em questão
E-ISSN 1808-5245

Algoritmo de classificação de especialistas em áreas na base de currículos Lattes

Fellipe de Melo Chagas
Graduando; Universidade de São Paulo (USP), São Paulo, SP, Brasil;
fellipe.chagas@usp.br

José de Jésus Pérez-Alcázar
Doutor; Universidade de São Paulo (USP), São Paulo, SP, Brasil;
jperez@usp.br

Luciano Antonio Digiampietri
Doutor; Universidade de São Paulo (USP), São Paulo, SP, Brasil;
digiampietri@usp.br

Resumo: Este trabalho propõe um algoritmo para construir um ranking de especialistas na base de currículos Lattes. Para isto, foi elaborado um algoritmo composto por três estágios de processamento: Score Alfa, que analisa os títulos dos documentos baseado na ontologia definida para selecionar os assuntos e importâncias destes assuntos para cada publicação; Score Beta, que analisa a qualidade das publicações utilizando conceitos definidos pela CAPES para beneficiar produções publicadas em veículos mais importantes; e Score Propagated, que analisa a importância de estar bem conectado a demais pesquisadores especialistas propagando o conhecimento obtido. Para o teste de precisão do algoritmo, utilizaram-se dados reais da área de Nanotecnologia.

Palavras-chave: Localização de especialistas. Currículos Lattes. Análise de Redes Sociais. Recuperação da informação.

1 Introdução

O termo ‘especialista’ é usado para se referir a uma pessoa/agente com um alto grau de conhecimento sobre certo assunto (LAPPAS; LIU; TERZI, 2011). A busca por especialistas é uma tarefa difícil, pois suas habilidades e conhecimentos são raros, custosos, amplamente distribuídos, difíceis de qualificar, em constante mudança, variando de níveis, normalmente se encontram culturalmente isolados e são muito procurados (MAYBURY, 2006). Além disso, segundo Lappas, Lui e Terzi (2011), encontrá-los virou um desafio devido à abundância de dados sobre os potenciais especialistas na internet.

Para facilitar o reuso do conhecimento humano e suas experiências, os artefatos de conhecimento precisam estar facilmente acessíveis, enquanto que os dados de recursos humanos têm de estar atualizados, explícitos e transparentes (JANEV, 2009).

Outro fator importante nessa necessidade pela busca de especialistas também está ligado ao novo paradigma, chamado de *Open Innovation*, que foi recentemente adotado por empresas e se constituí, basicamente, pela busca de novas ideias e soluções para o desenvolvimento de novas tecnologias fora das fronteiras das empresas, ou seja, ferramentas que auxiliam nesta busca de especialistas para a solução de problemas estão sendo cada vez mais requisitadas (STANKOVIC, 2010).

Para auxiliar a enfrentar os desafios expostos anteriormente, surgiu a necessidade da combinação do uso de técnicas já existentes para a elaboração de novas técnicas. Os *Expert Finding Systems* (EFS), também conhecidos como *Expertise Location Systems* (ELS), são desenvolvidos para dar auxílio nas seguintes atividades: identificar especialistas; classificá-los de acordo com suas especialidades e níveis de conhecimento; validar a profundidade da especialidade classificada; e recomendar especialistas por um *ranking* ordenado pelas habilidades, experiências, certificações e reputação (MAYBURY, 2006).

A produção intelectual registrada na internet pelos especialistas pode servir como base para avaliar a expertise deles. Esse tipo de problema é chamado de localização de especialistas sem restrições de grafos (LAPPAS, LIU; TERZI, 2011), ou relevância baseada em documentos (WANG et al., 2013). Entretanto, nesse cenário, não são consideradas as relações entre os vários especialistas, por exemplo, a relação de coautoria de um pesquisador com pesquisadores bem ranqueados faz deste pesquisador um potencial especialista em tal assunto. O problema de inferir o nível de conhecimento de um pesquisador usando suas conexões (de coautoria) com outros especialistas é chamado de localização de especialistas com propagação do score (LAPPAS; LIU; TERZI, 2011), ou relevância de autoridade (relevância social) (WANG et al., 2013).

A base de especialistas pode ser interpretada como uma rede social em que os especialistas (entidades), se relacionam a partir de citações, publicações, etc. Redes sociais são normalmente representadas por grafos nos quais os nós são os atores e as arestas são os relacionamentos entre esses atores (FREITAS, 2008).

No caso do Brasil, uma fonte importante de informação é a base de currículos do CNPq ("CV-Lattes") (DIGIAMPIETRI et al., 2012). Nessa base, há um registro das atividades dos pesquisadores brasileiros (ex: artigos, patentes, etc.). Com isso, o objetivo deste artigo é desenvolver um algoritmo de classificação dos membros da rede social formada pelos pesquisadores da base de currículos Lattes do CNPq (DIGIAMPIETRI et al., 2012), categorizando as publicações por assuntos de certa área do conhecimento, graus de importância dos veículos de publicação, além de abordagem baseada na propagação do *score* utilizando a estrutura de relacionamentos de coautoria em publicações, a identificação destes relacionamentos foi realizada utilizando-se a metodologia proposta por Digiampietri et al. (2014). Trata-se, assim, de uma abordagem híbrida que mistura a relevância baseada em documentos e a propagação do *score*.

