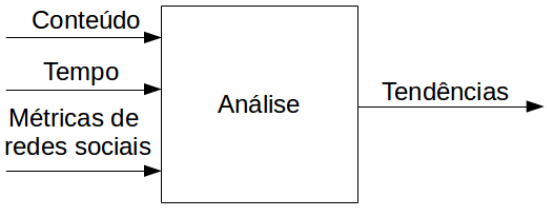
A partir dos estudos realizados na área de análise de tendências e dos trabalhos explorados na revisão sistemática, um plano metodológico básico foi desenhado para a realização dos experimentos em todas as fases de desenvolvimento do projeto. De forma geral, a construção do modelo pode ser dividida em duas fases: a fase inicial que consiste da análise de tendências baseada apenas no conteúdo e tempo e a fase posterior com a inclusão do fator de redes sociais. As figuras 10 e 11 ilustram as duas fases do projeto.

Figura 10 – Modelo de análise da primeira fase de análise de tendências do projeto



Fonte: Caio Cesar Trucolo, 2015

Figura 11 – Modelo de análise da segunda fase de análise de tendências do projeto

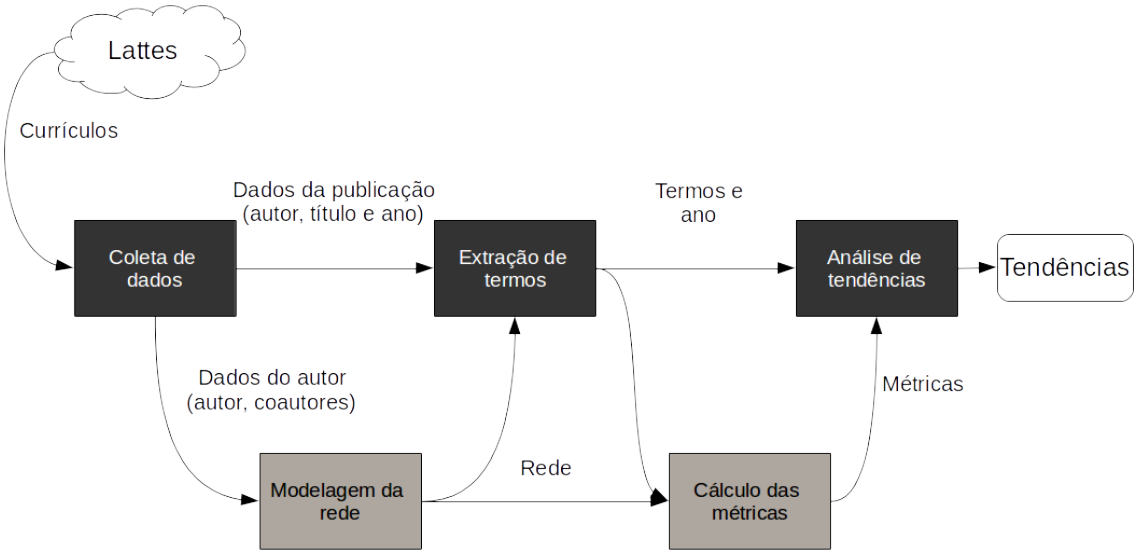


Fonte: Caio Cesar Trucolo, 2015

Os procedimentos utilizados como base para os experimentos iniciais e finais, que serão discutidos ao longo deste capítulo, são ilustrados pela figura 12. As etapas ilustradas pelas caixas mais claras são exclusivas do modelo final, já incluindo os fatores de redes sociais.

Nas próximas seções cada uma das etapas será detalhada.

Figura 12 – Esquema gráfico dos procedimentos básicos de todo o processo de análise de tendências



Fonte: Caio Cesar Trucolo, 2015

4.1 Coleta e Base de dados

A coleta de dados foi baseada no processo detalhado por Digiampietri et. al. (2012). Os currículos extraídos da plataforma Lattes foram organizados e armazenados em um banco de dados relacional. Dos currículos foram extraídas informações sobre publicações, autoria e coautoria, áreas de pesquisa, orientação, bancas, etc. Para essa dissertação foram utilizados as informações das publicações (título e ano de publicação) e autores (número de identificação). Os títulos foram escolhidos para a análise por servirem como um resumo do conteúdo das publicações. Os títulos das publicações também são usados por outros trabalhos de análise de tendências voltados a artigos científicos (Kawamae, 2010, Kawamae, 2012 e Abe e Tsumoto, 2009).

4.2 Extração automática de termos

Com o intuito de automatizar a preparação dos dados para a análise de tendências, foi implementada uma técnica de extração automática de termos. A abordagem utilizada, baseada em Nakagawa e Mori (2002), consiste no reconhecimento automático de termos

específicos do domínio a partir da frequência adjacente das palavras que formam o termo (simples ou composto).

No modelo desenvolvido, os títulos foram partidos em todos os conjuntos possíveis de termos simples e compostos de forma adjacente (termos candidatos) tendo as pontuações e a lista de *stop words* como separadores. Como exemplo, o título *Social Network Analysis for Digital Media*, teria o seguinte conjunto de termos: *Social, Network, Analysis, Digital, Media, Social Network, Network Analysis, Digital Media, Social Network Analysis.*

Para o cálculo da pontuação de cada termo candidato, a seguinte fórmula foi utilizada:

$$FED\,(TC) = f\,(TC) \times \left(\prod_{i=1}^{T} (FE\,(P_i) + 1) \times (FD\,(P_i) + 1) \right)^{1 \div T} > 1$$

em que $FED(TC)$ é a frequência esquerda-direita do termo candidato, $f(TC)$ é a frequência do termo candidato TC, e $FE(Pi)$ e $FD(Pi)$ indicam a frequência das palavras da esquerda e da direita, respectivamente.

Os termos selecionados foram os termos cujas frequências esquerda-direita (FED) fossem maior que 1. A pontuação maior que 1 indica que a frequência adjacente das palavras que compõe o termo possui uma relevância mínima.

Com os experimentos, observou-se, assim como também discutido por Nakagawa e Mori (2002), que os termos compostos tinham um melhor significado semântico do domínio em análise. Como o objetivo do estudo principal é identificar as tendências de pesquisa para as publicações, optou-se por se analisar apenas os termos compostos. A tabela 1 mostra alguns dos termos extraídos automaticamente para o domínio da área de ciência da computação.

A maioria dos termos extraídos são da língua inglesa porque a maior parte das publicações são nessa língua. Termos em português e espanhol também foram extraídos.

4.3 Análise histórica

Análise de tendências nesta dissertação, como apresentado na segunda seção do capítulo 2, possui um primeiro passo de análise histórica dos termos extraídos. O conjunto de documentos dessa análise histórica é composta pelos termos extraídos, pelos carimbos de tempo (ano de publicação) e um índice de importância para medir a importância ou popularidade do termo ao longo do tempo. O índice TF-IDF foi escolhido por ser um

Tabela 1 – Amostra de termos extraídos para a área de Ciência da Computação

Termo
neural network
genetic algorithm
information system
software development
web service
image segmentation
markov model
product line
sensor network
real-time system

Fonte: Caio Cesar Trucolo, 2015

dos índices mais utilizados para aferir a importância de termos (Abe e Tsumoto, 2009). Esse índice indica quão importante um termo é dentro de todo o conjunto de documentos levando em consideração a frequência dele dentro de cada documento e no conjunto como um todo. Por exemplo, se um termo aparece em poucos documentos mas nesses documentos ele tem uma frequência alta, para esses documentos ele vai ter grande importância, ou seja, um TF-IDF alto. Já um termo que aparece em muitos documentos mas com uma frequência parecida em cada documento, é um termo pouco importante e que possui um baixo TF-IDF.

