Minerando publicações científicas para análise da colaboração em comunidades de pesquisa

Kate Revoredo[1,2], Renata Araujo[1,2], Brunno Silveira[2], Thiago Muramatsu[2]
[1]Programa de Pós-Graduação em Informática
[2]Núcleo de Pesquisa e Prática em Tecnologia
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Av. Pasteur, 458 – Rio de Janeiro – RJ – Brazil

{katerevoredo,renata.araujo,brunno.silveira,thiago.muramatsu}@uniriotec.br

Abstract. *O trabalho apresenta uma abordagem para análise de uma comunidade científica a partir de suas publicações. O objetivo é traçar o perfil da comunidade ao gerar indicadores obtidos a partir da mineração dos textos das publicações científicas. Os indicadores são gerados de forma automática, sendo: redes de colaboração e o contexto (classificação) dessas publicações. Um sistema foi desenvolvido para auxílio da análise, sendo dividido em dois módulos principais: geração de grafos e classificação automática de texto. Realizou-se um estudo de caso com a comunidade brasileira de pesquisa em Sistemas de Informação.*

Resumo. *This work presents an approach for the analysis of scientific communities from its publications. The main objective is to understand the community profile, by obtaining indicators from scientific publications through text mining. These indicators are generated automatically, and can be: a collaborative network and the publications context (classification). A system was developed to support this analysis and comprises two main modules: graphs generation and document classification. A case study with the Brazilian research community in Information Systems was conducted.*

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 que são, em sua grande maioria, derivadas da teoria dos grafos e 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. 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. Compreender os assuntos que reúnem os interesses de uma dada comunidade, bem como as formas de relação e produção existentes entre seus membros é a forma de facilitar sua caracterização, criar meios de referência interna e externa de sua existência e, em última instância, gerenciá-la estrategicamente.

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. Propostas de análise de redes sociais de pesquisa podem ser observadas [ScriptLattes][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 o 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 está na pouca estruturação de sua produção e poucos parâmetros para agrupamento de temas de interesse.

Este trabalho tem por objetivo apresentar uma abordagem e ferramental para analisar uma comunidade científica através de suas publicações científicas com base em técnicas de mineração de textos [Feldman e Sanger, 2007] para a identificação de contextos das publicações e para a geração de sua rede de colaboração. A proposta foi avaliada considerando a comunidade nacional de Sistemas de Informação, através das publicações nas edições de 2008 a 2011 do Simpósio Brasileiro de Sistemas de Informação (SBSI) [SBSI 2012][CESI 2012].

O artigo está estruturado da seguinte forma: na Seção 2 são apresentados o objetivo da mineração de textos e os tipos de informação (classificação e agrupamento) que podem ser descobertas com sua aplicação. A Seção 3 apresenta a ferramenta construída para o apoio a análise. A seção 4 apresenta os resultados obtidos com a mineração e análise da comunidade de pesquisa em Sistemas de Informação no Brasil e a seção 5 encerra o artigo apresentando conclusões e perspectivas futuras.

2. Mineração de texto

O processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases – KDD) [Witten et.al., 2011] analisa e interpreta, de forma automática, dados para descobrir padrões compreensíveis, válidos, novos e potencialmente úteis. KDD é um processo iterativo onde as primeiras etapas correspondem a uma análise exploratória dos dados, com aplicação de funções relacionadas à captação, organização e ao tratamento dos dados. Essa fase tem por objetivo encontrar as características mais relevantes dos dados, reduzir a dimensionalidade e criar o conjunto de dados de entrada,

preparando os dados para os algoritmos da etapa seguinte, a mineração de dados. A mineração de dados é considerada a etapa mais importante do processo de KDD, pois é nela que é realizada a busca efetiva por conhecimentos úteis através da extração de padrões. A última etapa é a de pós-processamento e interpretação, abrangendo a apresentação e o 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 (Knowledge Discovery in Text – KDT) [Feldman e Sanger, 2007]. Neste caso, a etapa de pré-processamento inclui técnicas de processamento de linguagem natural [Manning e Schuetze, 1999], onde a compreensão automática da linguagem é considerada transformando-a em representações manipuláveis por programas de computador, envolvendo campos da lingüística, inteligência artificial, ciência da computação e lexicografia [Bates, 1995]. A etapa de mineração de dados passa a ser denominada mineração de texto.