Esse algoritmo produzirá um ranking de pesquisadores para cada assunto de uma determinada área analisada, ordenados de acordo com o grau de conhecimento no assunto estimado pelo algoritmo.

O restante deste artigo está organizado da seguinte forma: a seção 2 apresenta uma revisão de ferramentas existentes, conceitos básicos associados e alguns trabalhos correlatos; a seção 3 contém a metodologia utilizada, descrevendo cada uma das etapas de processamento do algoritmo desenvolvido; a seção 4 apresenta os resultados resultantes da execução do algoritmo; e, por fim, a seção 5 contém a conclusão a respeito das abordagens e rankings obtidos.

2 Revisão bibliográfica

Esta seção apresenta a revisão da literatura com enfoque em três assuntos: ferramentas existentes para a busca de especialistas; ontologias; e técnicas e algoritmos que serviram de base para a solução proposta.

2.1 Ferramentas existentes para busca de especialistas

Em Maybury (2006) são analisadas algumas ferramentas de *Expert Finding*: TACIT Active NET, AskMe, IDOL K2, Endeca, Recommind, SEE-K e Entopia. Essas ferramentas serviram como base para a especificação e desenvolvimento da ferramenta apresentada neste artigo.

As principais características de destaque dessas ferramentas são descritas a seguir: a análise de importância das palavras por meio do cálculo de frequência dos termos ou conjuntos de termos nos documentos é feita por diversas ferramentas como TACIT ActiveNET e SEE-K; a utilização de listas de termos e palavras-chave como um vocabulário controlado dá-se em ferramentas como o SEE-K; o processamento de documentos, que classifica-os com relação à qualidade da fonte em que foram publicados é feito pela AskMe; a análise de rede social, realizada com a ferramenta ENTOPIA, que cria uma rede social com os empregados se relacionando por meio de tópicos, tendo estes relacionamentos pesos definidos de acordo com o volume de informação encontrada (MAYBURY, 2006).

2.2 Ontologia

Para categorizar as publicações em assuntos de áreas do conhecimento, é necessária a existência de uma ontologia que contenha termos relacionados a esta área de conhecimento em que se deseja realizar a classificação. Assim, esses termos são utilizados como um vocabulário controlado no algoritmo de classificação.

Uma ontologia é definida como uma especificação formal e explícita de uma conceptualização compartilhada, que seja legível para os computadores e representem um contexto do mundo real (STAAB; STUDER, 2009). Para a aplicação teste do algoritmo, utilizou-se de uma ontologia disponível na

literatura (ALUISIO, 2005) para as áreas da Nanotecnologia e Nanociência. A ontologia, inclusive, foi desenvolvida a partir de uma análise linguística e lexicográfica, que aborda tópicos específicos para ontologias por meio da aplicação de ferramentas de software para extração automática de termos a partir de *corpus* diversos, além de utilizar conceitos da área de redes complexas para também auxiliar na extração automática de termos e na definição da própria ontologia.

A primeira versão da ontologia foi baseada num *corpus* volumoso de artigos, livros e resumos de fontes variadas em inglês (ALUISIO, 2005). Ou seja, a ontologia foi criada a partir dos resumos, títulos e palavras-chaves de artigos, tópicos de livros, etc. Esse processo de criação é descrito em (ALUISIO, 2005). Foi feita uma tradução da ontologia para português; tanto a ontologia em inglês quanto a versão em português foram usadas em nosso trabalho.

2.3 Técnicas e algoritmos básicos utilizados na solução proposta

Existem diferentes formas de se localizar um especialista em certos assuntos. Segundo Wang, et al. (2013) as classificações ocorrem:

a) com base em informações autorreveladas, isto é, declaradas pelo candidato a especialista;
b) com base nos documentos de sua autoria;
c) com base na análise de redes sociais ou propagação de autoria.

Como a base CV-Lattes é baseada nas informações autorreveladas pelos autores dos currículos, na nossa proposta foi utilizada a primeira técnica apresentada. A seguir, descreveremos as duas últimas técnicas por serem bastante interessantes do ponto de vista acadêmico.

2.3.1 Localização de especialistas baseada nos documentos

Sistemas tradicionais frequentemente usam técnicas de recuperação de informação para descobrir a expertise de uma grande coleção de textos. Nesse tipo de técnica, o perfil de expertise do candidato é construído pela junção de todos os documentos de sua autoria. Nesse processo, os documentos são

convertidos numa visão lógica formada por termos de indexação (são removidas palavras insignificantes – stop words – e palavras derivadas e convertidas a sua raiz – *stemming*). Depois de dada certa consulta (termos de busca), o modelo de recuperação de informação é empregado, e neste caso, utilizamos como exemplo o modelo de espaço vetorial (SALTON; MCGILL, 1983; BAEZA-YATES; RIBEIRO NETO, 1999), para achar o nível de similaridade entre o documento e a consulta. Nesse processo, o peso dos termos no documento é calculado por meio da técnica TF-IDF (Frequência do Termo – Inversa da Frequência nos Documentos) (SALTON; MCGILL, 1983; BAEZA-YATES; RIBEIRO NETO, 1999).