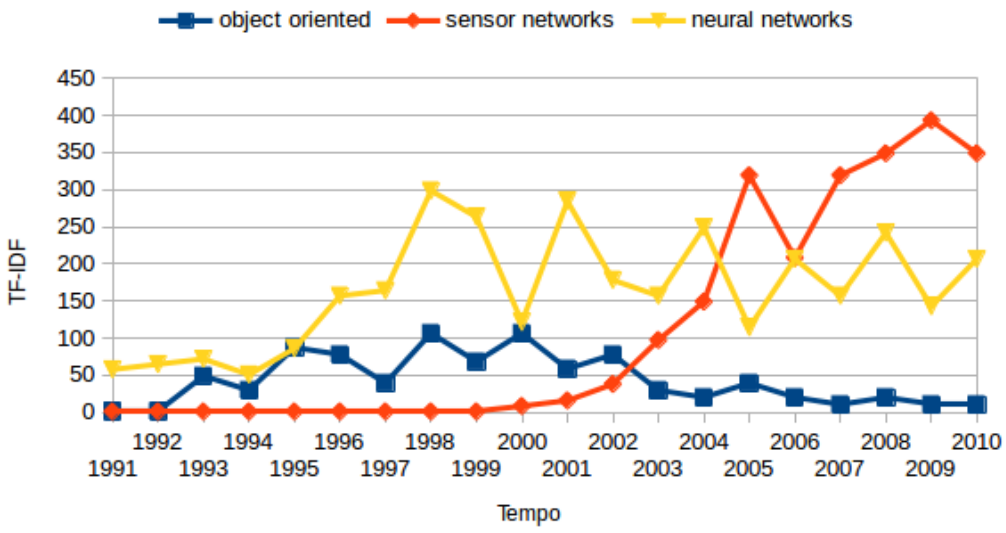
No caso específico desse trabalho, os títulos de artigos científicos publicados em cada ano assumem o papelo de ”documento”e o conjunto total s ao os títulos de artigos científicos de todos os anos.

A partir dessas informações, séries temporais foram modeladas. A figura 7 mostra o comportamento temporal de três termos. Na figura pode-se ver um claro comportamento de ascensão para o termo *sensor networks* e uma situação de queda a partir do ano de 2002 para o termo *object oriented*. Já o termo *neural networks* possui um comportamento de oscilação entre os anos mas não indica sinais de queda ou ascensão na parte final da série.

O passo seguinte à construção dos modelos é a previsão do índice de importância TF-IDF. Nos primeiros experimentos foram considerados apenas os fatores temporais e as técnicas de regressão foram utilizadas para essa tarefa. Análises de regressão seguem o formato

$$Y \approx f(X, \beta),$$

Figura 13 – Comportamento temporal de três termos no período de 1991 a 2011



Fonte: Caio Cesar Trucolo, 2015

em que a variável dependente Y pode ser aproximada pelas variáveis independentes X e seus respectivos parâmetros β para determinada função f. Nas análises de regressão desse trabalho a variável dependente é o índice TF-IDF do termo e a variável independente é o tempo (os anos de publicação dos artigos).

O método de mínimos quadrados foi utilizado para determinar as curvas de tendência que mais se adequavam às séries temporais de cada termo. Os tipos de regressão utilizados foram linear, exponencial, logarítmica, do tipo *power law* e polinomial de grau 2 a 5. Posteriormente, foi calculado o erro quadrático para cada curva de tendência gerada para se determinar a curva mais adequada à série temporal de cada termo.

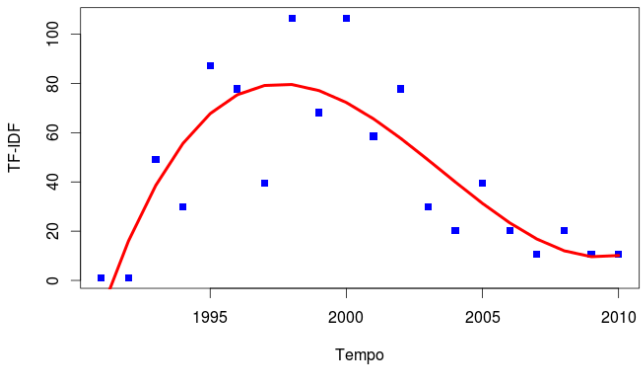
As figuras 14, 15 e 16 apresentam exemplos de curvas geradas pelas regressões mais adequadas a cada série temporal.

Essa forma de análise que considera apenas o fator temporal e os índices de importância dos termos foi o primeiro passo do trabalho que serviu de base para os experimentos posteriores.

4.4 Análise de redes sociais

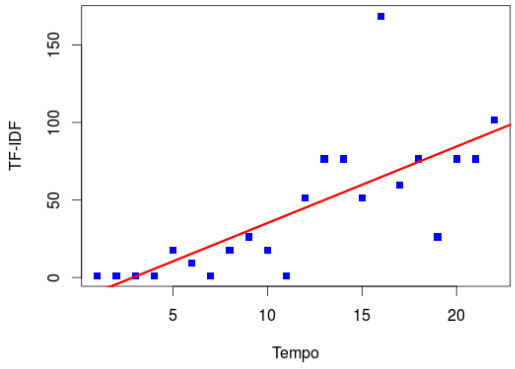
A análise de redes sociais é a forma de se estudar a estrutura de conexão entre indivíduos e grupos, e o comportamento social como um todo. A hipótese deste mes-

Figura 14 – Curva gerada pela regressão não linear polinomial de grau 3 para o termo *object oriented*



Fonte: Caio Cesar Trucolo, 2015

Figura 15 – Curva gerada pela regressão linear para o termo *comunicação científica*

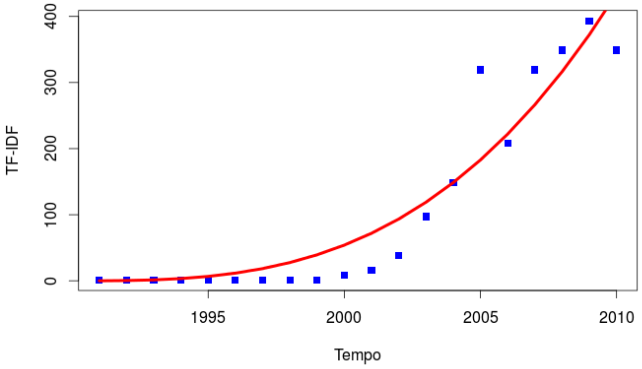


Fonte: Caio Cesar Trucolo, 2015

trado é que incluir as informações extraídas das análises de redes sociais na análise de comportamento temporal de conteúdos textuais resultará em uma maior aproximação da realidade, ou seja, uma maior acurácia nas previsões. O comportamento dos indivíduos e por consequência os conteúdos gerados por eles, como publicações acadêmicas, têm alta probabilidade de sofrerem influências sociais. Dessa forma, o comportamento temporal desses conteúdos podem ser melhor explicados tendo métricas de redes sociais como um fator adicional. Essas métricas de redes sociais serão adicionadas como novas características dentro do conjunto de dados.

O processo de seleção das métricas foi baseado em suposições sobre a capacidade de explicação de difusão de informação de cada uma delas. Por exemplo, uma das suposições é

Figura 16 – Curva gerada pela regressão não linear tipo *power law* para o termo *sensor network*



Fonte: Caio Cesar Trucolo, 2015

que um nó (indivíduo) inserido dentro de um componente gigante, que é o maior componente conexo da rede (no qual todos os indivíduos são ligados direta ou indiretamente), tem uma maior capacidade de disseminar informação do que um nó que não está inserido. As métricas selecionadas são: presença ou não do nó no componente gigante, comprimento do caminho mais curto para o nó mais central, centralidade de grau, centralidade *eigenvector*, centralidade *page rank*, centralidade *betweenness*, centralidade *closeness*, coeficiente de clusterização, equivalência estrutural com o nó mais central e centralidade média da comunidade. As métricas de centralidade explicam quão importante é o nó, dentro do contexto de influência e difusão de informação dentro da rede, a métrica de caminho mais curto indica quão distante está o nó do nó mais importante da rede e a métrica de equivalência estrutural mostra quão similar é o nó em relação ao nó mais importante da rede.

