Um modelo de percepção de conhecimentos para identificação de comunidades no contexto das organizações

Eliane Maria De Bortoli (UTFPR, PR, Brasil) – elianedb@fadep.br
• FADEP (Faculdade de Pato Branco) – R. Benjamim Borges dos Santos, 21, Bairro Fraron, CEP: 85530-350, Pato Branco-PR
César Augusto Tacla (UTFPR, PR, Brasil) – tacla@cpgei.cefetpr.br

Recebido em: 20/08/08 Aprovado em: 06/10/08

Resumo
Atualmente as redes de informação têm ocupado um papel fundamental às pessoas pela interação e troca de informações que proporcionam. Essas redes se organizam em comunidades formadas, normalmente, por interesses ou objetivos afins. Esse trabalho visa apresentar um modelo para propiciar a identificação de comunidades virtuais, sejam elas pertencentes a uma ou mais organizações, tendo como base o seu contexto de atividades, mais precisamente, o conteúdo dos documentos eletrônicos manipulados durante suas atividades de trabalho. O modelo proposto é avaliado por meio da realização de experimentos prévios. Eles permitem verificar a sua aplicabilidade e fundamentação para a gestão do conhecimento organizacional aumentando a comunicação e disseminação do conhecimento entre os indivíduos e proporcinando, por exemplo, a descoberta de competências e a formação de comunidades de prática.
Palavras-chave: *Comunidades Virtuais; Comunidades de Prática; Redes de Informação; Gestão do Conhecimento.*

Abstract
Information networks have currently played a fundamental role for the interaction and exchange of information they provide to people. These nets if organize in formed communities, normally, for similar interests or objectives. This work aims at to present a model to propitiate the identification of virtual communities, is pertaining they to one or more organizations, having as base its context of activities, more necessarily, the content of manipulated electronic documents during its activities of work. The considered model is evaluated by means of the accomplishment of previous experiments that allow to verify its applicability and recital for the management of the organizational knowledge of form to increase the communication and dissemination of the knowledge between the individuals, provide, for example, the discovery of competence and the formation of communities of practical.
Key-words: *Virtual Communities; Communities of Practical, Nets of Information; Management of Knowledge.*

GEPROS. Gestão da Produção, Operações e Sistemas – Ano 3, nº 4, Out-Dez/2008, p. 73-87

1. INTRODUÇÃO

As redes de informação têm ocupado um papel fundamental na sociedade atual. Essas redes estão cada vez mais desenvolvidas, gerando um fluxo de dados bastante elevado. Segundo Tuomi (1999), uma informação é convertida em conhecimento quando um indivíduo consegue ligá-la a outras informações, avaliando-a e entendendo seu significado no interior de um contexto específico. Sendo assim, pode-se dizer que os dados que trafegam pelas redes de informação, transformam-se em informação à medida que o indivíduo se apropria deles, passando ao nível do conhecimento quando o indivíduo se utiliza da informação processada.

Os níveis de produtividade de uma organização são influenciados pela troca e difusão do conhecimento (LALL, 2002) e, assim, a comunicação entres os indivíduos torna-se fundamental, pois é por meio dela que o conhecimento é gerado e difundido, impactando diretamente em vantagem competitiva para as organizações.

Visando proporcionar comunicação e maior engajamento entre os indivíduos atuantes em uma ou mais organizações, as redes de comunicação têm se tornado um elemento chave. Resultados obtidos por Gupta e Govindarajan (2000) e Storck e Hill (2000) já apontaram a importância das redes de trabalho entre indivíduos de uma organização, destacando seu papel na troca de conhecimentos (especialmente tácitos). Com essa finalidade, essas redes também denominadas comunidades de prática, normalmente espontâneas e informais em relação à estrutura formal da organização, podem envolver pessoas de dentro e de fora da empresa na troca de experiências e na busca de novas abordagens para problemas comuns, continuando a existir conforme seus membros se identifiquem com o propósito do grupo (WENPIN, 2000).

A dificuldade nesse caso é identificar ou perceber pessoas que se interessam pelas mesmas áreas para realizar essa troca de experiências. Muitas vezes, até mesmo no interior de empresas de grande porte, torna-se complicado reunir pessoas com interesses ou conhecimentos em comum, pois atuam em setores diferentes ou mesmo realizam trabalhos desvinculados.

Pesquisas recentes relatam a importância de se encontrar meios para facilitar a percepção em ambientes de trabalho locais ou distribuídos, enfatizando a representação de contextos de atividades (BUDZIK, 2002). Dessa forma, se tornam necessários meios para a representação de contextos de atividade facilitando a percepção de aspectos em comum entre os indivíduos e, conseqüentemente, a criação de canais de comunicação, o que pode ser facilitado pela identificação de comunidades.