Balog et al. (2006) definiram duas estratégias para a localização de especialistas a partir de uma coleção de documentos. A primeira, chamada de abordagem centrada no perfil do usuário, modela diretamente a expertise do especialista baseada nos documentos associados a ele. A segunda, chamada de abordagem centrada em documentos, localiza um documento relevante para um tópico de consulta e então acha o especialista associado. Ambas seguem o princípio de que a relevância do contexto textual de um candidato com respeito a uma consulta adiciona relevância à evidência de sua expertise. Nessas estratégias é utilizado um modelo probabilístico (BALOG et al., 2006) no lugar do modelo vetorial de representação de documentos.

Krulwich e Burkley (1996) desenvolveram o ContactFinder, que relaciona as consultas de usuários em boletins eletrônicos a pessoas que podem responde-las baseado no histórico das mensagens desses boletins (KRULWICH; BURKLEY, 1996). O sistema categoriza mensagens e extrai seus tópicos usando um conjunto de heurísticas.

2.3.2 Localização de especialistas baseado na análise de redes sociais

O ranking baseado na analise de enlaces (*links*) trouxe novas perspectivas aos problemas de localização de especialistas. Algoritmos populares de ranking de páginas web tais como *PageRank* (BRIN; PAGE, 1998) e *HITS* (KLEINBERG, 1999) têm sido utilizados para melhorar a identificação e ranqueio de especialistas. Nessa definição, uma nova classe de problemas de localização de

especialistas tem surgido, no qual o grau de expertise de um especialista depende do quão bem conectado este especialista está na rede (social) de especialistas. O *score*, neste caso, é calculado a partir da propagação do *score* dos vizinhos na rede.

Campbell et al. (2003) e Dom et al. (2003) utilizaram a rede gerada pelo serviço de e-mail para refinar sua identificação de especialistas. Nessa rede, cada nó corresponde a uma pessoa e cada aresta dirigida vai de um emissor a um receptor. Os melhores especialistas são aqueles que tendem a receber mais e-mails sobre um determinado tópico. Portanto, pessoas que tem recebido vários e-mails de perguntas são definidos como autoridades ou especialistas, e pessoas que reenviam perguntas a vários especialistas são definidos como concentradores (*hubs*). Assim, o score de autoridade calculado pelo algoritmo *HITS* sobre essa rede pode ser usado para ranquear os indivíduos na rede.

Zhang et al. (2007) utilizaram *PageRank* e *HITS* sobre uma rede social de perguntas e respostas. Nessa rede, um nó representa um usuário e arestas são construídas do usuário que fez o post inicial a todos os que o responderam. Os autores desenvolveram uma fórmula de *score* adaptada do algoritmo *PageRank*, que indica a probabilidade do usuário receber pedidos de resposta de outros usuários na rede. Os autores também usaram uma adaptação do algoritmo *HITS*, desta forma, um bom concentrador ("*hub*") é um usuário que ajudou vários especialistas. Similarmente, uma boa autoridade é um usuário que ajuda vários bons concentradores. Assim, o score de autoridade calculado pelo *HITS* é usado para quantificar a expertise do usuário.

2.3.3 Técnicas de localização de especialistas híbridas

Os sistemas de localização de especialistas recentes misturam as técnicas baseadas em documentos e na análise de redes sociais. Segundo Wang et al. (2013), existem diferentes estratégias de combinação, tais como:

a) combinação linear: método usado em vários trabalhos (MARROCCO et al., 2010; WU, 2011). Nessa estratégia, é assumido que tanto a relevância baseada em conteúdo quanto a baseada em redes sociais determinam um nível de expertise coletivamente e simultaneamente.

Entretanto, um alto ranking gerado utilizando essa estratégia não significa um alto nível tanto em relevância quanto em autoridade. Isto é, um especialista pode ter uma alta relevância e pouca autoridade ou vice-versa e mesmo assim ter um alto ranking;

b) ranking em cascata: este método usa uma sequência de funções de ranking para refinar progressivamente um *score* (WANG et al., 2011);

c) estratégia de expansão: usada em classificação supervisionada para combinar resultados de diferentes classificadores e que, em alguns casos, supera em desempenho à estratégia de combinação linear (TAX, 1997).

Neste trabalho é proposta e desenvolvida uma estratégia de ranking em cascata que usa o algoritmo TF-IDF para encontrar o peso de algum assunto (termo da ontologia) nas publicações (títulos) escritas pelo candidato a especialista. Não foi aplicado um modelo vetorial ou probabilístico para representar o documento já que as relevâncias serão calculadas para um assunto específico e com base nos títulos das publicações. Após o cálculo da soma dos pesos (TF-IDF) do assunto procurado em todas as publicações do candidato a especialista é adicionado o complemento, devido à importância da publicação de acordo com a tabela de conceitos Qualis fornecida pela CAPES, como será apresentado na próxima seção.