A utilização do nó mais importante de cada rede como uma referência é justificada pelo destaque que eles têm em relação aos outros. Para a rede social modelada para todos os doutores do Brasil na área de Ciência da Computação, o nó mais central têm um aumento de grau (número de vizinhos) de quase 49% em relação ao segundo nó com maior número de graus da rede. A tabela 2 mostra a comparação de grau e equivalência eigenvector para os primeiros dez nós mais centrais.

Após a escolha das métricas para análise, cada nó tem todas as métricas calculadas. Desse passo, resulta-se um arquivo com as métricas de todos os nós da rede. Além disso, é necessário identificar cada nó que utilizou cada termo extraído no período de análise.

Tabela 2 – Centralidade *eigenvector* e grau dos nós mais importantes da rede

Nós mais importantes	Centralidade *Eigenvector*	Grau
1	1.000	67
2	0.986	45
3	0.944	45
4	0.845	31
5	0.825	35
6	0.799	29
7	0.798	37
8	0.763	24
9	0.745	30
10	0.744	27

Fonte: Caio Cesar Trucolo, 2015

Dessa forma, também é montada uma base em que um nó pode ter utilizado zero, um ou mais termos e um termo pode ter sido utilizado por um ou mais nós.

Com o resultado das métricas para cada nó e a relação de nós e termos, as métricas de todos os nós são concatenadas para cada termo. Entretanto, em vez de construir as variáveis apenas com a soma das métricas dos nós individualmente, os nós foram agrupados a partir de comunidades identificadas dentro da rede. Dentro de comunidades a difusão da informação tem grande probabilidade de ser muita rápida tornando a informação um conhecimento geral de todos os membros. Não levar as comunidades em consideração pode induzir ao erro no sentido de se observar uma boa difusão de informação em toda a rede sendo que esta mesma difusão pode estar acontecendo dentro de uma única ou poucas comunidades. Identificando e agrupando os nós em comunidades faz com que a análise da difusão leve em conta toda a rede.

Existem diversos métodos para a identificação de comunidades. Nesse trabalho, optou-se por utilizar um método eficiente para uma grande quantidade de nós. Clauset et. al. (2004) propôs um método hierárquico aglomerativo em que cada nó inicia sendo uma única comunidade. De modo iterativo, o algoritmo determina a melhor composição local de nós que otimizam a função de modularidade (uma função que mede qualidade das comunidades) e os agrupa. Essa iteração acontece até que essa função não melhore mais. Ao final da iteração, as comunidades formadas são o resultado.

A especificação de cada variável é descrita a seguir.

1. Componente gigante: indica se o nó pertence ou não ao componente gigante;

2. Caminho mais curto para o nó mais central: Menor valor entre os caminhos mais curtos da comunidade para o nó mais central;
3. Centralidade de grau: Centralidade de grau média dos nós pertencentes à comunidade;
4. Centralidade eigenvector: Centralidade eigenvector média dos nós pertencentes à comunidade;
5. Centralidade Page Rank: Centralidade Page Rank média dos nós pertencentes à comunidade;
6. Centralidade Betweenness: Centralidade betweenness média dos nós pertencentes à comunidade;
7. Centralidade Closeness: Centralidade closeness média dos nós pertencentes à comunidade;
8. Coeficiente de clusterização: Valor médio de coeficiente de clusterização dos nós pertencentes à comunidade;
9. Equivalência estrutural com o nó mais central: Valor médio da equivalência estrutural com o nó mais central dos nós pertencentes à comunidade;
10. Centralidade média da comunidade: Centralidade média de todos os nós da comunidade.

Essa forma de se trabalhar com as métricas, que é a geração de um conjunto de dados a partir de um período de análise, pode causar erros de interpretação baseado na dinâmica da rede. Em um período longo de análise as métricas podem indicar uma tendência que pode já não ser mais considerada como tal. Termos e expressões, por exemplo, podem ter sido tendências no início do período de análise mas não mais no final do período. Com isso, foi iniciado um método de inclusão da dinâmica nessa metodologia que consiste em calcular as métricas para dois períodos adjacentes (2002 a 2006 e 2007 a 2011, por exemplo) e tirar a média ponderada das métricas de cada termo com um peso maior para os períodos mais atuais.

Além das variáveis de redes sociais, uma outra variável também é utilizada na formação dos conjuntos de dados para alguns experimentos, como será discutido na próxima seção. A variável é a *Previsão histórica*, que é o resultado da previsão de TF-IDF de cada termo feita com os modelos de regressão contendo apenas as informações de comportamento histórico. Essa variável, portanto, representa a parte de análise histórica dentro do conjunto de dados.

4.5 Experimentos

O principal objetivo deste trabalho é verificar se informações oriundas de redes sociais ajudam a explicar o comportamento temporal de conteúdos textuais gerados por pesquisadores. Assim, os experimentos contemplaram análises considerando e não considerando as informações de redes sociais. Diferentes conjuntos de dados foram gerados para testar períodos de análise, seleção de atributos e técnicas de previsão.

Os experimentos principais foram elaborados para prever o índice de importância TF-IDF para o ano de 2012, que é o ano seguinte as séries históricas analisadas, e dessa forma avaliar as melhores configurações para os modelos. Dados os experimentos principais, experimentos posteriores tiveram como meta averiguar a capacidade da abordagem desenvolvida para previsões de médio e longo prazo.

Para todos os experimentos, a área de estudo foi delimitada para Ciência da Computação. Foram utilizados currículos de 5.642 doutores no Brasil contendo 55.710 títulos de publicações para o período entre 1991 e 2012.

4.5.1 Conjuntos de dados

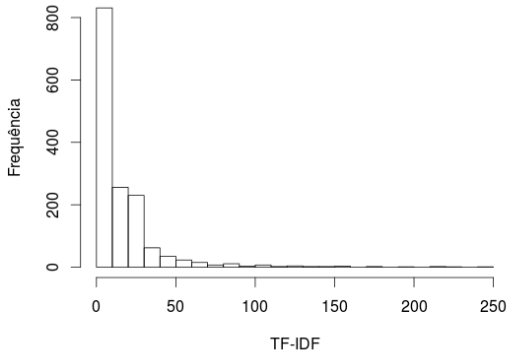
Para os experimentos foi construído um total de 16 conjuntos de dados diferentes. Os conjuntos de dados foram construídos baseados nos períodos de análise e seleção de atributos e se diferenciavam por esses mesmos fatores. O último conjunto foi elaborado para os testes de previsão de médio e longo prazo.

O primeiro período de análise trabalhado está contido entre 1991 e 2011. Esse período foi definido após a análise de número de publicações cadastradas para a área de Ciência da Computação. O número de publicações a partir de 1991 foi considerado adequado para a execução dos experimentos e após 2012 o número de publicações caia drasticamente por causa da não atualização dos currículos que levam certo tempo para serem atualizados pelos autores. Períodos de 10 anos (2002 a 2011) e 5 anos (2007 a 2011) também foram experimentados para se verificar se o período poderia influenciar nos resultados de previsão. Além disso, como descrito na seção anterior, também foram utilizados conjuntos de dados preparados para modelar a dinâmica das redes sociais.

Antes de se iniciar propriamente as análises de regressão, foram realizadas análises das variáveis independentes e da variável dependente. A análise mostrada nessa seção

possui apenas o conjunto de dados do período entre 2007 a 2011 tendo como resultados os TF-IDFs de 2012. 1500 termos foram usados para o conjunto de dados. A primeira variável a ser analisada foi a dependente, ou seja, os resultados reais para cada termo. O histograma da figura 17 mostra a distribuição do resultado real em vinte faixas de valores que vai do mínimo (zero) ao máximo (248,05). Pode-se notar a alta quantidade de valores baixos dessa variável.

