Revista Eletrônica de Sistemas de Informação
ISSN 1677-3071

v. 14, n. 1

jan-abr 2015 - Edição Temática sobre Análise de Redes Sociais e Mineração

doi:10.21529/RESI.2015.1401

Sumário

Editorial

BrASNAM

ISSN: 1677-3071

Esta revista é (e sempre foi) eletrônica para ajudar a proteger o meio ambiente, mas, caso deseje imprimir esse artigo, saiba que ele foi editorado com uma fonte mais ecológica, a *Eco Sans*, que gasta menos tinta.

This journal is (and has always been) electronic in order to be more environmentally friendly. Now, it is desktop edited in a single column to be easier to read on the screen. However, if you wish to print this paper, be aware that it uses Eco Sans, a printing font that reduces the amount of required ink.

MINERANDO PUBLICAÇÕES CIENTÍFICAS PARA ANÁLISE DA COLABORAÇÃO EM COMUNIDADES DE PESQUISA – O CASO DA COMUNIDADE DE SISTEMAS DE INFORMAÇÃO

MINING SCIENTIFIC PUBLICATIONS TO ANALYSE THE COLLABORATION IN RESEARCH COMMUNITIES – THE CASE OF THE INFORMATION SYSTEMS COMMUNITY

(artigo submetido em março de 2013)

Renata Mendes de Araujo
Coordenadora do Programa de Pós-graduação em Informática - Universidade Federal do Estado do Rio de Janeiro (Unirio)
renata.araujo@uniriotec.br

Brunno Athayde Silveira
Mestrando do Programa de Pós-graduação em Informática - Universidade Federal do Estado do Rio de Janeiro (Unirio)
brunno.silveira@uniriotec.br

Thiago Yusuke Muramatsu
Graduado em Sistemas de Informação pela Universidade Federal do Estado do Rio de Janeiro (Unirio)
thiago.muramatsu@uniriotec.br

Kate Revoredo
Professora do Programa de Pós-graduação em Informática - Universidade Federal do Estado do Rio de Janeiro (Unirio)
katerevoredo@gmail.com

ABSTRACT

Scientific communities are social structures composed by people and/or institutions connected through relationships, who have common interests and objectives, sharing information and knowledge. Following this view, social networks have been used to analyze and understand these communities, particularly through their scientific publications. From another perspective, recently born scientific communities are still consolidating their themes of interest and have little understanding about the existence as well as the potential of collaboration networks within the community. In these communities, the information analysis from publications face the challenge of low structure and the absence of consolidated parameters for grouping themes of interest. This work presents an approach to analyze scientific communities through text mining of their publications. The collaborative networks of authorship are automatically extracted and the publications context (classification) automatically found. A computational tool has also been developed to support these analyses. A case study with the Brazilian research community in Information Systems was conducted based on the past editions of the Brazilian Symposium on Information Systems.

Key-words: social network analysis; text mining; document classification; Brazilian IS research community.

RESUMO

Comunidades científicas são estruturas sociais compostas por pessoas e/ou instituições que se conectam através de relações e compartilham interesses comuns – informação, conhecimento e esforços em busca do mesmo objetivo. Esta característica tem levado ao uso de abordagens de análise de redes sociais para a compreensão destas comunidades, tendo como base suas publicações científicas. Contudo, comunidades de formação recente ainda estão consolidando seus assuntos de interesse e possuem pouco entendimento tanto da existência como do potencial de relações de colaboração entre seus participantes. Nestas comunidades, a análise de informação a partir de publicações enfrenta o desafio da pouca estruturação de sua produção e poucos parâmetros para agrupamento de temas de interesse. Este trabalho apresenta uma abordagem para a obtenção de indicadores para análise de comunidades científicas usando mineração de textos, que envolvem a rede de colaboração entre autores e o contexto de publicação (classificação de artigos). Uma ferramenta computacional para o apoio à análise foi também desenvolvida. A proposta foi avaliada considerando a comunidade nacional de Sistemas de Informação, com base nas publicações nas edições do Simpósio Brasileiro de Sistemas de Informação.

Palavras-chave: análise de redes sociais; mineração de textos; classificação de documentos; comunidade brasileira de Sistemas de Informação.

1 INTRODUÇÃO

A análise de redes sociais tem possibilitado diversas oportunidades para a compreensão da interação e da organização social de um grupo (BARABÁSI, 2003). Uma rede pode ser caracterizada segundo suas propriedades estruturais e topológicas sendo, em sua grande maioria, derivadas da teoria dos grafos, que explicam a sua estrutura. Um dos principais usos da análise de redes sociais é analisar tais propriedades, como, por exemplo: a centralidade de determinados nós da rede, a densidade de suas relações, sua capacidade de interconexão ou comunicação etc. (WASSERMAN E FAUST, 1994). Além disso, com a análise de redes sociais, busca-se entender os relacionamentos e o fluxo de informações entre pessoas, grupos e organizações. A unidade na análise de redes sociais não é o indivíduo, mas sim a coleção de indivíduos e os relacionamentos entre eles.

O movimento de surgimento e consolidação de comunidades de pesquisa é baseado principalmente na agregação de pesquisadores e resultados ao redor de uma mesma área/assuntos de interesse ao longo de um período de tempo e ações estratégicas de indução, consolidação e manutenção das associações entre estes grupos. Uma vez que estas relações de interesse e colaboração deixam de existir, a comunidade, em consequência, também declina. Comunidades científicas são estruturas sociais compostas por pessoas e/ou instituições que se conectam por meio de relações e compartilham interesses comuns – informação, conhecimento e esforços em busca do mesmo objetivo. Propostas de análise de redes sociais de pesquisa podem ser observadas (SCRIPTLATTES, s.d.; OLIVEIRA, LOPES E MORO, 2011; REIJERS *et al.*, 2009; FREIRE E FIGUEIREDO, 2011; MAIA *et al.*, 2012). A base destas propostas está nas possibilidades de mineração, visualização e análise de estruturas de redes sociais de pesquisadores, instituições, grupos e temáticas de pesquisa em uma determinada área, a partir de suas bases de produção, principalmente os artigos científicos produzidos por seus pesquisadores.

Em particular, comunidades científicas de formação recente possuem poucos parâmetros para classificação de assuntos de interesse e pouco entendimento tanto da existência como do potencial de relações de colaboração. Nestas comunidades, a compreensão de sua composição e tendências de interesse se beneficia de técnicas de descoberta de conhecimento a partir de seus artefatos principais de produção – publicações. O desafio de acompanhamento destas comunidades, no entanto, está na pouca estruturação de sua produção e nos poucos parâmetros para agrupamento de temas de interesse.

Este trabalho tem por objetivo apresentar uma abordagem para analisar uma comunidade científica com base em suas publicações científicas analisadas por meio de técnicas de mineração de textos (FELDMAN E SANGER, 2007) para a identificação do contexto das publicações e a geração de sua rede de colaboração. A proposta foi avaliada considerando a comunidade nacional de Sistemas de Informação, através das publicações

nas edições do Simpósio Brasileiro de Sistemas de Informação (SBSI, 2012; CESI, 2012). Uma primeira análise foi apresentada em Revoredo *et al.* (2012), na qual as edições do SBSI de 2008 a 2011 foram analisadas. Neste artigo, os resultados são complementados com a inclusão das análises dos dados de 2012. Além disso, o ferramental de apoio desenvolvido para a análise é apresentado, e são elaboradas discussões a respeito da evolução desta comunidade de pesquisa.