O resultado da soma desses *scores* é utilizado como valor inicial do candidato especialista para calcular seu ranking baseado em redes sociais ou score de autoridade. Para isso, foi criada uma rede a partir dos relacionamentos de coautoria entre os candidatos a especialista que permite a propagação de autoridade dependendo da importância do relacionamento.

O modelo de propagação utilizado no algoritmo formulado neste artigo se baseia na mesma ideia do algoritmo das páginas *HITS* e *PageRank*. Foi analisado basicamente que quanto mais arestas o nó possuir na rede, maior será o seu potencial de propagação de conhecimento (ZHANG, 2007). Esse modelo é baseado na seguinte equação:

$$s(v_i)^{n+1} = s(v_i)^n + \sum_{v_j \in U} \sum_{e \in R_{ji}} w((v_j, v_i), e) s(v_j)^n$$

Onde, *w((vj, vi), e)* é o coeficiente de propagação; *Rji* é um relacionamento entre o nó *vj* e *vi*; *U* denota a coleção de vizinhança para *vi* e *Rji* é a coleção de todos os relacionamentos entre *vj* e *vi*. *S(vi)* representa a expertise do candidato a especialista *i* no assunto *q*.

O *PageRank* é o algoritmo de propagação e classificação de páginas web utilizado pelo motor de buscas do Google, e se baseia no fato de que as páginas são consideradas os nós na rede e os *links* que conectam estas páginas são as arestas. Para realizar a classificação das páginas é utilizado um passeio aleatório entre elas, em que continuamente são escolhidos os links em cada uma delas, de forma que as páginas que forem mais visitadas possuirão maior grau na classificação (PAGE, 1998).

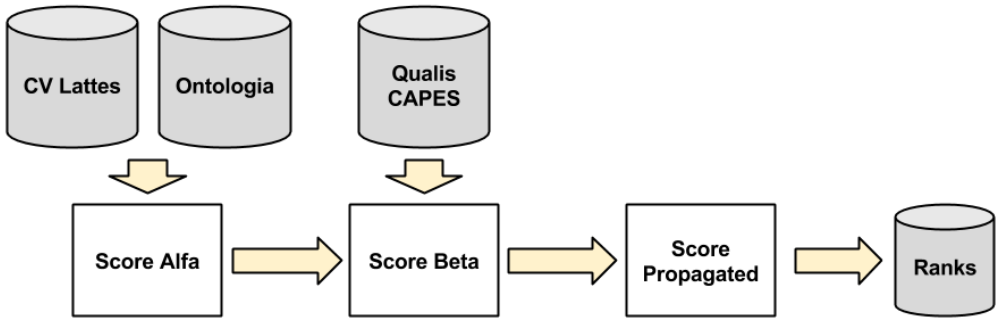
O algoritmo de *HITS* presume a divisão dos nós da rede social em servidores e autoridades de informação, em que os nós que apontam para muitos nós são considerados servidores e nós que recebem muitos nós são considerados autoridades da informação. Essa relação entre os dois tipos de nós, basicamente, se fortalece da seguinte maneira: bons servidores apontam para boas autoridades de conhecimento e boas autoridades de conhecimento são aquelas que são apontadas por diversos bons servidores (ZHANG, 2007).

Após cada uma das iterações do algoritmo de propagação é necessário normalizar os valores obtidos dividindo os mesmos pelo maior encontrado, dessa forma não serão criadas disparidades inexistentes, mantendo assim a qualidade dos pesos após a propagação (ZHANG, 2007). A normalização utilizada para atingir esse requisito foi a *min-max*, que consiste em padronizar os valores em um intervalo padrão de valores, que também se encarrega de mapear os valores atuais para essa faixa (HAN; KAMBER, 2001).

3 Metodologia

O algoritmo formulado baseia-se numa proposta híbrida que envolve uma análise de conteúdos das publicações e técnicas de redes sociais. O fluxo do algoritmo é demonstrado na Figura 1, em que as etapas de *Score Alfa* e *Beta* são as relacionadas ao conteúdo das publicações e a etapa de *Score Propagated* são relacionadas com as técnicas de redes sociais.

Figura 1 – Diagrama do funcionamento do algoritmo desenvolvido



Fonte: elaborada pelos autores.

Inicialmente, foram realizados pré-processamentos no banco de dados construído a partir de dados da plataforma Lattes por Digiampietri et al. (2012), filtrando apenas pesquisadores que indicam em seu currículo a área da Nanotecnologia como uma das suas áreas de atuação.

Em seguida, foi realizada a identificação de assuntos por meio do cálculo do TF-IDF, em que foram utilizadas apenas as palavras presentes nos títulos das publicações. Tendo-se em vista de que será utilizado o vocabulário controlado para definir quais termos serão usados para o cálculo do TF-IDF, os termos não presentes neste vocabulário serão automaticamente desconsiderados. Desse modo, será definido o que foi chamado de *Score Alfa* de cada assunto para os posteriores cálculos e aplicações de hipóteses. Entretanto, nessa fase só foram considerados os títulos, já que este era o único tipo de informação relevante disponível sobre os artigos no CV-Lattes. Trabalhos posteriores poderão considerar outros campos e fazer uso de outras fontes ou bases de dados.