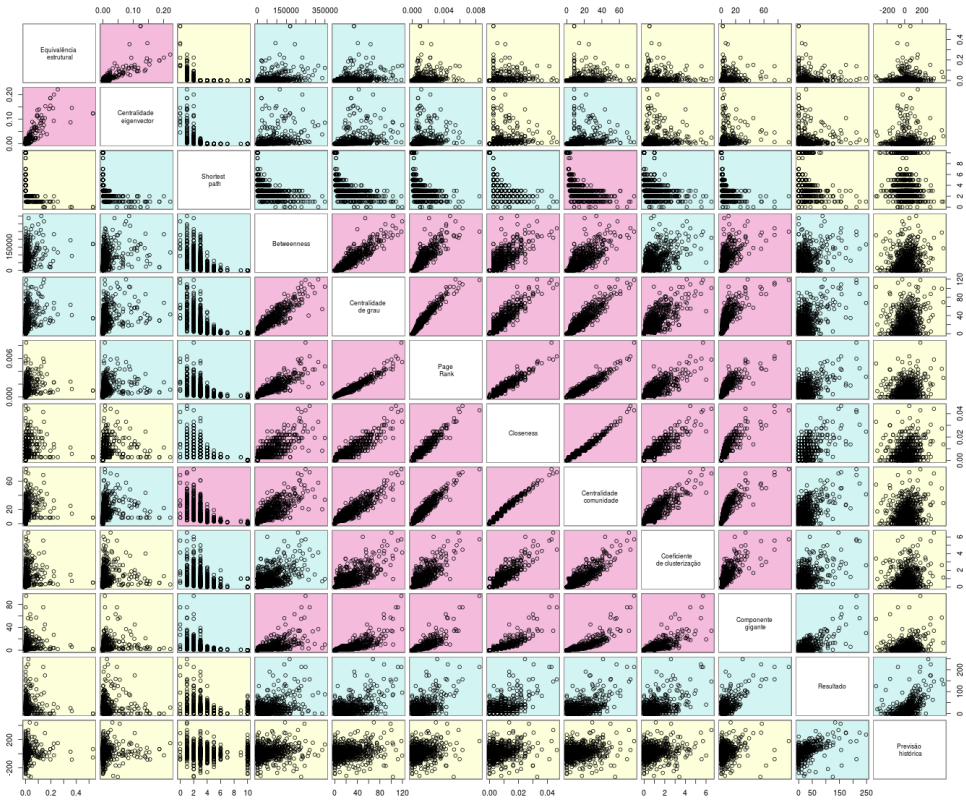
Figura 17 – Histograma da distribuição do resultado real



Fonte: Caio Cesar Trucolo, 2015

Em relação às variáveis preditoras, uma das primeiras análises que podem ser realizadas é a análise de correlação. Além da análise de distribuição das variáveis de forma individual, muitas vezes torna-se interessante analisar o relacionamento entre variáveis do conjunto. A correlação entre variáveis indica o grau de relacionamento das variáveis, ou seja, a similaridade do comportamento das mesmas. Assim, todas as variáveis foram comparadas com cada outra variável pelo gráfico de dispersão da figura 18. Diferentes níveis de correlação são apresentados pelas variáveis. Se por um lado as variáveis *Closeness* e *Centralidade média* da comunidade possuem uma correlação extremamente alta, por outro lado, as variáveis *Shortest path* e *Previsão histórica* possuem uma correlação quase nula. É interessante notar que a maioria das variáveis apresentou um nível médio de correlação com o resultado real, mas nenhuma apresentou uma nível alto. Em uma análise mais simplista é possível perceber que o treinamento de modelos de previsão com todas as variáveis não é interessante. Inserir em um mesmo modelo diferentes variáveis que possuem alta correlação muito provavelmente não elevará o poder de predição.

Figura 18 – Gráfico de dispersão de todas as variáveis do conjunto de dados. Cada gráfico de dispersão indica o nível de correlação para as variáveis: a cor amarela indica níveis baixos de correlação, a cor verde indica níveis médios de correlação enquanto a cor vermelha indica altos níveis de correlação.



Fonte: Caio Cesar Trucolo, 2015

Na regressão linear, seja ela múltipla ou simples, o objetivo consiste em determinar os pesos das variáveis independentes para se obter o melhor resultado de previsão possível da variável dependente. Esses pesos determinam a importância relativa das variáveis no modelo de previsão geral. Essa técnica também facilita identificar a influência de cada variável no modelo geral. Quando há mais de uma variável independente, é necessário entender a colinearidade entre essas variáveis independentes e não apenas com a variável dependente. Portanto, existem métodos que procuram identificar o melhor conjunto de variáveis que melhor predizem a variável dependente. *Stepwise estimation* é um dos métodos que procura determinar o melhor grupo de variáveis preditoras. Esse método pode ser classificado como *wrapper*, ou seja, método de seleção que usa o resultado do modelo como avaliador do subconjunto de atributos sendo testado. Entretanto, para outros tipos de técnicas de previsão, como as que serão apresentadas na seção seguinte, esse tipo de método de seleção pode ser inviável por tornar todo o processo computacionalmente muito

oneroso. Dessa forma, *Relief* foi escolhido para a seleção. O *Relief* é uma técnica do tipo filtro, que diferentemente das técnicas do tipo *wrapper*, determinam o conjunto de variáveis independentes antes do treinamento dos modelos de previsão. Além disso, a seleção manual também foi considerada por ser um bom método de seleção quando há certo conhecimento sobre os atributos do conjunto.

Na etapa de preparação de dados, é necessário entender a magnitude dos valores nulos no conjunto de dados e os possíveis problemas que seriam gerados por eles e então decidir se é preciso algum tipo de ajuste. Para este trabalho, como as variáveis foram criadas a partir da análise de redes sociais, os valores nulos não se tornaram um problema.

4.5.2 Técnicas de previsão

Foram escolhidas três técnicas de previsão para os experimentos: RNA, SVM e *Rotation Forest*. Estas técnicas foram treinadas utilizando dados "do passado" (utilizando diferentes intervalos de tempo) para tentar predizer os valores "do futuro". Alguns parâmetros dessas técnicas foram testados e combinados resultando em um total de 16 testes para RNA, 9 testes para SVM e 15 testes para *Rotation Forest*.

Para RNA, os parâmetros alterados foram a taxa de aprendizado (com os valores 0,01, 0,05, 0,1 e 0,5) e o termo *momentum* (com os valores 0, 0,01, 0,05 e 0,1).

As técnicas de SVM utilizadas foram específicas para problemas de regressão. Os testes foram realizados para diferentes *kernels* e o parâmetro alterado para cada um deles foi apenas a variável de complexidade C, que por linhas gerais, determina a quantidade aceitável de erro no treinamento de cada exemplo. Os *kernels* utilizados foram os *kernels* polinomial e polinomial normalizado, PUK e RBF.

Nos testes com *Rotation Forest*, diferentes implementações voltadas a regressão foram utilizadas e parâmetros específicos de cada implementação foram alterados. Os parâmetros específicos das implementações seguem o padrão dos parâmetros das técnicas de árvores de decisão, como, por exemplo, a opção pela poda. A primeira implementação usada foi *Random Forest*, que teve parâmetros de profundidade máxima da árvore e o número de árvores geradas alterados. A segunda implementação usada foi M5P que tiveram parâmetros de opção por poda e suavização de predição assim como o número mínimo de instâncias permitidas para cada folha. As implementações de *Random Tree* e aprendizado

por árvore de decisão rápida tiveram como únicos parâmetros alterados a quantidade de dados utilizados para realizar *backfitting*.

4.5.3 Métodos de avaliação

Em todos os experimentos, as técnicas de previsão foram avaliadas pelo método de validação cruzada dividindo o conjunto de entrada em 10 subconjuntos. As medidas de erro utilizadas em todos os experimentos foram MAE e RAE. MAE é uma medida absoluta simples e bastante utilizada. O MAE é calculado por:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|,$$

sendo f_i o valor predito e y_i o valor atual.

RAE é uma medida alternativa para resultados em termos percentuais. RAE é dada por:

$$RAE = \frac{\sum_{i=1}^{n}|f_i - y_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|},$$

em que:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

4.5.4 Resultados

Antes dos experimentos com os conjuntos de dados que incluem os fatores de redes sociais, foram realizados testes, a título de comparação, apenas com as previsões feitas a partir de regressões lineares e não lineares dos conjuntos de dados dos períodos de 1991 a 2011, 2002 a 2011 e 2007 a 2011. Os resultados para esses testes podem ser vistos na tabela 3.