O artigo está estruturado da seguinte forma: na Seção 2 são apresentados o objetivo e processos da mineração de textos e os tipos de informação (classificação e agrupamento) que podem ser descobertos com sua aplicação. A Seção 3 apresenta o método e ferramenta propostos para a análise de comunidades científicas, a partir de suas publicações. A seção 4 apresenta os resultados obtidos com a mineração e análise da comunidade de pesquisa em Sistemas de Informação no Brasil. A seção 5 apresenta os trabalhos relacionados e, finalmente, a seção 6 encerra o artigo apresentando conclusões e perspectivas futuras.

2 MINERAÇÃO DE TEXTO

O processo de Descoberta de Conhecimento em Bases de Dados (DCBD) (WITTEN *et al.*, 2011) analisa e interpreta, de forma automática, dados para descobrir padrões compreensíveis, válidos, novos e potencialmente úteis. DCBD é um processo iterativo e interativo como ilustrado na Figura 1.

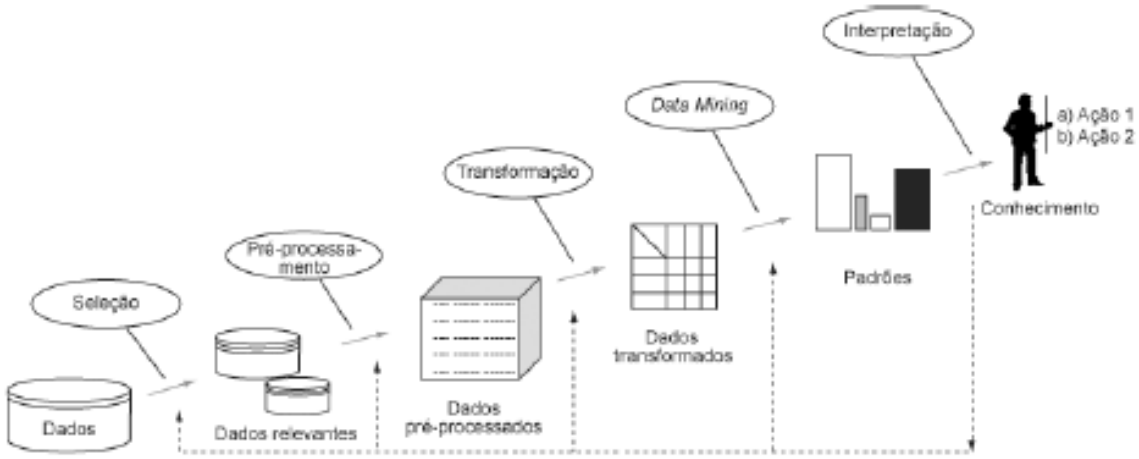

Figura 1: Etapas do processo de DCBD

Fonte: Fayyad *et al.* (1996).

As primeiras etapas (seleção, pré-processamento e transformação) comumente denominadas apenas de fase de *pré-processamento,* correspondem a uma análise exploratória dos dados, com aplicação de funções relacionadas à captação, organização e ao tratamento dos dados. Essa fase tem, entre outros, o objetivo de encontrar as características mais relevantes dos dados, reduzindo, quando possível, a dimensionalidade da base de dados. Assim, o conjunto de dados relevantes para a descoberta

de conhecimento é selecionado e então utilizado pelos algoritmos da etapa seguinte, a *mineração de dados*. Esta é considerada a etapa mais importante do processo de DCBD, pois é nela que algoritmos de aprendizado de máquina (MITCHELL, 1997) são aplicados para aprender modelos que descrevem os padrões existentes implicitamente nos dados. A última etapa é a de *pós-processamento,* em que o modelo aprendido é interpretado, e a melhor forma para apresentá-lo é selecionada. Ou seja, essa etapa cuida do tratamento do conhecimento obtido na mineração de dados, viabilizando a avaliação dos padrões e a utilidade do conhecimento.

Quando o conjunto de dados corresponde a dados não estruturados, o processo é denominado Descoberta de Conhecimento em Textos (DCT) (FELDMAN E SANGER, 2007). Neste caso, a etapa de pré-processamento inclui técnicas de Processamento de Linguagem Natural (PLN) (MANNING E SCHUETZE, 1999), onde a compreensão automática da linguagem e representações manipuláveis por programas de computador são consideradas. Esta etapa envolve os campos da linguística, inteligência artificial, ciência da computação e lexicografia (BATES, 1995). Além disso, a etapa de mineração de dados passa a ser denominada *mineração de texto*.

Os modelos gerados na etapa de mineração de texto são comumente utilizados para classificação. Nesta, o modelo aprendido adota uma estrutura que permite a classificação automática de um novo documento em uma ou mais categorias previamente definidas. Por exemplo, um classificador pode determinar o gênero (carta, artigo científico, relatório técnico) de um documento considerando as suas características, como comprimento e estrutura. A utilização de técnicas de DCBD para geração de um modelo que permita a classificação de um novo documento se justifica quando o repositório de documentos é muito grande, o que inviabiliza a classificação manual.

Para que o classificador seja aprendido, a fase de pré-processamento (preparação) dos documentos precisa ser executada. Nesta, os documentos são agrupados entre as categorias existentes para que então sejam analisados. Esta análise permite que características dos documentos que ajudem a distinguir uma categoria da outra sejam encontradas. Por exemplo, a existência de um destinatário pode auxiliar na distinção de uma carta de um artigo científico ou um relatório técnico.

Algumas técnicas de PLN facilitam o processo de seleção dessas características, tais como: retirada dos termos que não influenciam na definição da categoria do documento, retirada de símbolos não relevantes (ex: #,#,$,%,¨,&,*,(,), etc.), conversão de termos em radicais, entre outras. Por exemplo, na frase “Uma publicação científica deve conter uma contribuição significativa”, contida em um documento do gênero artigo científico, o termo “uma” pode ser desconsiderado. Além disso, pode-se entender que verbos não são relevantes para a determinação do gênero do documento em questão e também podem ser retirados.

A importância de cada um dos termos remanescentes é calculada utilizando alguma medida de relevância. Nesse trabalho foi utilizado o

escore de relevância (SALTON E BUCKLEY, 1988), que avalia a relevância de um termo para uma determinada categoria, comparado à relevância para as demais categorias. A técnica do escore de relevância foi apresentada e aplicada no estudo de Wiener *et al.* (1995). A seleção dos termos deve ter como objetivo encontrar um subconjunto de termos que se mostre o mais eficiente para a tarefa de classificação. Os termos, então, devem ser adequadamente discriminativos entre as categorias. Cada categoria é analisada individualmente, com base em seus documentos, para a seleção dos subconjuntos de termos que melhor discriminam os documentos desta categoria.

Para a seleção dos termos representantes de uma categoria, todos os termos recebem um valor (escore de relevância) de quão bem servem como preditores individuais da categoria. Valores altamente positivos e altamente negativos indicam termos úteis para a discriminação. No primeiro caso, o termo contribui fortemente para a indicação de que o documento pertence à categoria sendo analisada. Nesta situação, o termo é comum aos documentos de uma mesma categoria e incomum aos documentos das outras categorias. Já no segundo caso, o termo fornece indicação de que o documento não pertence à categoria analisada. Em seguida, os *k* termos mais relevantes são selecionados, onde *k* é um parâmetro definido por quem realiza a análise.