Sendo assim, esse trabalho visa apresentar um modelo para propiciar a identificação de comunidades virtuais, sejam elas pertencentes a uma ou mais organizações, tendo como base o seu contexto de atividades, mais precisamente, o conteúdo dos documentos eletrônicos manipulados durante suas atividades de trabalho. Uma comunidade virtual é um grupo de pessoas com interesses comuns que usam a Internet (sites web, e-mail, programas de mensagens instantâneas, etc.) para se comunicar, trabalhar juntos e buscar a realização de interesses (LEFEVER, 2003).

Este trabalho está dividido em seções. Na seção 2, são apresentados conceitos de percepção e na seção 3 estão definições associadas a comunidades de prática. Na seção 4 apresentam-se os trabalhos relacionados, a seção 5 descreve o modelo conceitual de percepção proposto e a seção 6 apresenta o diagrama funcional. A seção 7 apresenta o modelo experimental e descreve os experimentos realizados e os resultados obtidos. A seção 8 contém as considerações finais do trabalho.

2. PERCEPÇÃO E CONTEXTO

Quando duas ou mais pessoas colaboram na realização de alguma atividade, uma das principais dificuldades encontradas é a falta de conhecimento do contexto das atividades dos demais indivíduos. Um contexto consiste em alguma informação que pode ser utilizada para caracterizar a situação de uma entidade. Uma entidade é uma pessoa, lugar ou objeto considerado relevante para os usuários e as aplicações (DEY e ABOWD, 2000).

Usuários que trabalham juntos precisam de informações adequadas sobre o ambiente cooperativo, tais como: presença de outros membros e atividades, compartilhamento de artefatos, entre outros (GROSS e PRINZ, 2004). A percepção de contextos de atividades permite que os indivíduos ajam de forma pró-ativa no intuito de colaborar entre si. Em se tratando de comunidades virtuais no âmbito de uma organização, a percepção do contexto de atividades de um grupo de indivíduos possui papel fundamental na identificação de pessoas com interesses em comum ou que possuam conhecimentos ou competências relevantes para os demais indivíduos da organização. A percepção pode ser classificada dentro de quatro categorias principais (GROSS; STARY e TOTTER, 2005):

a) **Informal:** a percepção informal é o conhecimento de quem está ao redor, o que estas pessoas estão fazendo e o que provavelmente irão fazer. Essas informações podem ser conseguidas a partir do contexto de trabalho de cada indivíduo. A percepção informal é um pré-requisito para a interação espontânea.
b) **Social:** trata da percepção de diferentes tipos de informações subjetivas como interesse, atenção ou estado emocional de um indivíduo. Isso é freqüentemente percebido através de estímulos não verbais, ou seja, pelo contato visual, expressão facial e linguagem corporal.
c) **Grupo-estrutural:** esse tipo de percepção inclui informações sobre o próprio grupo e seus membros, tais como os papéis e responsabilidades dos membros, o posicionamento e o estado de um membro em relação a determinados assuntos ou em relação a um artefato compartilhado e os processos do grupo (ELLIS e GIBBS, 1991).
d) **Espaço de trabalho:** inclui informações sobre o espaço de trabalho em geral, tais como interações de outros participantes no espaço compartilhado e os artefatos nele contidos.

O modelo proposto neste artigo proporciona uma mescla de percepção informal e do espaço de trabalho. A percepção informal se deve ao fato de que os indivíduos terão a possibilidade de acompanhar a atuação dos demais pela representação de seu contexto de trabalho. Já a percepção de espaço de trabalho se configura pela disponibilidade de informações sobre o contexto de trabalho dos indivíduos através de seus artefatos. Estas informações contribuirão para a formação de comunidades com o intuito de favorecer a colaboração entre indivíduos.

3. COMUNIDADES DE PRÁTICA

Muitos são os exemplos de comunidades virtuais que podem ser formadas pelo agrupamento de membros de organizações dos mais diversos setores de atividade, em espaços on-line: intranet corporativa, contextos de trabalho cooperativo, sistemas de educação à distância, entre outros. Foram desenvolvidos alguns esforços de classificação das comunidades virtuais, considerando sua finalidade. Um deles seria comunidade de prática, enunciado como um grupo de pessoas com objetivos e interesses em comuns, cujo propósito é apoiar uns aos outros, aprender e promover seu entendimento através de colaboração eletrônica, empregando práticas comuns, trabalhando com ferramentas em comum, compartilhando crenças e sistemas de valores semelhantes (EFIOS, 2008).