Comumente, através da mineração, busca-se a classificação automática de texto ou agrupamento por semelhança. A classificação ou agrupamento podem sempre ser feitos de maneira manual, mas quando a quantidade de informação disponível a ser tratada cresce, sua realização de maneira manual torna-se inviável. Este cenário influenciou estudos que possibilitam a classificação automática de textos com bons resultados. Um sistema de classificação automática de documentos deve então ser capaz de associar um documento a uma ou mais categorias pré-definidas. Para isso um classificador precisa ser aprendido durante a fase de mineração de texto utilizando algum algoritmo de aprendizado de máquina [Mitchell, 1997] com uma abordagem supervisionada, ou seja, a categoria dos textos utilizados para o aprendizado do classificador é conhecida.

Para que o classificador seja aprendido a fase de pré-processamento (preparação) dos textos precisa ser executada. Os textos são então divididos entre as categorias existentes e cada documento dessa coleção é analisado. São aplicadas algumas técnicas que facilitam o processo de seleção de características dos textos, tais como: retirada de todos os termos que não influenciam para a definição da categoria do texto, retirada de símbolos (ex: #,#,$,%,¨,&,*,(,), etc.), conversão de termos em radicais, entre outras. Em seguida, são extraídos dos textos, através de funções que determinam a relevância dos termos (como por exemplo, frequência relativa dos seus termos com relação ao documento), todos os termos que expressam melhor suas características, ou seja, os termos relevantes para a definição da categoria. Esses termos definem a lista de termos relevantes para uma determinada categoria. Um termo comum aos documentos de uma mesma categoria e incomum aos documentos das outras categorias é um bom classificador desta categoria. Essa lista de termos compõe o índice que representa a categoria. Para a seleção desses termos, cada um deles recebe um valor de quão bem servem como classificadores individuais da categoria. Este valor é chamado de escore de relevância. 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 preparação dos textos. Quando

o classificador for bem avaliado ele passa a ser utilizado para categorizar novos textos recebidos, definindo então a fase de categorização propriamente dita.

Na fase de categorização, o novo texto a ser classificado também passa pela etapa de preparação já mencionada, onde uma possível função é utilizada para extrair a lista de termos relevantes do documento, considerando uma determinada função. A categorização pode ocorrer, por exemplo, através de uma comparação entre a lista de termos das categorias e a lista de termos do novo documento (através da utilização de uma métrica de similaridade). A categoria que possuir a lista de termos mais similar à lista do documento novo será escolhida como sua classe. Existem abordagens que permitem a classificação de um texto em mais de uma categoria. Nesses casos, é indicado o grau de pertinência do texto em cada uma das categorias. Métricas de similaridade difusa são utilizadas nesses casos.

3. Apoio à mineração e análise de publicações

Para o auxílio a análise de comunidades científicas, foi desenvolvida uma aplicação que implica técnicas de mineração de textos de publicações científicas. Esta aplicação oferece dois tipos de informação para análise: a geração de grafos de colaboração e a classificação automática de artigos científicos. Para a geração da primeira foi considerada uma ferramenta já existente, Pajek[1], em conjunto com a aplicação proposta para apoio a visualização do grafo de colaboração.

Tendo em vista os contextos de análise iniciais utilizados para treinamento e avaliação desta aplicação, sua utilização está 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 grafos de colaboração

Aqui o objetivo é usar técnicas de mineração para gerar dois tipos de grafos de colaboração – colaboração entre autores e colaboração entre instituições desses autores. No primeiro caso, os vértices do grafo representam autores das publicações e existirá uma aresta conectando dois vértices se eles tiverem publicado um artigo científico em co-autoria. O peso da aresta indica a quantidade de artigos que foram publicados em conjunto (veja Figura 2). Já no segundo caso, os vértices representam instituições e existirá uma aresta conectando dois vértices se existir um artigo escrito por autores dessas duas instituições. O peso da aresta vai indicar a quantidade de artigos publicados onde a afiliação dos autores envolvidos corresponde às instituições em questão (Figura 3).