Para a definição do *Score Beta* para cada assunto de cada publicação foram feitas as análises de veículos de publicação em comparação com a tabela de estratos Qualis fornecida pela CAPES e, de acordo com cada estrato, é adicionado um fator ao valor de *Score Alfa* já obtido para as publicações. Os estratos do Qualis são divididos por áreas, algumas revistas possuem notas diferentes em áreas diferentes. No caso do nosso exemplo de Nanotecnologia, por não existir uma área específica desta área no Qualis e por ser esta uma área bastante interdisciplinar, utilizamos as classificações da área INTERDISCIPLINAR, para contemplar artigos que relacionam várias áreas e envolvem Nanotecnologia.

Logo, é abordado o sentido de qualidade das publicações dando um maior diferencial a pesquisadores com suas pesquisas publicadas em periódicos renomados, ou seja, que possuem o conceito Qualis mais elevado.

Na última etapa do algoritmo, baseado no modelo de propagação desenvolvido por Zhang (2007), encontrou-se o *Score Propagated* de cada pesquisador por assunto, que consiste basicamente no resultado da etapa de *Score Beta* de cada pesquisador, propagado de acordo com seus relacionamentos.

Para isso, realizou-se um conjunto de iterações e normalizações previstas pelo algoritmo para cada assunto da área analisada, sendo que a condição de parada das iterações foi dada pela modificação dos *scores* já obtidos. Caso esteja abaixo de uma porcentagem indicada, as iterações são finalizadas.

Após a finalização de todas as iterações de um determinado assunto, os valores provenientes do *Score Beta* propagados na etapa de Score Propagated são ordenados de forma decrescente e armazenados na base de dados, constituindo o ranking para o assunto.

As fórmulas que resumem cada etapa do algoritmo podem ser representadas da seguinte maneira:

$$ScoreAlfa = TF\text{-}IDF(publicação, assunto)$$

$$ScoreBeta = ScoreAlfa(assunto) + Qualis(publicação)$$

$$ScorePropagated = ZhangPropagation(ScoreBeta(assunto), pesquisador)$$

A complexidade assintótica do algoritmo desenvolvido é calculada considerando seis variáveis diferentes: o número de palavras dos títulos (pal); a quantidade de assuntos (ass); a quantidade de publicações (pub); a quantidade de periódicos com Qualis na área em investigação (qual); o número de pesquisadores (pes) e a quantidade de relacionamentos de coautoria entre pesquisadores (coa). O algoritmo é limitado assintoticamente por pal*ass + qual + pub*pub + pes*coa. Sabendo-se que o número de veículos em periódicos com Qualis em uma determinada área costuma não ser muito alto e que o número de coautorias entre um dado número de pessoas é limitado pelo quadrado do número de pessoas, podemos dizer que o algoritmo é limitado assintoticamente por: pal*ass + pub*pub + pes*pes*pes cada termo multiplicativo desta fórmula correspondem, respectivamente, aos cálculos de TF-IDF das publicações por assunto, à identificação de coautorias e à propagação do *score* dentro da rede.

Os rankings a serem obtidos que possuem quantidade de pesquisadores superior a 1 elemento e com relacionamentos entre os membros serão validados comparando-se os resultados do ranqueamento com o resultado dos questionários, conforme será descrito a seguir.

Foi desenvolvida uma ferramenta de formulação de perguntas para os elementos da rede de forma que estes classifiquem quem (entre duas opções) conhece mais de um determinado assunto. A partir dos dados disponíveis foram identificados quais pesquisadores um determinado pesquisador conhece (pelas relações de coautoria) e estas informações foram utilizadas para a criação de perfis (um para cada pesquisador da rede). Para cada perfil foram formuladas perguntas aleatórias baseadas nas conexões do pesquisador, seguindo o padrão: "*No assunto Q, qual pesquisador possui mais conhecimento*?". Foram fornecidos como opções de escolha para cada pergunta dois nomes de pesquisadores com os quais o usuário já tenha se relacionado em publicações no assunto.

Os questionários foram divulgados para os pesquisadores via e-mails que continham a explicação do projeto, o endereço de acesso à ferramenta e as informações necessárias para a autenticação. A busca dos e-mails foi realizada

manualmente na internet, priorizando os pesquisadores que possuem mais conexões na rede social.

4 Resultados

Foram recuperados os currículos de 1.067 pesquisadores (de um total de 1.236.548 pesquisadores (DIGIAMPIETRI et al., 2012), que correspondem aos pesquisadores que possuem currículos Lattes e informaram atuar na área de Nanotecnologia. Estes pesquisadores registraram em seus currículos um total de 12.212 publicações dentre as 11.529.218 presentes na base total.

Com a execução do algoritmo, esta quantidade de pesquisadores e publicações foi diminuída, pois algumas publicações não se encaixaram nos termos da ontologia, ficando fora das classificações posteriores.