O termo comunidades de prática é utilizado para caracterizar redes informais, formadas dentro das organizações e entre elas, as quais visam a colaboração entre os seus membros. Uma comunidade de prática é uma comunidade que aprende (PÓR, 2005).

Existem inúmeras comunidades de prática geradas no cotidiano das pessoas e que estão dispersas pelos ambientes de trabalho, de lazer, de estudos ou outros, onde as pessoas possuem uma diversidade de conhecimentos, regras de convivência determinadas e metas comuns, gerando um fluxo de informações e ações que culminam com o resultado.

Segundo Wenger (2000), a prática reside em uma comunidade de pessoas e nas relações de engajamento mútuo. Os membros de uma comunidade de prática trabalham juntos, olham uns pelos outros, conversam entre si, trocam informações e opiniões e são diretamente influenciados pelo entendimento mútuo como uma questão de rotina.

O modelo proposto neste trabalho está de acordo com a caracterização dessas comunidades como redes informais, formadas dentro das organizações e entre elas, as quais visam a colaboração entre os seus membros.

4. TRABALHOS RELACIONADOS

É possível classificar as ferramentas de identificação de comunidades em três categorias em função da forma de construírem as representações de contexto e, implicitamente, do conteúdo das representações:

a) **Análise estrutural:** é baseada na análise de estruturas de links (normalmente no ambiente web), a partir das quais é possível identificar indivíduos com interesses similares, possibilitando a formação de comunidades.

b) **Por adesão:** utiliza-se de uma interface de sistema para oferecer suporte à criação de agrupamentos (sentido de comunidade) e interação entre seus membros, ou seja, só fazem parte de uma comunidade os indivíduos que aderirem ao sistema.

c) **Análise de conteúdo:** se baseia na verificação dos conteúdos manipulados pelos indivíduos a fim de identificar similaridades entre os mesmos. Indivíduos que manipulam conteúdos similares são considerados como tendo interesses em comum e, agrupados, constituem comunidades.

A seguir, são apresentados sistemas pertencentes a esse grupo (onde se enquadra o modelo proposto) nos quais a análise de conteúdo se baseia em documentos eletrônicos. A maior parte desses trabalhos se utiliza de percepção informal e espaço de trabalho.

a) **Cyclades** (GROSS, 2003)**:** É um ambiente distribuído de arquivos virtuais para colaboração que se baseia em arquivos abertos ou públicos, provendo diversos serviços de suporte para pesquisadores. Ele faz com que os usuários fiquem cientes de pessoas que tenham procurado a mesma informação, ou produzido documentos similares, propiciando a construção de comunidades, comunicação e colaboração.

b) **I2I** (BUDZIK, 2002)**:** Objetiva oferecer comunicação oportunista entre usuários, estabelecendo um contexto compartilhado pela visitação do mesmo local na web. Esse contexto é representado por um vetor de termos extraídos dos documentos eletrônicos textuais manipulados pelos usuários. É realizada a computação de similaridade entre esses contextos, os quais são agrupados dentro de um conceito de proximidade, formando comunidades afins.

c) **CUMBIA** (VIVACQUA; MORENO e SOUZA, 2005)**:** Se utiliza de agentes para detectar oportunidades para colaboração de forma dinâmica. Trata-se de um framework, baseado em agentes, onde cada usuário dispõe de um grupo de agentes para auxiliá-lo na gestão do conhecimento e tarefas colaborativas. Para determinar um contexto de usuário, o ambiente de trabalho atual é observado e analisado por agentes que coletam informações e as transmitem aos demais do grupo. Essas informações são comparadas para encontrar similaridades e, assim, indivíduos que executam tarefas similares são referenciados uns aos outros, possibilitando a formação de comunidades.

d) **SCE** (LEE; BORODIN e GOLSMITH, 2008)**:** Trata-se de uma abordagem semântica que combina características léxicas de páginas web (focando em blogs e páginas de fóruns) com informações de hiperlinks (grafo da web) para a descoberta de comunidades. Para representar a similaridade entre duas páginas de um grafo da web são usadas as medidas das arestas aliadas ao conteúdo das páginas. Para computar a similaridade, é construído um vetor canônico de representação da página e construído um vetor TF-IDF (Term Frequency – Inverse Document Frequency) representando as características de cada página. Na redução da dimensão do espaço de termos é usado um limiar de freqüência a fim de minimizar o impacto de termos irrelevantes para o desempenho global.