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

Para gerar os grafos de colaboração a partir de artigos científicos relacionados à área que se deseja analisar, um conjunto de artigos científicos é fornecido como entrada, estruturados em: título, autores, afiliação, resumo em português, resumo em inglês e o texto do artigo. Esses artigos devem seguir o modelo de artigos da Sociedade Brasileira de Computação [SBC] e ter formato *Adobe Portable Document Format* (PDF).

A primeira tarefa a ser executada é a de extração das informações, onde 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, artigos com um único autor são descartados. Com relação ao grafo de colaboração entre instituições, esses artigos também foram desconsiderados, mesmo que mais de uma instituição tenha sido mencionada. Além disso, empresas mencionadas na afiliação de um autor não são consideradas, a não ser que seja a única afiliação fornecida. Dois co-autores de uma mesma instituição não caracterizam uma colaboração entre instituições, logo não são considerados relacionamentos de uma instituição com ela mesma. Dessa forma, todos os grafos de colaboração gerados não apresentam laços.

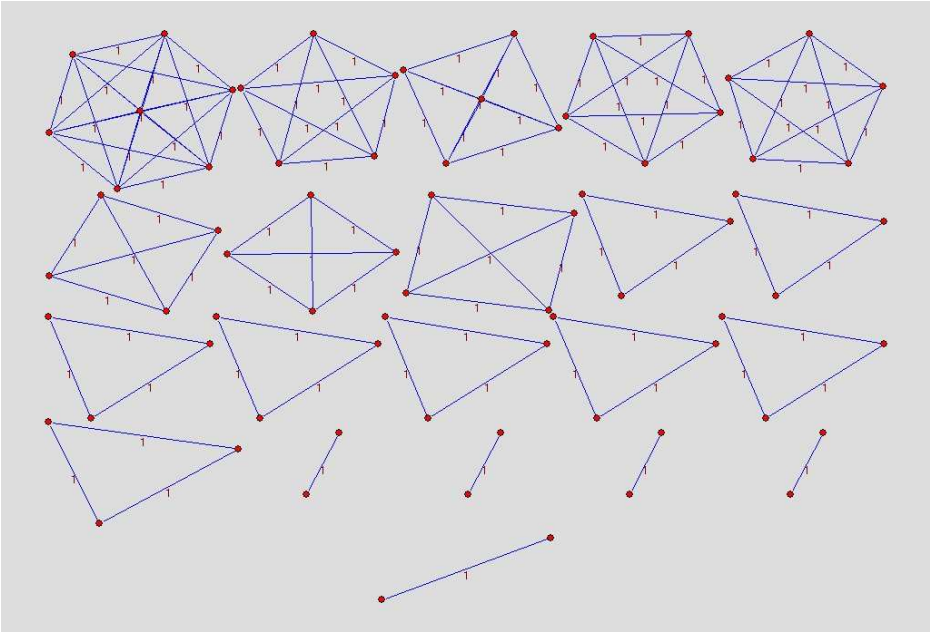

Figura 2 - Grafo de colaboração entre autores do SBSI (2008)