Tabela 3 – Resultados de regressão (linear ou não linear) para os três períodos

	MAE	**RAE**
1991-2011	*16,7626*	*113,16%*
2002-2011	21,3832	136,42%
2007-2011	45,7179	288,14%

Fonte: Caio Cesar Trucolo, 2015

Também para comparação foram realizados testes de predição com os algoritmos de *Rotation Forest*, SVM e RNA apenas com as informações de TF-IDF para cada ano. Para esses experimentos, o ano foi modelado como um atributo sendo o ano de previsão o resultado real. A diferença desse conjunto de experimentos em relação ao conjunto anterior são apenas as técnicas utilizadas, já que as técnicas utilizadas anteriormente podem ser classificadas como paramétricas porque a forma de relacionamento funcional entre a variável dependente e as variáveis independentes são conhecidas/estimadas enquanto as técnicas utilizadas para esse conjunto de experimentos carecem de conhecimento *a priori* sobre a forma da função sendo estimada. Os períodos utilizados foram os mesmos: 1991 a 2011, 2002 a 2011 e 2007 a 2011, tendo o ano de 2012 como o ano de previsão. Os resultados obtidos podem ser vistos na tabela 4. Observa-se uma diminuição da medida RAE em relação aos resultados apresentados na tabela 3.

Tabela 4 – Resultados de regressão (RAE) para as técnicas de SVM, RNA e Rotation Forest com os dados de entrada tendo como atributos os resultados de TF-IDF para cada ano.

	SVM	**RNA**	**Rotation Forest**
1991-2011	51,80%	59,29%	51,52%
2002-2011	51,01%	57,59%	*51,46%*
2007-2011	*50,31%*	*56,81%*	51,74%

Fonte: Caio Cesar Trucolo, 2015

Nas tabelas 5 e 6, são resumidos os melhores resultados, para MAE e RAE, respectivamente, de cada técnica de previsão para os conjuntos de dados que incluem os fatores de redes sociais nos períodos de 1991 a 2011, 2002 a 2011 e 2007 a 2011 além dos conjuntos construídos com o fator de dinâmica para os períodos adjacentes de 1991 a 2001 e 2002 a 2011, e também de 2002 a 2006 e 2007 a 2011.

O melhor desempenho (menor erro), como mostrado pelas tabelas 5 e 6, foi de 39,28% para o período de 2007 a 2011. Por outro lado, ao se calcular o erro médio de todas as técnicas para os diferentes períodos, os melhores desempenhos foram para o período de dez anos (2002 a 2011) como indicado pela tabela 7. Em uma primeira comparação de erro entre os resultados dos modelos paramétricos que não incluem o fator de redes sociais e os modelos que incluem (tabelas 3 e 5) houve reduções de erro de 42%, 70% e 85% para os períodos 1991 a 2011, 2002 a 2011 e 2007 a 2011, respectivamente.

Tabela 5 – Melhores resultados (MAE) para cada técnica de previsão

	Todos atributos	**Manual**	**Relief**
Reg. Linear 91-11	10,6985	-	-
RNA 91-11	10,8129	10,8848	10,8864
SVM 91-11	9,6660	*9,2213*	9,5084
Rot. Forest 91-11	10,5613	10,4245	10,5146
Reg. Linear 02-11	8,3497	-	-
RNA 02-11	7,1630	7,2386	7,2351
SVM 02-11	6,8066	6,4436	6,4816
Rot. Forest 02-11	6,3238	*6,2705*	6,2869
Reg. Linear 07-11	10,0095	-	-
RNA 07-11	9,2179	9,4090	9,2453
SVM 07-11	8,5012	8,3782	8,1839
Rot. Forest 07-11	6,6707	*6,2973*	6,4593
Reg. Linear dinâmico 91-11	15,0664	-	-
RNA dinâmico 91-11	14,4809	14,2334	14,1755
SVM dinâmico 91-11	13,5834	13,2614	*12,8769*
Rot. Forest dinâmico 91-11	13,4493	13,9017	13,6378
Reg. Linear dinâmico 02-11	18,2478	-	-
RNA dinâmico 02-11	18,0614	18,0584	17,8593
SVM dinâmico 02-11	*17,1096*	17,1915	17,1579
Rot. Forest dinâmico 02-11	17,6621	17,9282	17,7184

Fonte: Caio Cesar Trucolo, 2015

O comportamento entre os melhores resultados do modelo não paramétrico que não utiliza dados de redes sociais e do modelo proposto pode ser melhor observado pela figura 19. O modelo proposto só não obteve menores erros para o período entre 1991 e 2011. Sendo melhor para os períodos seguintes.

Em relação as técnicas, a que se sobressaiu foi *Rotation Forest* com os dois melhores resultados. Pelos experimentos, é possível notar que *Rotation Forest* obteve melhores desempenhos para períodos mais curtos de análise. Já SVM teve bons desempenhos para todos os períodos sendo até melhor que *Rotation Forest* para o período entre 1991 e 2011.

Observando as tabelas 5 e 6, observa-se que a utilização dos conjuntos de dados com todos os atributos tiveram os piores resultados. Já a seleção manual superou os resultados da seleção com *Relief*.

A tabela 8 mostra as variáveis selecionadas pelos métodos *Relief* e Manual.

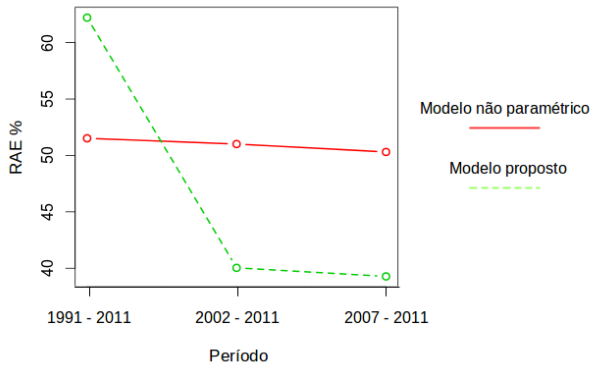
A tabela 9 compara os resultados de 15 termos para os modelos simples de regressão (sem o fator de redes sociais) e o modelo proposto (com o fator de redes sociais). Os 15

Tabela 6 – Melhores resultados (RAE) para cada técnica de previsão

	Todos atributos	**Manual**	**Relief**
Reg. Linear 91-11	72,18%	-	-
RNA 91-11	72,95%	73,43%	73,44%
SVM 91-11	65,21%	*62,21%*	64,15%
Rot. Forest 91-11	71,51%	70,57%	71,18%
Reg. Linear 02-11	53,22%	-	-
RNA 02-11	45,75%	46,25%	46,23%
SVM 02-11	43,45%	41,05%	41,30%
Rot. Forest 02-11	40,29%	*40,04%*	40,12%
Reg. Linear 07-11	62,76%	-	-
RNA 07-11	57,77%	59,02%	57,98%
SVM 07-11	53,31%	52,37%	51,22%
Rot. Forest 07-11	41,68%	*39,28%*	40,33%
Reg. Linear dinâmico 91-11	58,04%	-	-
RNA dinâmico 91-11	55,72%	54,99%	54,62%
SVM dinâmico 91-11	52,32%	51,42%	*49,63%*
Rot. Forest dinâmico 91-11	53,65%	55,73%	54,67%
Reg. Linear dinâmico 02-11	60,42%	-	-
RNA dinâmico 02-11	59,86%	59,83%	59,21%
SVM dinâmico 02-11	*56,68%*	57,26%	56,98%
Rot. Forest dinâmico 02-11	58,75%	59,61%	58,85%

Fonte: Caio Cesar Trucolo, 2015

Figura 19 – Comportamento dos erros para os melhores resultados dos experimentos do modelo proposto e do modelo não paramétrico de regressão sem a inclusão dos fatores de redes sociais



Fonte: Caio Cesar Trucolo, 2015

termos utilizados foram identificados como as principais tendências para 2012 a partir da previsão dos modelos simples de regressão. Para os testes com o modelo proposto, foi

Tabela 7 – Média de melhores resultados RAE para cada período

Período	**Média RAE**
1991-2011	69,68%
2002-2011	*43,77%*
2007-2011	51,57%

Fonte: Caio Cesar Trucolo, 2015

Tabela 8 – Variáveis selecionadas pelos métodos *Relief* e manual

Relief	**Seleção manual**
Componente gigante; Centralidade eigenvector; Centralidade Betweenness; Coeficiente de clusterização e Previsão de regressão.	Componente gigante; Centralidade eigenvector; Coeficiente de clusterização; Equivalência estrutural e Previsão de regressão.