5. MODELO CONCEITUAL DE PERCEPÇÃO

O modelo de percepção proposto está fundamentado no conteúdo dos recursos manipulados pelos usuários. A análise destes conteúdos permite a identificação de comunidades potenciais de indivíduos. Os recursos principais são os documentos eletrônicos textuais (ex: arquivos pdf, html) denominados artefatos textuais. Assume-se que os artefatos textuais são vestígios das atividades executadas pelos indivíduos e, portanto, podem ser utilizados para reconstruir o contexto das mesmas. O objetivo é encontrar similaridades entre os contextos que permitam o agrupamento desses usuários em comunidades afins.

Conceitualmente, propõe-se um modelo de percepção que permite aos indivíduos perceberem outros que realizam ou realizaram atividades em contextos similares.

Uma atividade é definida como um conjunto de ações (abrir um documento, conversar com um colega), realizadas por um indivíduo, visando atingir um objetivo (obter informação, construir um artefato). Um indivíduo realiza diferentes atividades ao longo do tempo, em contextos específicos que podem ser reconstruídos a partir dos artefatos textuais acessados ou produzidos.

A mesma atividade, realizada em momentos diferentes, pode apresentar contextos diferentes. Por exemplo, a atividade "manter-se informado sobre política", em um dia, pode-se acessar as páginas de política do jornal A porque houve uma notícia sobre um escândalo de corrupção na câmara dos deputados e, num outro dia, as páginas de política do jornal B para manter-se atualizado sobre as atividades dos candidatos à presidência da república. Assim, o contexto de uma atividade contém informações que descrevem o cenário no período de execução da atividade. Neste cenário, encontram-se:

a) Data, horário, local geográfico e físico (GROSS e PRINZ, 2004).
b) Os artefatos manipulados, como: documentos e imagens (GROSS e PRINZ, 2004).
c) As ferramentas utilizadas para manipular os artefatos, tais como, softwares aplicativos.
d) As pessoas envolvidas na atividade, seus estados emocionais, a sub-rede social destas pessoas com as relações inter-pessoais, informações de reputação e confiança.
e) Fatos e eventos que influenciam a execução da atividade, como, por exemplo, durante a redação de um relatório, receber um e-mail adiando a entrega do mesmo, assim a atividade pode temporariamente ser suspensa (GROSS e PRINZ, 2004).
f) Seqüência de ações executadas, como consultar páginas web, criar ou editar documentos.

Do ponto de vista de um indivíduo, diversos contextos de atividades podem coexistir em função das atividades em execução ou já executadas. Dois indivíduos quaisquer que realizam o mesmo tipo de atividade dificilmente terão contextos exatamente iguais. Portanto, a percepção de contextos deve buscar os similares (não exatamente iguais) e deve ser assíncrona (as atividades não precisam ocorrer ao mesmo tempo).

Em relação à análise temporal, o estudo da evolução (alterações) das comunidades ao longo do tempo possui diversas aplicações nas organizações:

a) Descoberta de lideranças: uma comunidade que permanece estável em relação aos seus membros apresenta, muito provavelmente, um membro condutor/líder.
b) Descoberta de competências: comunidades podem revelar competências existentes até então desconhecidas na organização. Os indivíduos podem desenvolver atividades paralelas relacionadas a temas que podem ser de interesse da organização.
c) Mapeamento do capital intelectual: a identificação de comunidades transversais aos departamentos de uma organização pode mostrar as áreas de conhecimento da mesma e a quantidade de pessoas envolvidas/interessadas por área. Isso pode auxiliar a determinar áreas para investimento e pessoas para desenvolver competências necessárias à organização.
d) Mensurar a evolução do conhecimento: o aumento/diminuição dos membros de uma comunidade ao longo do tempo pode indicar o nível de evolução do conhecimento em uma organização e a variação do número de comunidades novas ou extintas ao longo do tempo.

O modelo conceitual de percepção proposto visa identificar comunidades potenciais existentes a partir de uma população de indivíduos que executam atividades não modeladas à priori. A identificação de uma atividade é feita pelo seu contexto e dos membros de uma comunidade ocorre em função do cálculo da similaridade dos contextos das suas atividades. Pelo grau de similaridade, um indivíduo pode ser identificado como participante de nenhuma ou de diversas comunidades. A análise temporal delas pode revelar informações importantes sobre o capital intelectual da organização, apontando pessoas-chaves e direções para investir em comunidades que atuam em áreas importantes para a organização.