Durante esse aprendizado, a base de documentos considerada é dividida em duas. A primeira, denominada de treinamento é utilizada para aprender o classificador e a segunda, denominada de teste, é utilizada para avaliar o classificador aprendido. Caso a avaliação do classificador não seja satisfatória, o treinamento é feito novamente, considerando, por exemplo, novas técnicas de pré-processamento dos documentos. Quando o classificador for bem avaliado, ele passa a ser utilizado para categorizar novos documentos recebidos, definindo então a fase de classificação propriamente dita.

A classificação de um novo documento é feita a partir da lista dos termos discriminativos de cada categoria. Uma forma de efetuar essa classificação é verificar a quantidade de termos de cada categoria que o documento sendo classificado contém. Para isso, o documento sendo classificado também passa por uma etapa de pré-processamento, em que os termos mais relevantes do documento são extraídos. A categoria que tiver mais termos sendo mencionados no documento determina a categoria desse documento. Se for do interesse, também é possível associar mais de uma categoria a um mesmo documento: as *m* categorias que tiveram termos das suas listas de termos mencionados no documento são consideradas.

Um grau de pertinência em cada uma das categorias também pode ser indicado (abordagens de similaridade difusa), baseando-se por exemplo na quantidade de termos em cada categoria ou no escore de relevância. Essa ultima abordagem será a utilizada nesse artigo. Outras formas, de classifi-

cação podem ser consideradas, como por exemplo, a abordagem Rocchio ou Redes Bayesianas (BING, 2011).

3 ANÁLISE DE COMUNIDADE CIENTÍFICA POR MINERAÇÃO DE TEXTO

Para o auxílio à análise de comunidades científicas, foi desenvolvida uma abordagem que aplica técnicas de mineração de textos em conjuntos de publicações científicas e um ferramental de apoio para a análise. Entende-se que, particularmente para comunidades recentes, dois tipos de informação são relevantes para a análise de sua configuração e evolução no tempo: as colaborações e parcerias entre os pesquisadores da área e os temas de interesse compartilhados e para os quais a área aparenta convergir.

A hipótese estudada nesta pesquisa é a de que é possível extrair esta informação a partir da base de publicações científicas que delineia uma comunidade, por meio da mineração de textos. Na abordagem proposta, a mineração de artigos científicos extrai informação relacionada à colaboração de autoria entre pesquisadores que publicam na área e a classificação automática dos artigos em categorias de interesse da comunidade.

Tendo em vista os contextos de análise iniciais utilizados para treinamento e avaliação desta aplicação, sua utilização está inicialmente voltada às comunidades de pesquisa existentes no âmbito da Sociedade Brasileira de Computação (SBC). Nesta sociedade, as comunidades de pesquisa se estruturam ao redor de eventos científicos, com uniformidade de características e, sobretudo, formatação de suas publicações. A estratégia de formação, consolidação e acompanhamento das comunidades está bastante vinculada à dinâmica dos eventos a elas ligados. Identificar maneiras de descobrir padrões de colaboração e evolução de temáticas em cada área, a partir das publicações em eventos, pode se tornar uma ferramenta estratégica importante para estas comunidades.

3.1 GERAÇÃO DE REDES DE COLABORAÇÃO

Aqui o objetivo é minerar o repositório de publicações para gerar dois tipos de redes de colaboração – colaboração entre autores e colaboração entre as instituições às quais os autores estão vinculados. No primeiro caso, os nós da rede representam autores das publicações e existirá uma aresta conectando dois nós se eles tiverem publicado um artigo científico em co-autoria. O peso da aresta indica a quantidade de artigos que foram publicados em conjunto. Já no segundo caso, os nós representam instituições, e existirá uma aresta conectando dois nós se existir um artigo escrito por autores dessas duas instituições. O peso da aresta vai indicar a quantidade de artigos publicados em que a afiliação dos autores envolvidos corresponde às instituições em questão. A Figura 2 ilustra uma rede de colaboração entre pesquisadores e outra entre instituições.

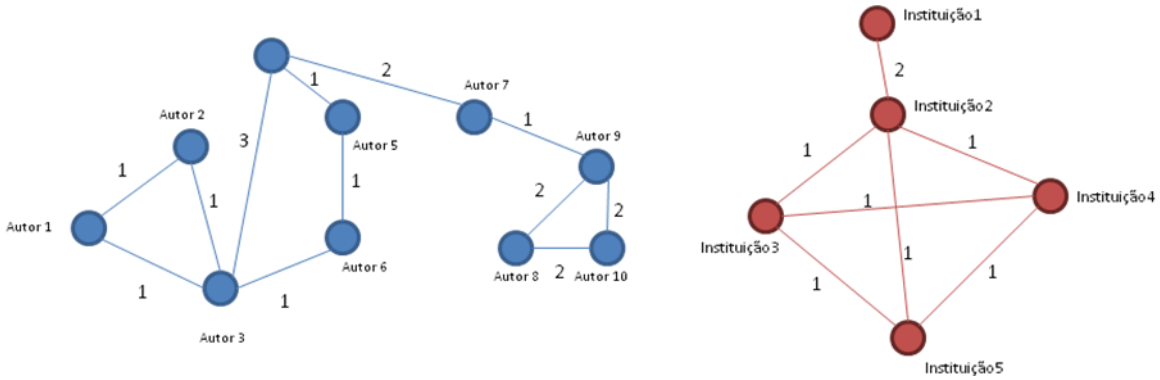

Figura 2: Exemplo de redes de colaboração

Fonte: elaborado pelos autores

Para gerar as redes de colaboração, um conjunto de artigos científicos é fornecido como entrada, estruturados em: título, autores, afiliação, resumo em português, resumo em inglês e texto do artigo. A primeira tarefa a ser executada é a de extração das informações, em que técnicas de reconhecimento de entidades (MANNING *et al.*, 2008) são utilizadas para extrair os nomes dos autores e das instituições envolvidas.

Como o objetivo é analisar a colaboração científica, algumas regras foram aplicadas: i) artigos com um único autor foram descartados; ii) dois co-autores de uma mesma instituição não caracterizam uma colaboração entre instituições; iii) autores com mais de uma afiliação, sendo uma delas uma empresa, tem a empresa desconsiderada como instituição.

3.2 CLASSIFICAÇÃO DE PUBLICAÇÕES

Como mencionado na Seção 2, antes de poder ser utilizado para classificação, um classificador precisa ser aprendido por meio de algum algoritmo de aprendizado de máquina. Na abordagem proposta, um conjunto de artigos científicos da comunidade sendo avaliada é utilizado para aprender um classificador (fase de treinamento do classificador). Em etapa seguinte, um determinado artigo científico da mesma comunidade (não constante no conjunto utilizado para a aprendizagem do classificador) é classificado.

Para a fase de treinamento, um algoritmo de aprendizado supervisionado foi considerado, ou seja, para o conjunto de artigos utilizados no treinamento do classificador, foi associado um conjunto preliminar de categorias (tópicos ou assuntos de interesse da comunidade). Os artigos foram associados a cada categoria e o processo de aprendizado foi executado para que os termos relevantes que definem cada uma das categorias pudessem ser definidos, utilizando-se o escore de relevância. A Figura 3 representa o esquema de criação das listas de termos do conjunto preliminar de categorias, em que as atividades ("Preparação do artigo" e "Seleção das características") foram realizadas de maneira automática.