O uso de títulos para a classificação das publicações pode fazer com que artigos sejam eliminados de forma incorreta (falso negativos) pelo simples fato de possuir um título muito genérico, que não transpareça todos os elementos abordados no trabalho ou que não enfatize assuntos relacionados à área.

A quantidade final de pesquisadores e publicações pertencentes ao resultado da execução e que será utilizada nas demais análises foi de 697 e 7.548, respectivamente.

Com relação à quantidade de publicações por pesquisadores, pode-se identificar a diminuição do número dos mesmos conforme o aumento do intervalo de publicações, o que parece bastante aceitável. Em um tipo de abordagem mais generalizada, poderia se concluir que os pesquisadores com mais publicações seriam os mais bem pontuados, porém, na abordagem utilizada, dado a possibilidade de divisão destas publicações em diversos assuntos, isso não pode ser afirmado.

A ontologia utilizada como vocabulário controlado para a classificação do algoritmo contém 1.796 termos distintos que fazem parte da área de Nanotecnologia. Após a execução do algoritmo, contabilizaram-se 591 termos diferentes, excluindo-se, desta maneira, 1.205 termos que não foram referenciados em nenhum dos títulos das publicações apresentadas.

A média de menções dos termos da ontologia nos títulos das publicações é de 22,3 menções por termo. Entre os termos com mais menções (tanto em inglês quanto em português), pode-se destacar: "nanopartículas" (656 menções); "nanotubos" (388 menções); tubos (381 menções); vidro (371 menções); nanotubos de carbono (311 menções); ferro (251 menções); espectroscopia (241 menções); eletrônicos (241 menções); e dispersão (215 menções).

Isso indica que o conjunto de títulos das publicações utilizado abrange poucos dos assuntos da área, porém, a média de menções de cada um desses termos utilizados é relativamente alta. Assim, o conjunto se mostra bastante consistente e evidencia que provavelmente são estes os termos mais importantes da área no cenário nacional.

Das publicações classificadas na etapa *Score Alfa*, 2.941 produções possuem veículos de publicação com estratos Qualis atribuídos pela CAPES. Desta forma, 4.607 publicações não sofreram influência nesta etapa do algoritmo por não possuir veículo de publicação classificados pelo Qualis Periódicos.

Do total de publicações, foram identificadas 632 publicações em coautorias entre os membros da rede. Os títulos dessas publicações contêm 251 termos diferentes da ontologia. Com isso, pode-se perceber que a rede possui uma baixa quantidade de coautorias por publicação, verificando-se que a grande maioria das publicações é produzida por apenas um pesquisador ou em coautoria com pesquisadores que não consideram a área de Nanotecnologia como área de atuação.

Entre as coautorias identificadas, relacionando-as com os termos da ontologia, podem ser identificados alguns termos (tanto em inglês quanto em português) que estão envolvidos em mais relações de coautorias, como: "nanopartículas" (82 coautorias); "tubos" (57 coautorias); "nanotubos" (57 coautorias); "nanotubos de carbono" (49 coautorias); "nanocápsulas" (37 coautorias). É importante também salientar alguns termos que não possuem essas relações, embora estejam colocados na produção de mais de um pesquisador, como os termos "sínteses químicas", "genes" e "ácidos graxos", presentes nos títulos das publicações de sete pesquisadores não têm nenhuma

ligação de coautoria entre eles; assim como termos que só aparecem nos títulos das publicações de um único pesquisador da amostra, por exemplo, "cabos elétricos", "nanopontos", "agulhas", entre outros.

Observa-se que uma pequena parcela dos termos está envolvida em muitas relações de coautoria, enquanto a maior parte dos termos apresenta pouca ou nenhuma presença. Além disso, os termos que possuem maiores quantidades de publicações estão entre os que possuem pesquisadores melhor classificados.

O resultado final do algoritmo é um conjunto de rankings que ilustram como se comporta a rede por assunto após todos os tratamentos realizados. Esse conjunto é composto por 591 rankings diferentes (um para cada termo oriundo da ontologia e presente nos títulos das publicações).

Destacam-se como rankings composto por mais pesquisadores: "nanopartículas" (207 elementos); "espectroscopia" (104 membros); "nanotubos" (103 membros); "ferro" (102 membros); "tubos" (97 membros); "óleo" (91 membros); "nanocompósitos" (86 membros); e "eletrônicos" (84 membros). A média de pesquisadores por ranking é de 11,54 membros. Ao se retirar deste cálculo os rankings compostos por um único pesquisador (totalizando 153 rankings), a média sobe para 15,22 membros por ranking.

Como foi mencionado na seção 3, os rankings obtidos que possuem uma quantidade de pesquisadores superior a um elemento e com relacionamentos entre os membros foram validados comparando-se os resultados do ranqueamento com os dos questionários.

Ademais, dos 1.067 pesquisadores selecionados, foram enviados e-mails para 129 pesquisadores, que eram alguns dos mais reconhecidos na área e tínhamos acesso ao seu email. Destes pesquisadores, 63 responderam.

As respostas desses 63 pesquisadores foram utilizadas como base para analisar a acurácia do sistema de identificação de especialistas desenvolvido. A Tabela 1 contém a porcentagem de acertos e erros iniciais comparando-se o ranqueamento produzido pelo sistema desenvolvido com as respostas dos questionários.