6. DIAGRAMA FUNCIONAL

De acordo com Paliouras (2002), o trabalho de construção de comunidades virtuais tem semelhanças com o trabalho de exploração do uso da própria Internet. Isso se justifica, pois as comunidades são construídas com a coleta de dados dos usuários, durante sua interação com o sistema computacional. O objetivo é identificar padrões comportamentais e de interesse na interação e basear os modelos da comunidade nesses padrões. Segundo o autor, os estágios constituintes do processo de identificação de comunidades são: coleta dos dados, pré-processamento dos dados, descoberta de padrões e pós-processamento dos padrões. Com base nisso, foram definidos seis estágios necessários à identificação de comunidades que são apresentados no diagrama em blocos da Figura 1, os quais podem ser divididos em dois processos principais: captura do contexto de atividades e identificação de comunidades.

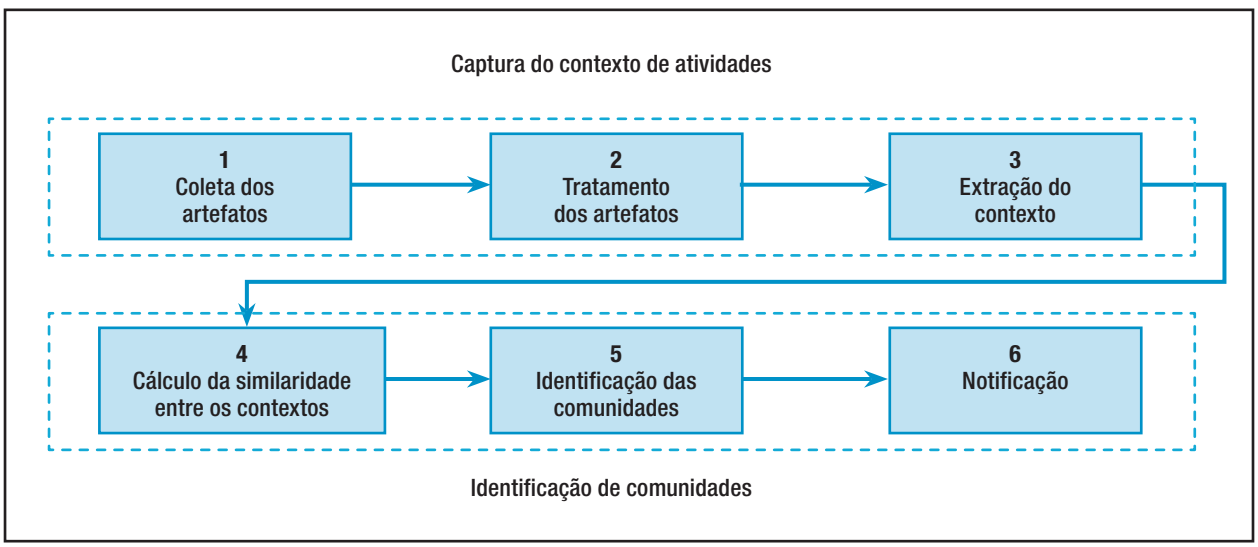

Figura 1 – Estágios necessários à identificação de comunidades. Fonte: Paliouras (2002).

a) **Coleta dos artefatos textuais:** percepção das atividades dos usuários a fim de extrair uma coleção de artefatos utilizados por eles em suas atividades.
b) **Tratamento:** seleção dos artefatos representativos da atividade em execução e preparação dos artefatos para posterior extração do contexto de atividades.

c) **Extração do contexto atual:** trata-se da submissão da coleção de artefatos textuais de cada indivíduo ao algoritmo a fim de extrair o contexto de atividades.

d) **Cálculo da similaridade entre os contextos:** realizado com base no contexto de atividades (estágio 3). Considerando os sistemas apresentados neste trabalho, na maioria das vezes, o contexto de atividades é constituído de um vetor de termos relevantes acompanhados de seus valores de TF/IDF (Term Frequency – Inverse Document Frequency).

e) **Identificação das comunidades:** obtida confrontando os valores de similaridade entre os usuários para encontrar intersecção com base em limite de similaridade pré-estabelecido.

f) **Notificação:** aviso aos usuários sobre as alterações ocorridas nas comunidades (ex: ingresso de novos membros, saída de um membro da comunidade).

7. MODELO EXPERIMENTAL

O modelo experimental consiste na representação do contexto de atividades de um usuário, o qual é composto de artefatos. Nesta implementação, utiliza-se somente artefatos textuais (at) – documentos (.txt, .doc, .pdf) e páginas web. O contexto é composto pelos dados básicos (data, hora) e pelos ats criados ou acessados pelos usuários em suas atividades. A partir da extração do contexto de cada usuário, esse experimento visa realizar o cálculo da similaridade entre os contextos obtidos e manter as comunidades.