Figura 3: Processo de identificação das categorias

Fonte: elaborada pelos autores

A partir da lista de termos associada a cada categoria, é possível classificar novos artigos científicos. Primeiro os termos relevantes desse artigo são extraídos, preparando o novo artigo. Em seguida, a similaridade do artigo sendo classificado com cada uma das categorias é estabelecida com base na lista de termos identificados. Cada artigo pode ser definido como similar a mais de uma categoria, ou seja, ele pode estar associado a mais de uma categoria. A Figura 4 mostra o esquema de classificação de um novo artigo.

Figura 4: Processo de classificação de um novo artigo

Fonte: elaborada pelos autores

A abordagem considera ainda que a fonte principal para a identificação de termos seja o conteúdo dos resumos dos artigos. O resumo de um artigo científico tem como característica ser conciso, expressando em poucas palavras o objetivo do trabalho, ou seja, é uma boa sumarização do artigo. Dessa forma, é natural considerar que o resumo contém as palavras relevantes para determinação dos tópicos de interesse do artigo em questão, ou seja, para a sua classificação. Com isso, a lista de termos relevantes é gerada a partir dos resumos extraídos de cada um dos artigos, desconsiderando-se seu corpo principal.

3.3 FERRAMENTA E3SUITE

Um ferramental específico para o apoio à análise com o uso da abordagem proposta foi desenvolvido. Esta ferramenta, denominada Suite para Estudo da Evolução de Eventos - E3Suite (rspsi.uniriotec.br) oferece duas funcionalidades principais: a geração de redes de colaboração e a classificação automática de artigos científicos. A ferramenta requer que os documentos utilizados como entrada sejam artigos que sigam o modelo de artigos da Sociedade Brasileira de Computação (SBC, s.d.), no formato *Adobe Portable Document Format* (PDF).

A geração de redes de colaboração se inicia com a entrada do conjunto de artigos que se deseja analisar, a serem processados via mineração de textos para a extração de informações sobre as relações de coautoria nos artigos. A visualização da rede de colaboração é possível por meio da utilização da ferramenta Pajek[1].Esta ferramenta aceita como entrada um arquivo do tipo texto formatado, onde são especificados os nós, as arestas e o peso das arestas, para a geração da rede. Dois arquivos no formato esperado pelo Pajek são gerados a partir da mineração das publicações - um que considera os autores e outro que considera as instituições. Alterações simples na formatação permitem a manipulação dos nós e arestas, como alteração de cor, tamanho e posição.

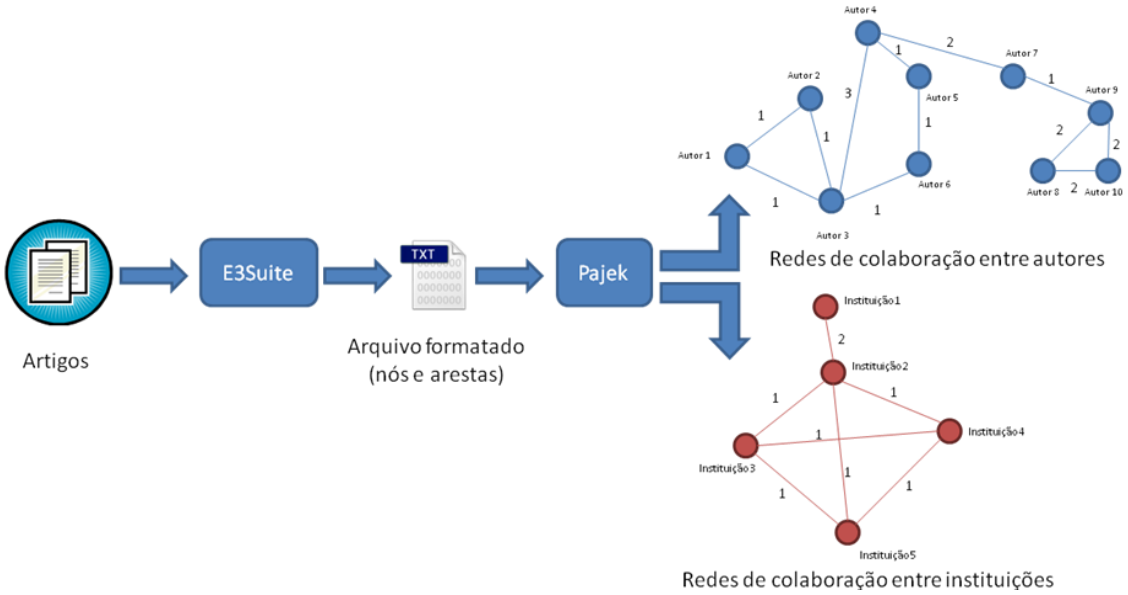

Figura 5: Esquema de geração de redes de colaboração a partir da mineração de texto na ferramenta E3Suite

Fonte: elaborada pelos autores

Para a classificação de artigos, a aplicação permite, a partir de uma determinada base de categorias conhecidas, classificar os artigos científicos de acordo com os tópicos de interesse da comunidade. Para identificação das categorias iniciais, o usuário deve entrar com os artigos (treinamento) e suas respectivas categorias. O algoritmo então encontra palavras importantes (ou seja, termos que distinguem uma categoria em relação à outra), formando dicionários por categorias. Sendo assim, é possível a partir da análise de um novo artigo, ainda não categorizado, classificá-lo.

A partir da tela do módulo de classificação (Figura 6), o usuário deve indicar a pasta com os artigos a serem utilizados para formação das categorias (base de treinamento) e a pasta dos artigos que serão classificados

[1] http://pajek.imfm.si/doku.php?id=start

(base de teste). Os artigos devem estar em formato PDF, diretamente nas pastas indicadas. Além disso, o usuário deve entrar com uma planilha em formato EXCEL, indicando os tópicos dos artigos.