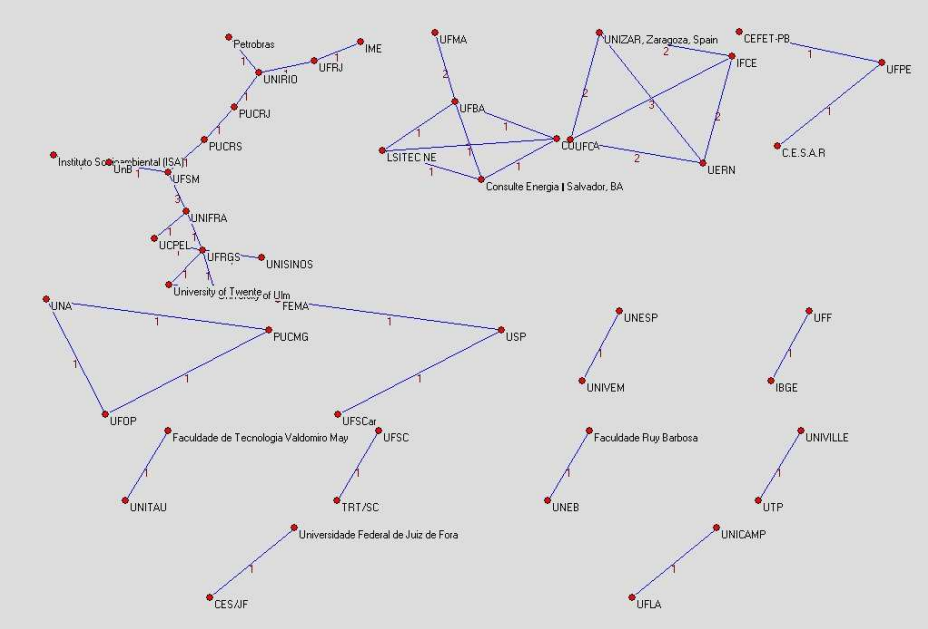

Figura 3 - Grafo de colaboração de instituições SBSI (2008 a 2010)

O segundo passo após a extração das informações trata-se da visualização do grafo, para o que foi utilizada a ferramenta Pajek. Esta ferramenta aceita como entrada um arquivo do tipo texto formatado, onde são especificados os vértices, as arestas e o peso do grafo a ser gerado. Dessa forma, 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 vértices e arestas, como alteração de cor, tamanho e posição.

3.2. Classificação de publicações

Este módulo da aplicação permite classificar os artigos científicos de acordo com os tópicos de interesse da comunidade. Como mencionado na Seção 2, antes de poder ser utilizado para classificação, um classificador precisa ser aprendido através de algum algoritmo de aprendizado de máquina. O módulo de classificação da aplicação tem então duas funcionalidades: na primeira o conjunto de artigos científicos da comunidade sendo avaliada é utilizado para aprender um classificador, é a fase de treinamento do classificador, e na segunda um determinado artigo científico da mesma comunidade (não constante no conjunto utilizado pela primeira funcionalidade) é classificado de acordo com o classificador aprendido previamente. A aplicação permite também a classificação de um conjunto de artigos científicos de uma única vez.

Para a fase de treinamento, um algoritmo de aprendizado supervisionado foi considerado, ou seja, o conjunto de artigos utilizados para o treinamento do classificador tem associado os seus tópicos de interesse. Dessa forma, é necessário que o tópico de cada artigo seja fornecido. Os artigos são então divididos por tópicos e o processo de descoberta de conhecimento em texto é executado para aprender a lista de termos que define cada uma das categorias (tópicos). Para essa tarefa foi utilizado o escore de relevância.

A partir da lista de termos associada a cada categoria é possível classificar novos artigos científicos. Primeiro os termos relevantes desse artigo são extraídos. Em seguida, a técnica de similaridade difusa é utilizada para definir a similaridade do artigo sendo classificado com cada uma das categorias através da lista de termos destas. Cada artigo pode ser definido como similar a mais de uma categoria, ou seja, ele pode estar associado a mais de um tópico.

O resumo de um artigo científico tem como característica ser conciso, expressando em poucas palavras o objetivo dos trabalhos, 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 classificação do mesmo. Com isso, a primeira tarefa executada é a de extração do resumo de cada um dos artigos.

4. Análise da comunidade de pesquisa em SI

A área de pesquisa em Sistemas de Informação tem sido vista 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 em áreas como a Ciência da Informação, Administração, Economia e Negócios, 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. 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 os artigos científicos aceitos para publicação nas edições de 2008 a 2011, totalizando 84 artigos para a análise da comunidade de SI.