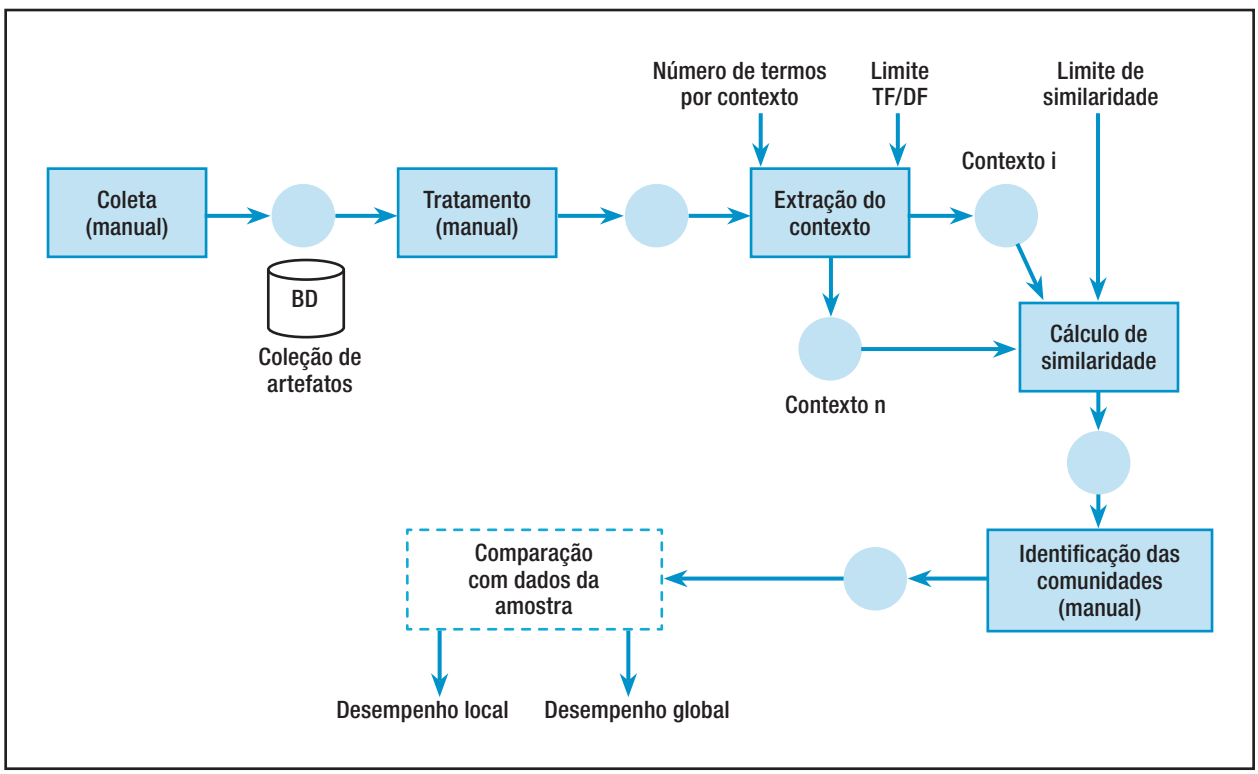

Figura 2 – Modelo experimental.

O mecanismo de percepção é constituído por dois processos principais: a construção do contexto de atividade de um usuário – que captura ats e extrai termos representativos dos conteúdos – e o processo de identificação de comunidades. A Figura 2 apresenta os elementos do modelo experimental explicados em seguida:

a) **Variáveis independentes:** número de termos no vetor de contexto e limite de similaridade para considerar dois usuários como participantes de uma mesma comunidade. As moderadoras são tipo de variável independente e indicam: tamanho da coleção de artefatos, limite TF/IDF para incluir um termo e vetor de stop words, utilizado pelo minerador de textos.

b) **Variáveis dependentes:** traduzem o resultado do modelo e são classificadas em desempenho local e desempenho global.

c) **Variáveis espúrias:** interferem no resultado do experimento e ocorrem em função de fenômenos ocasionais não previstos. O tamanho da coleção de artefatos é uma variável espúria, pois podem interferir nos valores de similaridade gerados. Se este for o caso, é preciso controlá-la, por exemplo, estabelecendo um número de documentos e uma tolerância.