Fonte: Caio Cesar Trucolo, 2015

utilizada a configuração que gerou o menor erro como observado nas tabelas 5 e 6, ou seja, *Rotation Forest* no período de 2007 a 2011 com seleção manual.

Tabela 9 – Comparação dos resultados dos modelos para as 15 principais tendências identificadas pelo modelo simples de séries temporais para o ano de 2012

Termo	**Obtido**	**Séries-para-métrico**	**Erro**	**Séries-não para-métrico**	**Erro**	**Mo-delo pro-posto**	**Erro**
service discovery	135,17	441,52	306,35	58,43	76,74	123,39	11,77
based approach	155,19	424,16	268,97	249,40	94,21	161,10	5,91
information systems	147,32	334,29	186,97	182,08	34,76	148,37	1,05
supply chain	174,31	298,37	124,06	145,71	28,6	143,96	30,35
web services	225,28	297,74	72,46	190,14	35,14	201,05	24,23
product line	174,99	291,57	116,57	481,68	306,69	154,73	20,26
motion estimation	107,78	274,36	166,58	174,73	66,95	99,00	8,78
social network	249,05	269,42	20,38	327,70	78,65	198,94	50,11
business process	131,75	240,09	108,34	264,25	132,5	119,61	12,14
time series	150,79	217,76	66,97	196,08	45,29	147,03	3,76
neural network	213,36	178,86	34,51	565,81	352,45	198,85	14,51
sign language	108,21	176,83	68,62	76,97	31,24	101,69	6,52
são paulo	191,93	172,84	19,09	71,51	120,42	145,79	46,15
genetic programming	128,25	156,64	28,39	104,18	24,07	107,98	20,26
routing problem	101,11	147,16	46,05	195,61	94,5	83,75	17,36

Fonte: Caio Cesar Trucolo, 2015

Pela tabela 9, fica evidente o ganho de predição entre as duas abordagens. O erro gerado pelo modelo com fator de redes sociais corresponde a apenas 17% do erro total

gerado pelo modelo paramétrico de regressão e a 18% do erro total gerado pelo modelo não paramétrico.

Tabela 10 – Flutuação das posições das principais tendências em curto, médio e longo prazo

Posição	**2012**	**2015**		**2020**	
1	web service	neural network	2	neural network	0
2	social network	social network	0	social network	0
3	neural network	web service	-2	web service	0
4	based approach	information system	2	based approach	1
5	product line	based approach	-1	product line	2
6	information system	business process	4	business process	0
7	time series	product line	-2	supply chain	1
8	são paulo	supply chain	1	information system	-3
9	supply chain	service discovery	1	service discovery	0
10	service discovery	genetic programming	2	são paulo	1
11	business process	são paulo	-2	sign language	3
12	genetic programming	routing problem	3	routing problem	0
13	sign language	motion estimation	1	genetic programming	-3
14	motion estimation	sign language	-1	time series	1
15	routing problem	time series	-8	motion estimation	-2

Fonte: Caio Cesar Trucolo, 2015

Em um passo posterior a avaliação dos modelos e a escolha dos melhores a partir dos erros e estatísticas relacionadas, passa-se a utilizar de fato as previsões geradas. Além de previsões imediatas, as análises podem ser feitas para períodos de médio e longo prazo. A tabela 10 mostra a flutução de posições dos mesmos 15 termos analisados anteriormente.

A fim de se verificar a qualidade da abordagem proposta por esse mestrado para além das previsões a curto prazo, ou seja, as previsões para o ano seguinte, os últimos experimentos foram executados com um corte de análise histórica mais no passado, mais especificamente, no ano de 2005. Assim, o treinamento foi executado para o período entre 1991 e 2005 e as previsões foram realizadas para os anos de 2006, 2007, 2008, 2009, 2010 e 2011. Os algoritmos de previsão utilizados para esses testes foram apenas SVM e *Rotation Forest* com as configurações que geraram os menores erros para os primeiros experimentos.

A tabela 11 mostra os resultados obtidos para esses experimentos. Como esperado, os erros aumentam conforme o intervalo entre o ano de predição e o período de treinamento, com exceção do ano de 2010. Entretanto, o erro não aumenta de forma drástica para os períodos posteriores a um ano. O comportamento dos erros ao longo dos anos se aproxima de um formato linear. Comparando novamente os resultados da tabela 3 com a abordagem

Tabela 11 – Erros RAE para testes de previsão de médio e longo prazo com modelo treinado para o período entre os anos de 1991 e 2005

Técnica	2006	2007	2008	2009	2010	2011
Rot. Forest	60,65%	62,25%	64,21%	65,54%	64,31%	67,87%
SVM	55,91%	57,43%	57,77%	58,70%	57,28%	59,29%

Fonte: Caio Cesar Trucolo, 2015

proposta os erros continuam bem menores, mesmo para testes que não são para o ano seguinte.

5 Conclusões

O presente trabalho apresentou uma nova proposta para problemas de previsão de tendências científicas utilizando séries temporais e métricas de redes sociais. O objetivo foi analisar e desenvolver um modelo computacional para análise de tendências tendo como novo fator agregado métricas de redes sociais. A hipótese trabalhada foi de que as métricas de rede ajudariam a explicar o comportamento temporal das produções científicas nacionais, aumentando a acurácia da predição de tendências. Essa hipótese foi confirmada a partir dos experimentos que mostraram uma diminuição de erro de até 85% para o modelo proposto em relação apenas a predição com técnicas de séries temporais.

Além da confirmação da hipótese para resultados de curto prazo, ou seja, os experimentos para o ano seguinte ao treinamento dos modelos, a hipótese continua válida para situações com intervalos entre o período de treinamento e teste.

Alguns problemas específicos também foram explorados dentro das etapas de construção do modelo. O primeiro foi sobre quais métricas de redes sociais utilizar e a melhor forma de inseri-las em um modelo de predição. Em seguida, já na fase de experimentação, foi necessário identificar o melhor período para resultados mais acurados. Diferentes períodos foram testados e os resultados analisados.

O desenvolvimento consistiu inicialmente de uma divisão do problema em duas partes: análise histórica e considerando informações da análise de redes sociais. Todo o trabalho baseado nas técnicas de séries temporais estava contido na análise histórica. Já a modelagem da rede, extração das métricas e construção das variáveis foi realizada na etapa de redes sociais. Posteriormente, os dois resultados foram agregados para dar início aos experimentos. Nessa etapa, diferentes técnicas de inteligência artificial foram utilizadas para realizar as previsões.

Além da apresentação de uma nova abordagem para previsão de tendências, as principais contribuições desse trabalho também foram a revisão sistemática sobre análise de tendências e publicação de resultados de previsão de tendências para segmentos acadêmicos contidos na base da plataforma Lattes.

5.1 Publicações

O presente trabalho contribui na publicação de cinco artigos científicos.