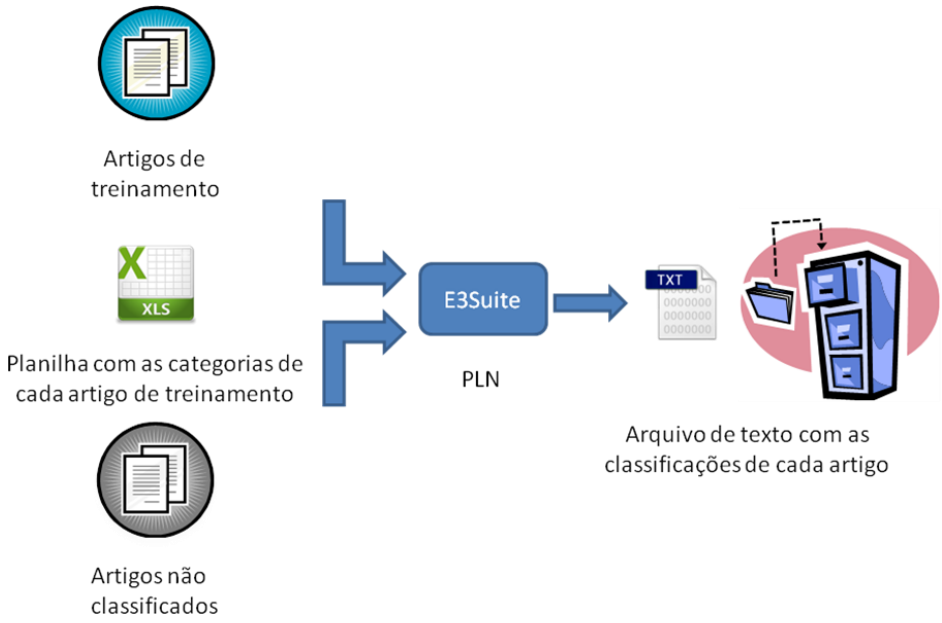

Figura 6: Esquema módulo de classificação da E3Suite

Fonte: elaborada pelos autores

4 ANÁLISE DA COMUNIDADE DE PESQUISA EM SISTEMAS DE INFORMAÇÃO

A área de pesquisa em Sistemas de Informação tem sido vista, no contexto da Sociedade Brasileira de Computação, como a interseção de áreas consolidadas da Ciência da Computação, notadamente as áreas de Engenharia de Software e Banco de Dados, com o viés principal de entender a construção destes sistemas. No entanto, ao se confrontar com os problemas do mundo, uma visão integrada não só destas duas áreas, mas de várias outras áreas da Computação (Redes de Computadores, Inteligência Artificial, Algoritmos e Otimização, para citar algumas) tem se tornado inevitável, a fim de compreender o objeto Sistema de Informação aplicado em um contexto específico de demanda e utilização.

Em seu caráter técnico, além da contribuição da própria Computação, a área de SI em muitas situações precisa se apropriar das soluções e referenciais teórico-práticos de áreas como a Ciência da Informação, Administração, Economia, entre outras. Por envolver ainda aspectos não totalmente técnicos, também tangencia áreas relacionadas à Psicologia, Sociologia e demais áreas das Ciências Sociais. O desafio de caracterização da área de pesquisa em Sistemas de Informação está principalmente em como tratar a complexidade inerente à pesquisa nesta área, por sua característica multidisciplinar e exigência de aplicação prática.

Em âmbito nacional, o Simpósio Brasileiro de Sistemas de Informação (SBSI) tem sido o evento principal para a apresentação de trabalhos científicos e a discussão de temas relevantes nesta área, aproximando pesquisadores, estudantes e empresários da comunidade de Sistemas de Informação, no âmbito da Sociedade Brasileira de Computação[2]. O evento ocorre desde o ano de 2004, mas somente a partir de 2008 dados sobre sua produção científica puderam ser organizados sistematicamente. Os resultados apresentados nesta seção compreendem primeiramente os artigos científicos aceitos para publicação nas edições de 2008 a 2012, totalizando 114 artigos.

4.1 REDE DE COLABORAÇÃO

A análise das redes de colaboração pode ser realizada para um ano específico de edição do evento. A rede apresentada na Figura 4 mostra a colaboração entre os autores do SBSI no ano de 2011, em que 23 artigos foram considerados. O peso da aresta representa o número de vezes que os autores colaboraram entre si nesta edição do evento. Os nomes dos autores foram omitidos, por não serem relevantes para a análise pretendida. Podemos perceber, pela análise da rede, que, nos 23 artigos analisados, nenhum pesquisador foi autor em mais de um artigo. Os *clusters* apresentados são redes totalmente distribuídas, nenhum dos nós é centralizador, separando outros *clusters*, ou seja, nenhum dos nós é ligado a outros dois nós que não se ligam entre si. Isso significa que nenhum dos autores tem uma posição centralizadora, de articulação de autoria. A comparação das redes ao longo dos anos de realização do evento permite observar a evolução da colaboração entre os membros da comunidade.

A análise das redes de colaboração também pode abranger um conjunto de publicações de vários anos. Por exemplo, a rede de colaboração de instituições para as edições 2008, 2009 e 2010 do SBSI pode ser visualizada na Figura 5. Nesta rede é possível perceber a existência de agrupamentos de instituições. A análise destes agrupamentos demonstra que a comunidade de pesquisa apresenta uma tendência à produção entre instituições em uma mesma região, pouca participação internacional e parcerias esporádicas (pouca quantidade de artigos entre os mesmos autores).

[2] Os pesquisadores da área de Administração também têm eventos consolidados que discutem o tema, mas a análise ora realizada se restringiu ao SBSI, que é o principal evento de Sistemas de Informação organizado pela comunidade da Computação e referendado pela Sociedade Brasileira de Computação.

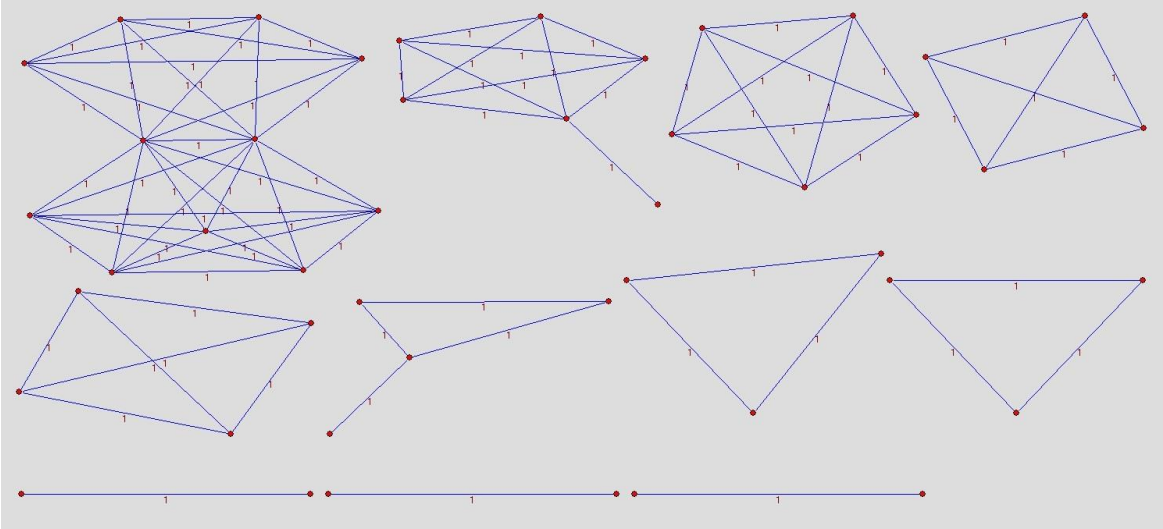

Figura 7: Rede de colaboração entre autores do SBSI (2011)

Fonte: elaborada pelos autores a partir dos dados de campo

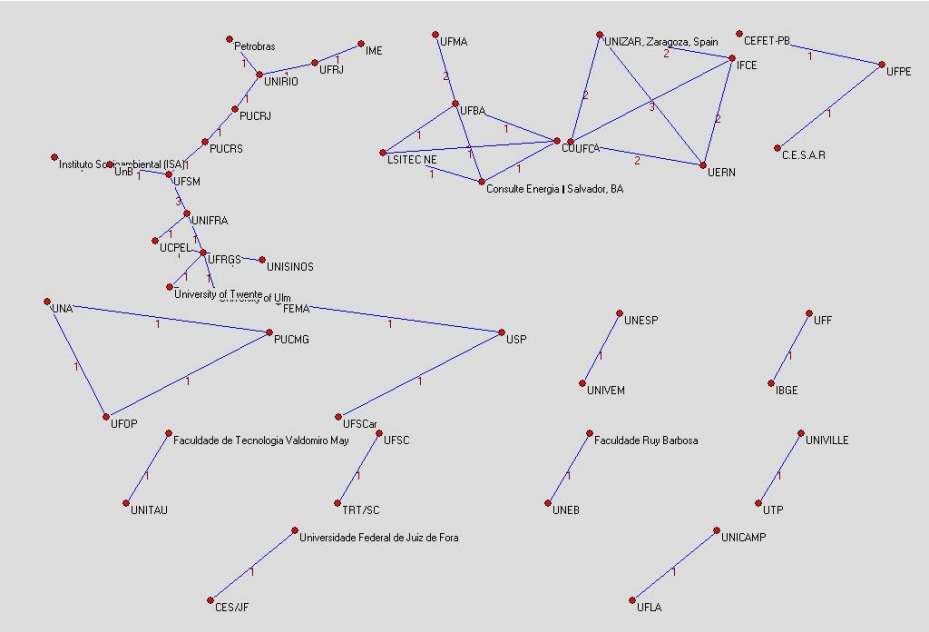

Figura 8: Rede de colaboração de instituições SBSI (2008 a 2010)