d) **Coleta dos artefatos textuais:** a coleta dos artefatos textuais ocorreu de forma manual. Cada indivíduo selecionado na amostra, composta por 12 pessoas, realizou a seleção e classificação dos artefatos textuais (ats) produzidos ou acessados durante um período de 6 meses. Um at típico é um documento eletrônico textual (.txt, .doc, .pdf, .htm, .html) com número ilimitado de páginas e na língua portuguesa. Cada usuário classificou os artefatos textuais listados para cada mês, utilizados na realização de suas atividades - artefatos acessados, modificados ou criados - portanto, o número de artefatos de cada pessoa, em cada um dos meses, foi variado.

e) **Tratamento:** os artefatos foram classificados para garantir que todos estivessem em português e, em seguida, foram convertidos para o formato de texto (.txt) para que pudessem ser processados pelo algoritmo. A conversão foi realizada de forma manual, utilizando-se de aplicativos como o Microsoft Word, Adobe Acrobat e Internet Explorer.

f) **Extração do contexto atual:** os artefatos convertidos foram submetidos ao algoritmo para a mineração dos textos (remoção das stop words, stemming) e cálculo do TF/IDF para cada um dos termos retornados. Para que um termo fosse considerado num contexto, foi aplicado um limite de 0.1 para os valores de TF/IDF. Cada um dos termos foi armazenado em um vetor de contexto classificado em ordem decrescente de tamanho. Esse procedimento foi realizado a todos os indivíduos pertencentes à amostra.

As subseções a seguir, contêm os procedimentos utilizados para a extração do contexto, cálculo de similaridades e identificação de comunidades.

7.1. Extração de um contexto de atividade

Os contextos de atividades (CAs) são representados como vetores de termos relevantes. Uma medida comum de relevância para termos é a TF/IDF (Term Frequency – Inverse Document Frequency) (SALTON, 1989). TF/IDF especifica que a relevância de um termo em um artefato textual (at) está em proporção direta para sua freqüência no at e em proporção inversa à sua incidência em toda a coleção de ats, representada por AT.

O elemento IDF para o i-ésimo termo é dado por $\log(|AT|/DF_i)$, onde DF_i é a quantidade de ats contendo o termo i. TFi designa, por sua vez, a freqüência do i-ésimo termo em um at particular. A fórmula TF/IDF é dada pela equação 1.

$$TFIDF(i) = TF_i \times log\left(\frac{|AT|}{DF_i}\right) \quad (1)$$

Um at é considerado como sendo um vetor, conforme mostra a equação 2.

$$\boldsymbol{at} = \{TF_1 * log(AT/DF_1), TF_2 * log(AT/DF_2), ..., TF_m * log(AT/DF_m)\} \quad (2)$$

Um vetor médio representa o contexto de atividade do usuário. A equação 3 retorna um vetor c (centróide) para uma coleção AT de ats pertencente a um certo usuário.

$$c = \frac{1}{|AT|} \times \sum_{at \in AT} at \quad (3)$$

Tabela 1 – Contextos das atividades dos usuários 1, 2 e 3.

Termo	c_1	c_2	c_3
T_1	0,6361		
T_2	3,2283		
T_3	0,4771	0,9542	
T_4		0,2347	0,4771
T_5		0,6361	0,2347
T_6			0,3180
T_7			0,6361

Cada vez que um at é gerado e adicionado a um contexto de atividade, ou um at já existente no contexto é modificado, torna-se necessário atualizar os seus vetores de representação. Cada vez que um contexto (o vetor dado pela equação 3) sofre alterações, há alteração dos valores de similaridade e a necessidade de atualização dos seus vetores.

A tabela 1 ilustra a representação dos contextos de atividades para três usuários (c_1, c_2 e c_3), obtidos com a equação 3. Assim, o usuário 1 possui um contexto formado pelos termos T_1, T_2 e T_3. Cada posição dessa tabela contém o valor médio do TF/IDF por termo e por usuário.

7.2. Cálculo da Similaridade

Os contextos de atividade de cada usuário possuem diferentes termos e, conseqüentemente, diferentes dimensões. Naturalmente, os contextos de atividade podem possuir termos comuns, dependendo da similaridade dos conteúdos dos seus ats (ex: religião e ateísmo). Para realizar o cálculo da similaridade, primeiramente é preciso normalizar os contextos de atividade para comparar e descobrir quais termos melhor discriminam esses contextos. Um termo que é comum, ou seja, importante para muitos usuários (ex. "projeto") não seria um bom discriminador.