Tabela 1 – Relação de acurácia do ranqueamento

	Acertos	**Erros**
Valores	41 (65,08%)	22 (34,92%)

Fonte: elaborada pelos autores.

Ao analisar os erros apresentados, verificou-se a existência de vários resultados que apontam erros no ranqueamento entre pesquisadores que são diferenciados por apenas uma posição. Isto é, o pesquisador A está situado na posição 5 do ranking do assunto X e o pesquisador B na posição 4, porém, o indivíduo A foi apontado como possuidor de mais conhecimento do que o especialista B na pesquisa. Esse tipo de resultado foi encarado como possivelmente uma zona de dúvida, já que, ao analisar alguns dos casos, reparou-se que as pontuações são muito semelhantes. Após a retirada desses resultados do conjunto, obtem-se a relação de acurácia de mais de 83%, conforme apresentado na Tabela 2.

Tabela 2 – Relação de acurácia dos rankings desconsiderando-se divergências de apenas uma posição no ranqueamento

	Acertos	**Erros**
Valores	41 (83,67%)	7 (14,29%)

Fonte: elaborada pelos autores.

Por apresentar, nestes testes, uma taxa de erro inferior a 15\%, o algoritmo pode ser considerado preciso.

Os resultados aqui produzidos também foram comparados com a ordenação dos pesquisadores produzida pela ferramenta de busca de pesquisadores do Portal da Inovação do Ministério da Ciência, Tecnologia e Inovação. Nesse portal, são gerados cálculos de relevância por meio de análise de frequências de termos no currículo (BRASIL, 2013); a ferramenta não esclarece a aplicação da abordagem de análise de qualidade dos artigos e

relacionamentos de coautoria. Além disso, não são aplicadas filtragens por área de atuação e pesquisa, gerando valores muito diferentes para termos mais abrangentes, por exemplo, no caso de Nanotecnologia, o termo "cabos" pode aparecer em pesquisas de diversas áreas, tornando impossível recuperar apenas os elementos relacionados com área alvo da classificação.

Devido a este fato, para que haja uma comparação mais coerente, o algoritmo desenvolvido com as ferramentas apresentadas neste artigo, utilizou-se apenas os ranqueamentos de termos bastante específicos da área. Foram selecionados quatro termos em particular: "nanoanéis", "nanocatalisadores", "nanomembranas" e "nanopós", que apresentam os resultados ilustrados na Tabela 3, mostrando uma alta convergência entre eles.

Tabela 3 – Comparação entre o ranqueamento resultante e os dados do portal da inovação

	Classificação Convergente	**Classificação Divergente**
Valores	15 (88,24%)	2 (11,76%)

Fonte: elaborada pelos autores.

5 Conclusão

Este trabalho apresentou um algoritmo de classificação de especialistas que utiliza como base os dados extraídos dos currículos Lattes, a classificação dos periódicos feita pela CAPES e uma ontologia de domínio.

Os resultados iniciais foram bastante satisfatórios e indicaram uma alta acurácia da solução proposta.

A pesquisa também demonstrou que a utilização de uma ontologia como forma de vocabulário controlado para a classificação apresenta vantagens e desvantagens. Ela facilita o processamento automático e filtragem de termos, fato especialmente útil quando a base de dados utilizada possui poucas informações sobre o conteúdo da produção, como no caso da base de currículos Lattes, em que só estão disponíveis os títulos das publicações.

Porém, esse tipo de filtragem pode acarretar em distorções pela não classificação de publicações da área que não possuem nenhum de seus termos na ontologia. Isso pode ser melhorado, entretanto, considerando a área de publicação do trabalho, mas este dado não é um campo muito usado pelos pesquisadores quando registram suas publicações no currículo Lattes.

O acréscimo de outras etapas de processamentos em trabalhos futuros, como a consideração do tempo da presença do pesquisador na rede, bem como outras variáveis adicionais, podem aumentar a acurácia do algoritmo. Também é válida a consideração da busca de outras fontes de dados que possam ser combinadas aos dados já recuperados, de forma a aumentar a informação disponível a respeito de cada publicação (por exemplo, o texto do resumo ou as palavras-chave), ou mesmo as informações sobre patentes.

Referências

ALUISIO, S. M. et al. **Desenvolvimento de uma estrutura conceitual (ontologia) para área de nanociência e nanotecnologia**. 2005. Relatório Científico. Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo, São Carlos, 2005.

BAEZA-YATES, R.; RIBEIRO NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999.

BALOG, K. et al. Formal models for expert finding in enterprise corpora. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 29., 2006, Seatle, USA, **Proceedings…** Seatle, 2006. p. 43-50.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. **Portal da inovação**. Disponível em: <http://www.portalinovacao.mct.gov.br/>. Acesso em: 10 jan. 2013.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB (WWW'98), 7., 1998, Brisbane, Australia, **Proceedings…** Brisbane, 1998. p. 107-117.