1. TRUCOLO, C. C.; DIGIAMPIETRI, L. A.. Análise de Tendências da Produção Científica Nacional da Área de Ciência da Computação. *Revista de Sistemas de Informação da FSMA*, v. 14, p. 2-10, 2014. Artigo que aplica a análise de tendências a produção científica brasileira em ciência da computação, sem considerar as informações oriundas da rede social acadêmica.
2. TRUCOLO, C. C.; DIGIAMPIETRI, L. A.. Análise de tendências da produção científica nacional na área de Ciência da Informação: estudo exploratório de mineração de textos. *AtoZ: novas práticas em informação e conhecimento*, v. 3, p. 87-94, 2014. Artigo que aplica a análise de tendências a produção científica brasileira em ciência da informação, sem considerar as informações oriundas da rede social acadêmica.
3. DIGIAMPIETRI, L. A.; ALVES, C. M. ; Trucolo, C. ; Oliveira; R.. Análise da Rede dos Doutores que Atuam em Computação no Brasil. *III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2014)*, Anais do XXXIV Congresso da Sociedade Brasileira de Computação (CSBC 2014), 2014. Artigo que utiliza a análise de redes sociais para quantificar algumas características da rede brasileira de doutores que atuam na área de ciência da computação.
4. DIGIAMPIETRI, L. A.; ALVES, C. M.; TRUCOLO, C. C.; VALDIVIA-DELGADO, K.; MUGNAINI, R.. Análise da Rede de Relacionamentos dos Doutores Brasileiros. *VIII Brazilian e-Science Workshop (BreSci 2014)*, Anais do XXXIV Congresso da Sociedade Brasileira de Computação (CSBC 2014), 2014. Artigo que utiliza a análise de redes sociais para quantificar algumas características da rede brasileira de doutores que atuam no Brasil.
5. TRUCOLO, C. C. ; DIGIAMPIETRI, L. A. . Uma Revisão Sistemática acerca das Técnicas de Identificação e Análise de Tendências. *X Simpósio Brasileiro de Sistemas de Informação*, p. 639-650, 2014. Artigo que apresenta a revisão sistemática desenvolvida durante este mestrado.

Ainda não foram publicados os resultados obtidos combinando a análise de tendências tradicional com informações oriundas da análise da rede social acadêmica.

5.2 Trabalhos Futuros

Pontos de aprimoramento foram levantados e discutidos ao longo do desenvolvimento do projeto. O primeiro deles é referente à análise de redes sociais de forma dinâmica. Cada rede possui certas características, mas grande parte delas possui comportamento dinâmico. Apesar do trabalho apresentar uma abordagem para essa situação, muito ainda pode ser feito para melhorar esse modelo. Já referente a etapa de mineração de texto, uma forma de agrupamento em tópicos dos termos extraídos pode ser feito para melhorar a qualidade semântica da aplicação dos modelos.

Previsão de tendências faz parte de um grupo maior de trabalhos recentes relacionados a análise de grandes quantidades de dados. Entretanto, a inserção do fator de redes sociais nesse contexto ainda é mínima. Esse trabalho, portanto, possui um papel de inicialização. Os resultados, como visto, são promissores. O modelo proposto pode ser utilizado e analisado por especialistas da área de computação ou mesmo aplicado para problemas gerais por especialistas de outras áreas.

Referências[1]

ABE, H.; TSUMOTO, S. Evaluating a method to detect temporal trends of phrases in research documents. In: *2009 8th IEEE International Conference on Cognitive Informatics*. IEEE, 2009. p. 378–383. ISBN 978-1-4244-4642-1. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5250711>.

BARBAGALLO, D. et al. Semi-automated Methods for the Annotation and Design of a Semantic Network Designed for Sentiment Analysis of Social Web Content. In: *2011 22nd International Workshop on Database and Expert Systems Applications*. IEEE, 2011. p. 222–226. ISBN 978-1-4577-0982-1. ISSN 1529-4188. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6059821>.

BAWAB, Z. A.; MILLS, G. H.; CRESPO, J.-F. Finding trending local topics in search queries for personalization of a recommendation system. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. New York, New York, USA: ACM Press, 2012. p. 397. ISBN 9781450314626. Disponível em: <http://dl.acm.org/citation.cfm?id=2339530.2339594>.

BOLELLI, L. et al. Finding topic trends in digital libraries. In: *Proceedings of the 2009 joint international conference on Digital libraries - JCDL '09*. New York, New York, USA: ACM Press, 2009. p. 69. ISBN 9781605583228. Disponível em: <http://dl.acm.org/citation.cfm?id=1555400.1555411>.

CHEN, L.; ZHANG, C.; WILSON, C. Tweeting Under Pressure: Analyzing Trending Topics and EvolvingWord Choice on SinaWeibo. In: *Proceedings of the first ACM conference on Online social networks - COSN '13*. New York, New York, USA: ACM Press, 2013. p. 89–100. ISBN 9781450320849. Disponível em: <http://dl.acm.org/citation.cfm?id=2512938.2512940>.

CHEONG, M.; LEE, V. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In: *Proceeding of the 2nd ACM workshop on Social web search and mining - SWSM '09*. New York, New York, USA: ACM Press, 2009. p. 1. ISBN 9781605588063. Disponível em: <http://dl.acm.org/citation.cfm?id=1651437-.1651439>.

CHI, Y.; TSENG, B. L.; TATEMURA, J. Eigen-Trend: Trend Analysis in the Blogosphere based on Singular Value Decompositions. In: *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*. New York, New York, USA: ACM Press, 2006. p. 68. ISBN 1595934332. Disponível em: <http://dl.acm.org/citation.cfm?id=1183614.1183628>.

CHRISTIANSEN, L. et al. Modeling topic trends on the social web using temporal signatures. In: *Proceedings of the twelfth international workshop on Web information and data management - WIDM '12*. New York, New York, USA: ACM Press, 2012. p. 3. ISBN 9781450317207. Disponível em: <http://dl.acm.org/citation.cfm?id=2389936.2389940>.

CLAUSET, A.; MARK, E. N.; MOORE, C. Finding community structure in very large networks. *Physical review E*, 2004.

[1] De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

CVIJIKJ, I. P.; MICHAHELLES, F. Monitoring Trends on Facebook. In: *2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing*. IEEE, 2011. p. 895–902. ISBN 978-1-4673-0006-3. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6118886>.

DEY, L.; MAHAJAN, A.; HAQUE, S. M. Document Clustering for Event Identification and Trend Analysis in Market News. In: *2009 Seventh International Conference on Advances in Pattern Recognition*. IEEE, 2009. p. 103–106. ISBN 978-0-7695-3520-3. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4782752>.

DIGIAMPIETRI, L. et al. Análise da rede dos doutores que atuam em computação no brasil. In: *CSBCBraSNAM 2014*. Brasília: [s.n.], 2014. Citado 3 vezes nas páginas 25, 26 e 27.

FELDMAN, R.; SANGER, J. *The Text Mining Handbook*. [S.l.]: Cambridge, 2007.

GLOOR, P. A. et al. Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. In: *2009 International Conference on Computational Science and Engineering*. IEEE, 2009. v. 4, p. 215–222. ISBN 978-1-4244-5334-4. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5284145>.

GOLBANDI, N. G. et al. Expediting search trend detection via prediction of query counts. In: *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*. New York, New York, USA: ACM Press, 2013. p. 295. ISBN 9781450318693. Disponível em: <http://dl.acm.org/citation.cfm?id=2433396.2433435>.

GOLLAPUDI, S.; SIVAKUMAR, D. Framework and algorithms for trend analysis in massive temporal data sets. In: *Proceedings of the Thirteenth ACM conference on Information and knowledge management - CIKM '04*. New York, New York, USA: ACM Press, 2004. p. 168. ISBN 1581138741. Disponível em: <http://dl.acm.org/citation-.cfm?id=1031171.1031208>.

GONG, J.; SUN, S. A New Approach of Stock Price Prediction Based on Logistic Regression Model. In: *2009 International Conference on New Trends in Information and Service Science*. IEEE, 2009. p. 1366–1371. ISBN 978-0-7695-3687-3. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5260596>.

GUAN, H.; JIANG, Q. Pattern matching of time series and its application to trend prediction. In: *2008 2nd International Conference on Anti-counterfeiting, Security and Identification*. IEEE, 2008. p. 41–44. ISBN 978-1-4244-2584-6. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4688342>.