4.1. Grafos de colaboração

O grafo apresentado na Figura 2 mostra a colaboração entre os autores do SBSI no ano de 2008, onde 21 artigos foram considerados. O peso da aresta representa o número de vezes em que os autores colaboraram entre si nesta edição do evento. Os nomes dos autores foram omitidos. Podemos perceber pela análise do grafo que, nos 21

artigos analisados, nenhum pesquisador foi autor em mais de um artigo. Os *clusters* apresentados são redes totalmente distribuídas, nenhum dos nós é descentralizador, 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 nós, ou seja, nenhum dos autores tem uma posição centralizadora, de poder, sobre os outros.Os grafos para as edições de 2009 e 2010 apresentam as mesmas características, com a ausência de lideranças significativas em termos de pesquisadores na área.

O grafo de colaboração de instituições para as edições 2008, 2009 e 2010 do SBSI pode ser visualizado na Figura 3. Neste, é possível perceber a existência de clusters de instituições. A análise destes clusters 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).

4.2. Classificação de publicações

Foi utilizada uma base de treinamento considerando três edições dos anais do SBSI (2008-2010) e para validação do classificador aprendido os anais da edição de 2011. Dessa forma, foram considerados 57 artigos para treinamento, e 27 para validação. Como o aprendizado é supervisionado, os artigos precisam estar associados ao seu conjunto de tópicos. Para que essa associação não fosse feita de forma manual, foi utilizada a mesma classificação associada pelos autores dos artigos quando da submissão dos mesmos. Como foram considerados os artigos referentes às edições de 2008 até 2011 do SBSI, um tratamento dos tópicos foi necessário, para alinhar tópicos sintaticamente diferentes, mas semanticamente semelhantes. Ao todo, foram levadas em consideração 30 categorias/tópicos.

Em um primeiro resultado gerado, para um total de 27 artigos analisados, a aplicação foi capaz de sugerir a classificação de 22 artigos (81% dos artigos), ou seja, encontrou semelhança entre o artigo e pelo menos um dos tópicos de interesse cadastrados. Este resultado não significa que o algoritmo sugeriu como categoria mais semelhante uma das referentes ao tópico de interesse cadastrado pelo autor. Na verdade, este resultado aconteceu duas vezes, sendo que em outras três vezes a segunda ou terceira categoria mais semelhante segundo o classificador fazia referência a um dos tópicos de interesse, que são as categorias criadas a partir da análise da base formada pelos artigos de 2008, 2009 e 2010.

Foram analisados os tópicos de interesse que o classificador não conseguiu relacionar aos artigos, ou seja, os tópicos cadastrados pelos autores e que não apareceram como semelhante (Tabela 1).

Tabela 1. Tópicos com divergências de classificação

Tópico	Número de citações pelos autores	Número de sugestões pelo classificador
“E-Business, E-Commerce e E-Government”	4	1
“Tecnologia da Informação no Governo Federal”	4	1

Modelagem conceitual de Sistemas de Informação"	6	3
"Gerência de dados e metadados/ Representação e gerência informação, dados e metadados"	3	0
"Sistemas de apoio à decisão"	7	1
"Planejamento estratégico de Sistemas e Tecnologia da Informação"	4	0

A explicação para este resultado pode estar no tamanho reduzido da base de treinamento para formação da lista de termos de cada categoria. Por exemplo, a categoria formada pelos tópicos “E-Business, E-Commerce e E-Government” e “Tecnologia da Informação no Governo Federal” teve sua lista de termos formada apenas por três artigos (ou seja, apenas três dos artigos do SBSI (edições 2008, 2009 e 2010) foram cadastrados pelos autores como referentes a estes tópicos. Esta explicação se sustenta ainda pelo fato de que, das outras categorias citadas na tabela, a que obteve o melhor desempenho foi "Modelagem conceitual de Sistemas de Informação”. Esta categoria foi formada após a análise de sete artigos existentes na base de treinamento. Relaciona-se o desempenho superior ao maior número de artigos de entrada, possibilitando a criação de um dicionário mais relevante para diferenciação da categoria em relação às demais.

Já uma categoria que teve sua lista de termos formada por um grande número de artigos obteve um comportamento diferente do classificador. A categoria "Metodologias e abordagens para engenharia de Sistemas de Informação" foi formada pelos termos de 17 artigos, sendo a categoria mais relacionada pelos autores no período. Citada oito vezes pelos autores de 2011, em todas estas vezes o classificador citou a categoria como uma das mais semelhantes, sendo duas vezes como a mais semelhante. Outro comportamento notado foi que, mesmo quando o autor não classificava seu artigo como pertencente ao tópico "Metodologias e abordagens para engenharia de Sistemas de Informação", o classificador tendia a apresentá-lo como semelhante.