Fonte: elaborada pelos autores a partir dos dados de campo

4.2 CLASSIFICAÇÃO DE PUBLICAÇÕES

Foi utilizada uma base de treinamento considerando quatro edições dos anais do SBSI (2008-2011) e para validação do classificador aprendido os anais da edição de 2012. Dessa forma, foram considerados 80 artigos para treinamento, e 32 para validação. O número de artigos utilizados por edição encontra-se na Tabela 1.

Tabela 1. Artigos considerados na classificação de 2012

Ano	Número artigos considerados	Base
2008	21	Treinamento
2009	20	Treinamento
2010	16	Treinamento
2011	23	Treinamento
2012	32	Teste

Fonte: elaborada pelos autores a partir dos dados de campo

A associação de artigos em tópicos foi feita por meio da classificação realizada pelos próprios autores dos artigos quando da submissão ao simpósio. Como foram considerados os artigos referentes às edições de 2008 até 2012 do SBSI, um tratamento dos tópicos foi necessário, para alinhar tópicos sintaticamente diferentes, mas semanticamente semelhantes entre as edições do evento. Ao todo, foram levadas em consideração 32 categorias/tópicos.

- Classificações semelhantes

Para um total de 32 artigos analisados, a aplicação foi capaz de sugerir a classificação de 25 artigos (78% dos artigos), ou seja, encontrou semelhança entre o artigo e pelo menos um dos tópicos de interesse cadastrados. Uma categoria que teve sua lista de termos formada por um grande número de artigos permitiu que o classificador classificasse corretamente artigos associados a essa categoria. Um exemplo é a categoria "Metodologias e abordagens para engenharia de Sistemas de Informação", a categoria mais relacionada pelos autores em 2011.

- Classificações divergentes

Classificações divergentes compreendem os artigos que a ferramenta classifica em categorias diferentes das estipuladas pelos autores no momento de submissão do artigo. Em geral, todos os artigos de 2012 aos quais o autor associou mais de um tópico, tiveram como sugestão da ferramenta uma categoria que representasse um tópico estipulado pelo autor.

Artigos que foram classificados pelos autores em apenas um tópico tiveram categorias sugeridas de forma divergente, por exemplo, para os tópicos: *Administração e negócios*; *Metodologias e abordagens para Engenharia de Sistemas e Informação*; *Infraestrutura de Tecnologia da Informação* e *Psicologia Aplicada a Sistemas de Informação, Aspectos Sociais e Humanos em SI.* Esta análise pode indicar que estes temas são demasiado genéricos para indicar objetivamente o conteúdo do artigo, dificultando a classificação pelos autores ou podem indicar temas novos a serem incentivados para a pesquisa da comunidade; podem, ainda, denotar uma falsa indicação de interesse e convergência nestes temas pela comunidade.

A dificuldade de classificação dos artigos em algumas categorias pode também ser um reflexo do treinamento da base. Tópicos como *Infraestrutura de Tecnologia da Informação*, por exemplo, foram treinados a partir de somente um artigo na base de treinamento. Isto diminui a margem de entendimento da ferramenta sobre como classificar artigos nesta mesma categoria.

Outro ponto a se levar em consideração nessa análise é que nem sempre a classificação do autor pode ser a mais adequada dentre a lista de tópicos do evento. Neste caso, o classificador permite a classificação em categorias mais adequadas. Por exemplo, um dos artigos não classificados da categoria "Administração e negócios", foi classificado como pertinente ao tópico *E-Business, E-Commerce e E-Government*.

Outra peculiaridade percebida: todos os nove artigos de 2012 que os autores classificaram como *"Planejamento, Auditoria, Alinhamento Estratégico, Segurança e Risco, Qualidade, Gerência de projetos e Gestão de Processos de Negócio de Sistemas de Informação*" tiveram entre as sugestões do classificador a categoria "*Gestão da informação*". Nesta situação, o classificador pode estar indicando a revisão de tópicos que envolvam muitos assuntos e sua substituição por um tópico mais genérico e de mais fácil identificação pela área.

4.3 REDES DE COLABORAÇÃO E CLASSIFICAÇÕES COMO FERRAMENTAS DE GESTÃO DO SBSI

A partir das análises apresentadas nas seções anteriores, podemos delinear como os dados oferecidos pela abordagem podem ser utilizados para promover a gestão de atividades do SBSI e, consequentemente, a evolução da comunidade de Sistemas de Informação no contexto da Sociedade Brasileira de Computação.

As análises das redes de colaboração demonstram que a comunidade segue pouco integrada em termos de produção científica, novos autores e instituições surgem a cada ano e lideranças são ainda tênues. Isto pode indicar aos organizadores do evento e coordenadores da área, a promoção de ações de integração de discussão e pesquisa, promovendo o trabalho em conjunto. Análises mais detalhadas da colaboração entre autores do evento podem ser encontradas em Oliveira e Dias (2012).

No que diz respeito às classificações, o refinamento contínuo dos tópicos obtidos em escores de relevância pode auxiliar a comunidade a aperfeiçoar a identificação de tendências temáticas em seu âmbito. Por outro lado, a análise das classificações automáticas pode levar ao delineamento de ações direcionadas pelos gestores da comunidade para o fortalecimento de temas reconhecidos como relevantes pela própria comunidade, a eliminação de temas ultrapassados ou a indução de temas estratégicos.

Os resultados apresentados mostram que o sistema classificador automático implementado poderia ser utilizado para sugestão de tópicos

aos autores no momento da classificação, tendo a possibilidade de sugerir em quase 90% das vezes um tópico de interesse relevante ao autor, que muitas vezes, em um primeiro momento, poderia não associar aquele tópico ao artigo. Isso facilitaria a classificação do artigo pelo autor e poderia diminuir possíveis erros de classificação.

4.4 DIFICULDADES, LIMITAÇÕES E APERFEIÇOAMENTOS

Na etapa de extração das informações, a principal dificuldade encontrada foi com relação à padronização no formato das publicações. Os artigos submetidos ao SBSI devem seguir o padrão SBC, mas nem todos seguiram. Como exemplo, um dos artigos apresentava o termo "*abstract*" escrito de maneira incorreta, o que fez com que o algoritmo ao verificar, caractere por caractere, não encontrasse o termo desejado, e assim não conseguisse definir o momento de coletar o dado. Sendo assim, alguns dos artigos tiveram de ser descartados. Considera-se utilizar os artigos descartados em uma futura versão do classificador, após a sua correção.