HAN, M. K. . J. P. J. *Data Mining: Concepts and Techniques*. [S.l.]: Morgan Kaufmann, 2011.

HUANG, H.; PASQUIER, M.; QUEK, C. Financial Market Trading System With a Hierarchical Coevolutionary Fuzzy Predictive Model. *IEEE Transactions on Evolutionary Computation*, v. 13, n. 1, p. 56–70, fev. 2009. ISSN 1089-778X. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4769012>.

JAYASHRI, M. et al. Topic clustering and topic evolution based on temporal parameters. In: *2012 International Conference on Recent Trends in Information*

Technology. IEEE, 2012. p. 559–564. ISBN 978-1-4673-1601-9. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6206816>.

KALEEL, S. B.; ALMESHARY, M.; ABHARI, A. Event detection and trending in multiple social networking sites. In: *Proceeding CNS '13 Proceedings of the 16th Communications & Networking Symposium Article No. 5*. Society for Computer Simulation International, 2013. p. 5. ISBN 978-1-62748-031-4. Disponível em: <http://dl.acm.org/citation.cfm?id=2499986.2499991>.

KAMATH, K. Y.; CAVERLEE, J. Discovering trending phrases on information streams. In: *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, New York, USA: ACM Press, 2011. p. 2245. ISBN 9781450307178. Disponível em: <http://dl.acm.org/citation.cfm?id=2063576.2063937>.

KARANIKAS, H.; THEODOULIDIS, B. *Knowledge Discovery in Text and Text Mining Soft- ware*. [S.l.], 2002.

KATO, D.; YATA, N.; NAGAO, T. Evolutionary trend prediction using plural technical indicators for foreign exchange transaction. p. 1170–1175, 2010.

KAWAMAE, N. Theme Chronicle Model: Chronicle Consists of Timestamp and TopicalWords over Each Theme. In: *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*. New York, New York, USA: ACM Press, 2012. p. 2065. ISBN 9781450311564. Disponível em: <http://dl.acm.org/citation.cfm?id=2396761.2398573>.

KAWAMAE, N.; HIGASHINAKA, R. Trend detection model. In: *Proceedings of the 19th international conference on World wide web - WWW '10*. New York, New York, USA: ACM Press, 2010. p. 1129. ISBN 9781605587998. Disponível em: <http://dl.acm.org/citation.cfm?id=1772690.1772838>.

KHAN, M. A. H.; IWAI, M.; SEZAKI, K. A robust and scalable framework for detecting self-reported illness from twitter. In: *2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2012. p. 303–308. ISBN 978-1-4577-2040-6. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper-.htm?arnumber=6379425>.

KHAN, M. A. H.; IWAI, M.; SEZAKI, K. Towards urban phenomenon sensing by automatic tagging of tweets. In: *2012 Ninth International Conference on Networked Sensing (INSS)*. IEEE, 2012. p. 1–7. ISBN 978-1-4673-1786-3. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6240529>.

KITCHENHAM, B. The problem with function points. *IEEE Software*, v. 14, n. 2, p. 29, mar. 1997. Citado na página 28.

KONTOSTATHIS, A.; GALITSKY, L.; POTTENGER, W. A survey of emerging trend detection in textual data mining. *Survey of Text ...*, p. 1–44, 2004. Disponível em: <http://link.springer.com/chapter/10.1007/978-1-4757-4305-0_9>.

LEMIEUX, V.; OUIMET, M. *Análise Estrutural das Redes Sociais*. [S.l.]: Instituto Piaget, 2008. 128 p. ISBN 9727719333. Citado na página 23.

MARTIE, L. et al. Trendy bugs: Topic trends in the Android bug reports. In: *2012 9th IEEE Working Conference on Mining Software Repositories (MSR)*. IEEE, 2012. p. 120–123. ISBN 978-1-4673-1761-0. ISSN 2160-1852. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6224268>.

MATHIOUDAKIS, M.; KOUDAS, N. TwitterMonitor: Trend Detection over the Twitter Stream. In: *Proceedings of the 2010 international conference on Management of data - SIGMOD '10*. New York, New York, USA: ACM Press, 2010. p. 1155. ISBN 9781450300322. Disponível em: <http://dl.acm.org/citation.cfm?id=1807167.1807306>.

MENA-CHALCO, J.; DIGIAMPIETRI, L.; JR., R. C. Caracterizando as redes de coautoria de currículos lattes. In: *CSBC 2012 - BraSNAM ()*. [s.n.], 2012. Disponível em: <http://XXXXX/99923.pdf>.

PARK, H. et al. Detection and Analysis of Trend Topics for Global Scientific Literature Using Feature Selection Based on Gini-Index. In: *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*. IEEE, 2011. p. 965–969. ISBN 978-1-4577-2068-0. ISSN 1082-3409. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper-.htm?arnumber=6103457>.

PERVIN, N. et al. Fast, Scalable, and Context-Sensitive Detection of Trending Topics in Microblog Post Streams. *ACM Transactions on Management Information Systems (TMIS)*, ACM, v. 3, n. 4, p. 19, jan. 2013. ISSN 2158-656X. Disponível em: <http://dl.acm.org/citation.cfm?id=2407740.2407743>.

POURKAZEMI, M.; KEYVANPOUR, M. A survey on community detection methods based on the nature of social networks. *Iccke 2013*, Ieee, n. Iccke, p. 114–120, out. 2013. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6682855>.

SHA, X. et al. Spotting Trends: The Wisdom of the Few. In: *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12*. New York, New York, USA: ACM Press, 2012. p. 51. ISBN 9781450312707. Disponível em: <http://dl.acm.org/citation.cfm?id=2365952.2365967>.

SINGH, S. Survey of Various Techniques for Determining Influential Users in Social Networks. n. Iceccn, p. 398–403, 2013.

SONNENWALD, D. Scientific collaboration. *Annual review of information science and technology*, v. 41, n. 1, p. 643–681, 2007.

TEIXEIRA, L. A.; OLIVEIRA, A. L. I. de. Predicting stock trends through technical analysis and nearest neighbor classification. In: *2009 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2009. p. 3094–3099. ISBN 978-1-4244-2793-2. ISSN 1062-922X. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper-.htm?arnumber=5345944>.

VEJLGAARD, H. *Anatomy of a Trend*. [S.l.]: McGraw-Hill, 2008.

VOIGT, M.; ALEYTHE, M.; WEHNER, P. Towards topics-based, semantics-assisted news search. In: *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics - WIMS '13*. New York, New York, USA: ACM Press, 2013. p. 1. ISBN 9781450318501. Disponível em: <http://dl.acm.org/citation.cfm?id=2479787.2479822>.

WU, Y.-P.; WU, K.-P.; LEE, H.-M. Stock Trend Prediction by Sequential Chart Pattern via K-Means and AprioriAll Algorithm. In: *2012 Conference on Technologies and Applications of Artificial Intelligence*. IEEE, 2012. p. 176–181. ISBN 978-1-4673-4976-5. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6395026>.

YONGHONG, Y.; WENYANG, B. Forecasting Model over Random Interval Data Stream Based on Kalman Filter. In: *2009 First International Workshop on Database Technology and Applications*. IEEE, 2009. p. 448–451. ISBN 978-0-7695-3604-0. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5207720>.

ZAFARANI, M. A. A. . H. L. R. *Social Media Mining*. [S.l.]: Cambridge University Press, 2014.

ZHU, Q. et al. Commodities Price Dynamic Trend Analysis Based on Web Mining. In: *2011 Third International Conference on Multimedia Information Networking and Security*. IEEE, 2011. p. 524–527. ISBN 978-1-4577-1795-6. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6103828>.

ZUBIAGA, A. et al. Classifying Trending Topics: A Typology of Conversation Triggers on Twitter Arkaitz. In: *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, New York, USA: ACM Press, 2011. p. 2461. ISBN 9781450307178. Disponível em: <http://dl.acm.org/citation.cfm?id=2063576.2063992>.