Este resultado demonstra que, apesar de existirem os tópicos e os mesmos serem usados para classificação, avaliação e posterior agrupamento para sua apresentação durante o evento, o conteúdo das publicações, de fato, não aponta estes tópicos como relevantes. Isto reflete a dificuldade dos pesquisadores em classificar suas pesquisas como dentro da área de SI, pelo menos a partir dos tópicos apresentados. A definição pela área de itens conceituais e áreas temáticas ainda é difusa, característica esperada para uma área que possui grande amplitude temática e incertezas quanto à sua caracterização.

O refinamento contínuo dos tópicos obtidos em escores de relevância pode auxiliar à 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 a 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.

4.3. Dificuldades, limitações e aperfeiçoamentos

Na etapa de extração das informações, a principal dificuldade encontrada diz respeito à padronização de formato das publicações. Os artigos devem seguir o padrão SBC, mas pequenas variações no padrão, ou mesmo pequenos erros atrapalharam o processo. Como exemplo, um dos artigos apresentava o termo “abstract” escrito de maneira incorreta. O algoritmo, ao verificar caractere por caractere, não encontrava a palavra desejada, não conseguindo definir o momento de coletar a informação. Sendo assim, alguns dos artigos tiveram de ser descartados. Considera-se utilizar os artigos descartados em uma futura versão do classificador. Outra possibilidade para atenuação e/ou resolução do problema, seria a correção por parte da aplicação dos artigos com problemas na padronização.

Foi necessário realizar um pequeno tratamento antes da geração dos arquivos em formato aceito pela ferramenta Pajek. Alguns nomes de autores não apresentavam a mesma grafia em diferentes artigos e o sistema não conseguia associá-los. Após o tratamento realizado, o sistema passou a associar os que puderam ser identificados manualmente como um único autor. 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.

O grafo de colaboração de instituições exigiu um tratamento maior se comparado ao grafo de autores. Isso porque os nomes das instituições costumam aparecer das mais variadas maneiras, muitas vezes contendo o departamento, por exemplo, enquanto os nomes dos autores costumam seguir o padrão escolhido por eles para publicação. Apesar da geração automática de um arquivo no formato aceito pelo Pajek, o tratamento foi realizado de forma manual, dada a dificuldade.

A definição das categorias nunca é uma tarefa trivial, ela demanda a análise manual e é essencial para um bom resultado. Diferente de outras pesquisas na área, neste trabalho os artigos são relacionados muitas vezes a mais de uma categoria, havendo uma sobreposição de classificações. Na definição das listas de termos relevantes das categorias, são levados em consideração os termos das outras categorias, sendo que muitas vezes estamos falando de artigos comuns a ambas as categorias. Na classificação automática de texto, informações relevantes são justamente as que distinguem uma categoria de outra, mas, no caso de artigos comuns a diversas categorias, uma informação pode ser relevante a mais de uma categoria, dificultando o estabelecimento do domínio.

Além disso, os artigos da base de treinamento tiveram suas categorias, definidas a partir dos tópicos de interesse, assinaladas pelos autores no momento da submissão. Um possível erro de classificação dos autores gera resultados indesejados. Esta foi uma das motivações para o desenvolvimento do classificador, identificar possíveis erros de classificação e sugerir automaticamente novas categorias.

Para melhorar o resultado da classificação, foi experimentada uma mudança na abordagem. 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, as categorias então passaram a ser formadas por artigos exclusivos. Isso quer dizer que cada artigo passa a ser relacionado a uma única categoria para formação da lista dos termos relevantes. Nesta situação, o número de categorias levadas em consideração (ou seja, o número de categorias utilizadas no treinamento, portanto o número de categorias às quais o classificador pode relacionar) diminuiu para 24, isso porque, se antes podiam ser consideradas quatro categorias por artigo e agora apenas uma, algumas foram descartadas.