Com relação às redes de colaboração geradas, foram necessários alguns ajustes para não prejudicar a análise. Dessa forma, antes de submeter as redes de colaboração à ferramenta Pajek, um tratamento nos arquivos foi realizado, inclusive de forma manual, em alguns casos. Um deles corresponde a ajustes nos nomes dos autores, já que alguns nomes apresentavam grafias diferentes nos diferentes artigos, prejudicando a analise da rede, já que um mesmo autor aparecia em vários nós distintos da rede. Uma possível evolução do sistema é a descoberta automática dessas equivalências, como por exemplo, considerando a informação de citações possíveis constantes no currículo Lattes do pesquisador. A rede de colaboração de instituições também exigiu um tratamento, inclusive maior, se comparado à rede de autores. Isso porque o nome das instituições permite mais variações que o dos autores. Por exemplo, a inclusão do nome do departamento, do Programa de Pós-graduação e do endereço. Apesar da geração automática de um arquivo no formato aceito pelo Pajek, esses tratamentos foram realizados de forma manual, dada a dificuldade de se obter resultado satisfatório utilizando-se alguma forma automatizada.

Outro ponto a ser mencionado, refere-se à particularidade do domínio. Um mesmo artigo pode ser associado a mais de um tópico. Como consequência, foi observada uma sobreposição de termos associados a cada uma das categorias, ou seja, na etapa de pré-processamento, quando as listas de termos para cada uma das categorias foram geradas, alguns termos foram incluídos em mais de uma lista, indicando que eles são relevantes em todas. Isso pode prejudicar a classificação, já que a classificação automática de documentos, espera contar com termos que permitam distinguir um tópico de outro. Uma possível evolução do sistema, para equacionar essa problemática, é considerar uma base de treinamento, em que o atributo de classe indica, além da categoria a que o artigo está associado, a pertinência desse artigo à categoria em questão. Já que o resultado é dado em grau de pertinência, a classificação inicial

também poderia ser por grau de pertinência. Por outro lado, se esses termos são relevantes a mais de uma categoria, então os documentos ondes eles são observados são associados a todos esses tópicos. Casos como esse sendo muito frequentes podem levantar questionamentos sobre se a definição das categorias não deveria ser revista, por exemplo, considerando a união de todos essas categorias com interseção em uma única. Além disso, as categorias foram associadas aos artigos a partir dos tópicos de interesse, assinaladas pelos autores no momento da submissão. Um possível erro de atribuição dos autores pode gerar resultados indesejados. Esta foi uma das motivações para o desenvolvimento do classificador, identificação de possíveis erros de classificação e verificação dos benefícios da sugestão automática dos tópicos.

Para melhorar o resultado da classificação, mudanças na abordagem têm sido experimentadas. Por exemplo, partindo do pressuposto que, ao marcar o tópico de interesse ao qual seu artigo é relevante, o autor sempre escolhe primeiro aquele com o qual o artigo possui mais semelhança, os tópicos então passaram a ser formados por artigos exclusivos. Isso quer dizer que cada artigo passa a ser relacionado a um único tópico. O objetivo era diminuir a relevância que determinados termos poderiam ter em mais de uma categoria, diminuindo a interseção em categorias.

5 TRABALHOS RELACIONADOS

Trabalhos que avaliam uma determinada comunidade ou evento a partir da rede de colaboração entre os seus atores foram propostos na literatura. Alguns exemplos são descritos nessa seção e podemos citar como principais pontos de diferença quando comparados com a nossa abordagem, o fato de: (i) nenhum deles ter avaliado a área de Sistemas de Informação; (ii) nenhum deles ter disponibilizado uma ferramenta de apoio para a geração automática da rede de colaboração e classificação das publicações o que permite analises futuras. O restante dessa seção resume esses trabalhos.

Em Daoshu *et al.* (2012) a estrutura da rede de colaboração composta pelos autores de artigos submetidos ao evento CSCWD (*International Conference on Computer Supported Cooperative Work in Design*) é analisada para que as principais características sejam identificadas. Apesar desse trabalho também utilizar uma rede de colaboração para efetuar análises, o objetivo principal dele é extrair as características do evento, entendendo como ele evolui no tempo, sem uma preocupação em avaliar uma determinada comunidade.

Em Procópio *et al.* (2011), a área de Banco de Dados é avaliada, com base nos artigos do seu principal evento nacional, o Simpósio Brasileiro de Banco de Dados (SBBD). Para tanto, uma rede de coautoria baseada nos artigos publicados no SBBD em 25 anos é construída. As suas características estruturais e a sua evolução temporal são discutidas. Algumas estatísticas, como média de artigos por autor, média de artigos por edição do

simpósio, média de coautores por artigo, também são apresentadas, permitindo a avaliação da área.

Bazzan e Argenta (2011) investigam as características da rede de coautoria construída a partir dos dados do DBLP considerando os pesquisadores membros de comitês das conferências patrocinadas pela SBC. A analise foi feita considerando métricas estruturais, como grau dos nós, maior componente conexo e coeficiente de clusterização. Foi percebido que a rede analisada não segue nenhum padrão de colaboração.

Em Freire e Figueiredo (2011), uma abordagem para indicar a importância de um pesquisador é apresentada. Nessa abordagem, a rede de coautoria construída é avaliada levando em conta grupos de indivíduos e, assim, a importância de um pesquisador em particular é medida. Essa analise é feita considerando pesquisadores da área de Ciências da Computação incluindo os ligados a instituições brasileiras.

Em Maia *et al.* (2012), a área de Distribuição e Redes é avaliada a partir dos trinta anos de publicações no principal evento da área, o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos. Uma rede de coautoria foi construída e analisada, tanto com relação as suas características estruturais, quanto a sua evolução ao longo da história.

Por outro lado, alguns trabalhos na literatura abordam a questão de classificação de artigos científicos. Em Kurach *et al.* (2013) um classificador multi-classe é aprendido, onde classificador multi-classe é um classificador capaz de associar mais de uma classe a um determinado objeto. A proposta foi avaliada considerando artigos científicos da area médica (*PubMed Central biomedical articles database*). Assim como a nossa abordagem, este trabalho permite que um determinado artigo seja associado a mais de um tópico. Entretanto, sem associar um grau de pertinência a cada um dos tópicos.

Em Corumba e Macedo (2011), um serviço Web que organiza automaticamente mensagens de *call-for-papers* recebidas em contas de correio eletrônico é descrito. As mensagens são classificadas em uma dentre seis grandes áreas da computação, a saber: Inteligência Artificial, Banco de Dados, Redes de Computadores, Teoria da Computação, Arquitetura e Engenharia de Software. Para isso, mineração de texto, mais especificamente o algoritmo K-vizinhos mais próximos, é utilizada. Além da categorização, o serviço extrai informações relevantes e recomenda *call-for-papers* baseando-se na similaridade com o currículo Lattes do usuário.

6 CONCLUSÃO

Este artigo apresentou uma abordagem para análise de uma comunidade científica a partir de mineração de textos nas suas publicações científicas. A rede de colaboração correspondente e a classificação dos tópicos relevantes à publicação são feitas de forma automática. A visualização destas informações é considerada como fonte para análise de

tendências de formação e evolução de uma comunidade de pesquisa, buscando seu direcionamento estratégico e fortalecimento. O artigo se concentrou em apresentar os resultados da aplicação das técnicas de mineração sugeridas e delinear formas de sua utilização para a análise de comunidades, tendo como exemplo o caso da comunidade nacional de Sistemas de Informação. Além disso, os principais resultados obtidos a partir dessa análise também foram descritos.