CAMPBELL, C.S. et al. Expertise identification using email communications. In: INTERNATIONAL CONFERENCE ON INFORMATION AND

KNOWLEDGE MANAGEMENT, 12., 2003, New Orleans, USA. **Proceedings…** New Orleans, 2003.

DIGIAMPIETRI, L. A. et al. Minerando e caracterizando dados de currículos lattes. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 2., 2012, Curitiba. **Proceedings…** Curitiba, 2012.

DIGIAMPIETRI, L. A. et al. BraX-Ray: an X-Ray of the Brazilian Computer Science Graduate Programs. **PLoS ONE**, v. 9, n. 4, p. e94541**,** 2014.

DOM, B. et al. Graph-based ranking algorithms for e-mail expertise analysis. In: ACM SIGMOD WORKSHOP ON RESEARCH ISSUES IN DATA MINING AND KNOWLEDGE DISCOVERY, 8., 2003, San Diego, USA. **Proceedings…** San Diego, 2003. p. 42-48.

FREITAS, C. M. D. S. et al. Extração de conhecimento e análise visual de redes sociais. In: SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE (SEMISH), 33., 2008, Belém. **Proceedings…** Belém, 2008. p. 108.

HAN, J.; KAMBER, M. **Data mining:** concepts and techniques. Burlington: Morgan Kaufmann, 2001.

JANEV, V. et al. Semantic web based integration of knowledge resources for expertise finding. **International Journal of Enterprise Information Systems**, Philadelphia, v. 5, n. 4, p. 53-71, 2009.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. **Journal of the ACM**, New York, v. 46, n. 5, p. 604-632, 1999.

KRULWICH, B.; BURKEY, C. Contactfinder agent: answering bulletin board questions with referrals. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 13., 1996, Portland, USA. **Proceedings…** Portland, 1996.

LAPPAS, T.; LIU, K.; TERZI, E. A survey of algorithms and systems for expert location in social networks. In: AGGARWAL, C. C. (Org.). **Social network data analytics**. Berlin: Springer Verlag, 2011. p. 215-241.

MARROCCO, C. et al. A linear combination of classifiers via rank margin maximization. In: JOINT IAPR INTERNATIONAL CONFERENCE ON STRUCTURAL, SYNTACTIC, AND STATISTICAL PATTERN RECOGNITION, 20., 2010, Cesme, Turkey. **Proceedings…** Cesme, 2010. p. 650-659.

MAYBURY, M. T. **Expert finding systems**. Nov. 2006. Technical paper, The MITRE Corporation. p. 1-35. Disponível em: <http://www.mitre.org/publications/technical-papers/expert-finding-systems>. Acesso em: 10 jan. 2013.

PAGE, L. et al. **The pagerank citation ranking:** bringing order to the web. 1998. Technical Report Computer Science. Stanford: University of Stanford, 1998.

SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval**. New York: McGraw-Hill, 1983.

STAAB, S.; STUDER, R. (Org.). **Handbook on ontologies**. 2. ed. Berlin: Springer Verlag, 2009.

STANKOVIC, M. Open innovation and semantic web: problem solver search on linked data. In: INTERNATIONAL SEMANTIC WEB CONFERENCE (ISWC), 7., 2010, Shanghai, China. **Proceedings…** Shanghai, 2010.

TAX, D.M.J. et al. Comparison between product and mean classifier combination rules. In: WORKSHOP ON STATISTICAL PATTERN RECOGNITION, 1., 1997, Prague, Czech Republic. **Proceedings…** Prague, 1997.

WANG, L. et al. A cascade ranking model for efficient ranked retrieval. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 34., 2011, Beijing, China, **Proceedings…** Beijing, 2011. p. 105-114.

WANG, G. A. et al. Expertrank: a topic-aware expert finding algorithm for online knowledge communities. **Decision Support Systems**, Amsterdam, v. 54, p. 1442-1445, 2013.

WU, S. Linear combination of component results in information retrieval. **Data & Knowledge Engineering**, Amsterdam, v. 71, n. 1, p. 114-126, 2011.

ZHANG, J. et al. Expert finding in a social network. In: INTERNATIONAL CONFERENCE ON DATABASE SYSTEMS FOR ADVANCED APPLICATIONS (DASFAA) 12., 2007, Bangkok, Thailand. **Proceedings…** Berlin: Springer, 2007, p. 1066-1069.

An algorithm for expert finding in areas based on Lattes' curriculum.

Abstract: This paper proposes to formulate an algorithm made for building a ranking of experts based on Lattes' curriculum data base. For this purpose, it was developed an algorithm based in three processing stages: Score Alfa, which analyzes the titles of documents based on ontology established to define the issues and priority of these subjects for each publication; Score Beta, that analyzes the publications' quality using concepts defined by CAPES to benefit productions that are published in most important vehicles; and Score Propagated, which examines the importance of being well connected to other specialist researchers, spreading the knowledge gained through this contact. In order to validate the accuracy of the algorithm, it was used the actual data of the Nanotechnology's area.

Keywords: Expert finding. Lattes' Curriculum. Social Network Analysis. Information retrieval.

Recebido em: 30/12/2014

Aceito em: 13/05/2015