Nesta situação, 24 dos 27 artigos puderam ser classificados (89%). Este resultado não é necessariamente melhor. O que aconteceu é que, ao mudar a abordagem, considerando que cada artigo de treinamento é único a uma categoria, o classificador teve maior facilidade para diferenciar uma categoria em relação à outra. Ao mesmo tempo, o classificador passou a ter mais dificuldade de identificar as categorias pouco citadas durante o treinamento.

Ao conseguir diferenciar mais as categorias entre si, o total de artigos em que o classificador sugeriu um dos temas do autor aumentou, à medida que o classificador passou a conseguir relacionar mais facilmente o artigo a uma de suas categorias mais semelhantes. Porém, apresentou mais dificuldade em citar como mais semelhante um dos tópicos citados pelo autor. Na primeira abordagem foram duas vezes (sendo em outras três oportunidades a segunda ou terceira), nesta segunda abordagem, isso ocorreu apenas uma vez para a categoria mais semelhante (sendo cinco para a segunda ou terceira).

Um treinamento com uma base maior de artigos pode melhorar a precisão do classificador. Além disso, uma possível evolução da abordagem é a consideração de um grau de pertinência para cada um dos tópicos associados a um artigo. Os artigos classificados podem futuramente ser utilizados para evoluir as listas de termos associadas a cada um dos tópicos.

5. Conclusão

Este artigo apresenta uma abordagem para apoio à mineração de textos de publicações científicas visando a construção de grafos de colaboração em pesquisa e a classificação de temas de interesse. 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.

Como trabalhos futuros, espera-se: aprimorar a aplicação construída para aperfeiçoamento dos algoritmos utilizados; realizar seu uso contínuo pela comunidade de Sistemas de Informação (incluindo a ampliação das fontes de publicações além do SBSI); construir 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.

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. (2003) "Linked. How everything is connected with everything else and what it means for business, science and everyday life." Plume. USA. 1a edição.

Bates, M. (1995) Models of natural language understanding. Em: *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 92, No. 22, pp. 9977–9982.

CESI (2012) "Comissão Especial em Sistemas de Informação. Sociedade Brasileira de Computação". http://www.sbc.org.br

Freire, V. e Figueiredo, D.R. (2011) " Ranking in collaboration networks using a group based metric" . *Journal of the Brazilian Computer Society*, 17:255-266.

Maia, G., Guidoni, D., Silva, T., Souza, F., Melo, P., Soares, C., Almeida, J., Loureiro, A. (2012) "Análise da Rede de Colaboração do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos: As primeiras 30 edições", *Anais do XXX Simpósio Brasileiro de Redes de Computadores*, pp. 14-27.

Reijers, H.A., Song, M., Romero, H., Dayal, U., Eder, J. e Koehler, J. (2009) A Collaboration and Productiveness Analysis of the BPM Community. Em: *Proceedings of the Business Process Management Conference*, Lecture Notes in Computer Science 5701, pp, 1-14, Springer-Verlag.

SBC. Modelos para publicação de artigos. Sociedade Brasileira de Computação. http://www.sbc.org.br.

SBSI (2012) Simpósio Brasileiro de Sistemas de Informação. Edição 2012. http://www.each.usp.br/sbsi2012/

ScripLattes – http://scriptlattes.sourceforge.net

Feldman, R., Sanger, J. (2007) "The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data" Cambridge University Press. 1a edição.

Witten, I.H., Frank, E. e Hall, M. A. (2011) "Data Mining: Practical Machine Learning Tools and Techniques". Morgan Kaufmann. 3a edição.

Manning, C. e Schuetze, H. (1999) "Foundations of Statistical Natural Language Processing". MIT Press. 1a edição.

Manning, C., Raghavan, P. e Schuetze, H. (2008) "Introduction to Information Retrieval". Cambridge Press. 1a edição.

Mitchell, T. (1997) "Machine Learning". McGraw Hill. 1a edição.

Oliveira, J.P.M, Lopes, G.R, Moro, M. M. (2011) "Academic Social Networks". Em: *ER Workshops 2011*: 2-3