Como trabalhos futuros, pretende-se realizar novos alinhamentos das categorias para treinamento do classificador, considerando-se as categorias - tópicos de interesse - surgidas nas próximas edições do SBSI. Este alinhamento visa a eliminar as redundâncias encontradas nas modificações dos tópicos de interesse que ocorrem ano a ano e nas próprias classificações dos autores, que indicam mais de uma área relacionada ao seu artigo. Busca-se, então, que as novas categorias de treinamento sejam as mais distintas possíveis entre si, na tentativa de aumentar a eficácia da resposta.

Além disso, espera-se aprimorar a aplicação construída para aperfeiçoamento dos algoritmos utilizados por meio do: uso contínuo pela comunidade de Sistemas de Informação (incluindo a ampliação das fontes de publicações além do SBSI); construção de ferramentas que apoiem, além da mineração, a visualização de informações direcionadas às comunidades de pesquisa e ao seu público-alvo, bem como ferramentas que apoiem a análise dos dados apresentados por seus gestores. Além disso, durante a etapa de DCBD intencionamos sempre considerar uma base de artigos maior, permitindo assim melhorar a precisão do classificador aprendido.

AGRADECIMENTOS

Este artigo é resultado do projeto Redes Sociais em Pesquisa de Sistemas de Informação (rspsi.uniriotec.br), financiado pela FAPERJ e do Instituto Brasileiro de Pesquisa em Ciência da Web. O trabalho conta também com financiamento parcial do CNPq.

REFERÊNCIAS

BARABÁSI, A. *Linked. How everything is connected with everything else and what it means for business, science and everyday life*. Plume, 2003.

BATES, M. Models of natural language understanding. In: Proceedings of the National Academy of Sciences of the United States of America, v. 92, n. 22, p. 9977–9982, 1995. http://dx.doi.org/10.1073/pnas.92.22.9977

BAZZAN, A.; ARGENTA, V. Network of collaboration among PC members of Brazilian computer science conferences. *Journal of the Brazilian Computer Society*, n. 17, p. 133–139, 2011. http://dx.doi.org/10.1007/s13173-011-0033-7

BING, Liu. Web data mining: exploring hyperlinks, contents, and usage data. 2. ed. Springer, 2011.

CESI - Comissão Especial em Sistemas de Informação. Sociedade Brasileira de Computação. 2012. Disponível em: http://www.sbc.org.br/index.php?option=com_content&view=category&layout=blog&id=298&Itemid=917 Acesso:24/03/2013.

CORUMBA, D.; MACEDO, H. Categorização automática de mensagens de call-for-papers. Revista Eletrônica de Sistemas de Informação, v. 10, n. 2, 2011. http://dx.doi.org/10.5329/RESI.2011.1002006

DAOSHU, Li.; JIANGUO, Li.; YONG, T.; JINJIA, Z.; JIEMIN, C. The structure analysis of the CSCWD conference's collaboration network. In: Proceedings of the IEEE 16th International Conference on Computer Supported Cooperative Work in Design. p. 713-718, 2012.

FAYYAD, U. M.; PIATETSKY-SHAPHIRO, G.; SMYTH, P.; UTHURUSAMY, R. *Advances in knowledge discovery & data mining*. 1. ed. American, 1996.

FELDMAN, R.; SANGER, J. *The text mining handbook: advanced approaches to analyzing unstructured data*. Cambridge University Press, 2007.

FREIRE, V.; FIGUEIREDO, D. R. Ranking in collaboration networks using a group based metric. *Journal of the Brazilian Computer Society*, n. 17, p. 255-266, 2011. http://dx.doi.org/10.1007/s13173-011-0041-7

KURACH, K.; PAWLOWSKI, K.; ROMASZKO, L.; TATJEWSKI, M.; JANUSZ, A.; SONNGUYEN, H. Multi-label classification of biomedical articles. In: Intelligent Tools for Building a Scientific Informaiton Platform, Springer p. 199-214, 2013. http://dx.doi.org/10.1007/978-3-642-35647-6_15

MAIA, G.; GUIDONI, D.; SILVA, T.; SOUZA, F.; MELO, P.; SOARES, C.; ALMEIDA, J.; LOUREIRO, A. Análise da rede de colaboração do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos: As primeiras 30 edições. In: Anais do XXX Simpósio Brasileiro de Redes de Computadores, pp. 14-27. 2012.

MANNING, C.; SCHUETZE, H. *Foundations of statistical natural language processing*. MIT Press. 1999.

MANNING, C.; RAGHAVAN, P.; SCHUETZE, H. *Introduction to information retrieval*. Cambridge Press, 2008. http://dx.doi.org/10.1017/CBO9780511809071

MITCHELL, T. *Machine learning*. McGraw Hill, 1997.

OLIVEIRA, E. A.; DIAS, V. M. F. Redes sociais do SBSI e o corte de vértices como base para identificar atores importantes na coesão de grupos de pesquisa. In: Anais do Simpósio Brasileiro de Sistemas de Informação. São Paulo: Sociedade Brasileira de Computação. 2012.

OLIVEIRA, J. P. M.; LOPES, G. R.; MORO, M. M. Academic social networks. In: ER Workshops 2011: 2-3, 2011. http://dx.doi.org/10.1007/978-3-642-24574-9_2

PROCÓPIO, P. S.; LAENDER, A. H. F.; MORO, M. M. Análise da rede de coautoria do Simpósio Brasileiro de Bancos de Dados. In: Anais do Simpósio Brasileiro de Banco de Dados, 2011.

REIJERS, H. A., SONG, M., ROMERO, H., DAYAL, U.; EDER, J.; KOEHLER, J. A collaboration and productiveness analysis of the BPM community. In: Proceedings of the Business Process Management Conference, Lecture Notes in Computer Science 5701, p. 1-14, Springer-Verlag, 2009. http://dx.doi.org/10.1007/978-3-642-03848-8_1

REVOREDO, K.; ARAUJO, R. M.; SILVEIRA, B.; MURAMATSU, T. Minerando publicações científicas para análise da colaboração em comunidades de pesquisa. In: Brazilian Workshop on Social Networks Analysis and Mining, 2012, Curitiba. I Brazilian Workshop on Social Networks Analysis and Mining. Porto Alegre: Sociedade Brasileira de Computação, 2012.

SALTON, G.; BUCKLEY, C. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, v. 24, n. 5, p. 513-523, 1988. http://dx.doi.org/10.1016/0306-4573(88)90021-0

SBC. Modelos para publicação de artigos. Sociedade Brasileira de Computação. s.d. Disponível em: http://www.sbc.org.br. Acesso em: 24/03/2013.

SBSI - Simpósio Brasileiro de Sistemas de Informação. Edição 2012. Disponível em: http://www.each.usp.br/sbsi2012/. Acesso em 24/03/2013

SCRIPLATTES, s.d. Disponível em: http://scriptlattes.sourceforge.net. Acesso em 24/03/2013.

WASSERMAN, S.; FAUST, K., *Social network analysis: methods and applications*, Cambridge University Press, 1994. http://dx.doi.org/10.1017/CBO9780511815478

WIENER, E. D., PEDERSEN, J. O.; WEIGEND, A. S. A neural network approach to topic spotting. In: Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, NV, 1995), p. 317-332, 1995.

WITTEN, I. H., FRANK, E.; HALL, M. A. *data mining: practical machine learning tools and techniques*. Morgan Kaufmann, 2011.