7.2.1. Cálculo do poder de discriminação dos termos

Para medir o poder de discriminação dos termos encontrados nos contextos de atividade é utilizada a técnica de índice Gini (SHANKAR e KARYPIS, 2000). Para o seu cálculo considera-se:

a) $\{c1, c2, ..., cm\}$ como sendo um conjunto de contextos de atividade computados de acordo com a equação (3);
b) T_i é o vetor derivado a partir da relevância do termo i em todos os contextos – $T_i = \{c_{1i}, c_{2i}, \ldots, c_{mi}\}$;
c) T'_i é o vetor normalizado, tal que $T'_i = \{c_{1i} / \| T_i \|_1, c_{2i} / \| T_i \|_1, c_{mi} / \| T_i \|_1\}$ e $\| T_i \|_1$ é a norma unitária do vetor T_i (somatório do módulo de todos os elementos do vetor T_i);
d) o poder de discriminação do termo *i* – denominado *pi* – é dado pela equação 4.

$$p_i = \sum_{j=1}^{m} T'^{2}_{ij} \qquad (4)$$

Para cada termo *i*, *pi* é igual ao somatório dos quadrados dos elementos do vetor T'_i. O valor de pi está sempre no intervalo $[1/m, 1]$. *pi* apresenta valor mais baixo quando $T'_{1i} = T'_{2i} = \ldots = T'_{mi}$, ou seja, quando o termo possui a mesma relevância em todos os contextos. O valor mais alto de pi ocorre quando apenas um contexto de atividade possui o termo *i*.

Tabela 2 – Vetores normalizados T'_i e índice Gini.

Termo	c_1	c_2	c_3	p_i
T'_1	1,0000			1,0000
T'_2	1,0000			1,0000
T'_3	0,3333	0,6667		0,5555
T'_4		0,3298	0,6702	0,5579
T'_5		0,7304	0,2696	0,6061
T'_6			1,0000	1,0000
T'_7			1,0000	1,0000

Seguindo o exemplo apresentado na tabela 1, o cálculo do *pi* é ilustrado na última coluna da tabela 2, para cada um dos termos de acordo com a equação 4. Os vetores T'_i, necessários para o cálculo de *pi*, também estão ilustrados na tabela 2 (os elementos da coluna *pi* não fazem parte dos vetores T'_i). Nota-se que os termos T_1, T_2, T_6 e T_7 são os melhores discriminadores, pois só aparecem em um dos contextos.

7.2.2. Similaridade

Para quantificar a similaridade entre dois contextos de atividade c_1 e c2, é criado um vetor comparável c'_2, da seguinte forma: para cada termo c_{1i}, o correspondente c'_{2i} é comparado com o c_{2i}. Quando um termo c_{1i} existe em c_2, então c'_{2i} é o resultado de mínimo ($c_{1i,}$ c_{2i}), caso contrário, atribui-se zero à c'_{2i}. Termos existentes somente em c_2 não são copiados para c'_2. Isso significa que, para calcular a similaridade entre contextos, é preciso construir vetores comparáveis de mesma dimensão. Dessa forma, dando seqüência ao exemplo das tabelas 1 e 2, a tabela 3 mostra estes vetores tomando-se como vetor base o contexto do usuário 1 (c_1).

Tabela 3 – Vetores para comparação com o contexto c_1.

Termos	c'_1	c'_2	c'_3
T_1	0,6361	0	0
T_2	3,2283	0	0
T_3	0,4771	0,4771	0

A similaridade entre c_1 e c_2 é computada utilizando-se do poder de discriminação dos termos (*pi*), de acordo com a equação 5. Assume-se que o valor máximo de similaridade é alcançado quando se compara um vetor c_i com ele mesmo. Daí a utilização de mínimo(c_{1i}, c_{2i}) na composição de c'_2 no parágrafo anterior.

$$similaridade\ (c_1, c'_2, p) = \frac{\sum_{i=1}^{|c_1|} c_{1i} \times c'_{2i} \times p_i}{|c_1|} \qquad (5)$$

A tabela 4 mostra os resultados da aplicação da equação 5 para o exemplo. Pode-se constatar que a maior similaridade é obtida quando se compara o vetor c_1 com ele mesmo. Em seguida, c_1 com c_2 e, c_1 com c_3, que não tem nenhum termo comum com c_1 e apresenta similaridade zero.

Tabela 4 – Cálculo da similaridade.

Termos	c_1 x c'_1	c_1 x c'_2	c_1 x c'_3
T_1	0,40470	0	0
T_2	10,4221	0	0
T_3	0,12646	0,126469	0
Similaridade	10,9533	0,126469	0

7.3. Identificação de Comunidades

Para identificar uma comunidade é preciso definir um limite de similaridade mínimo entre os contextos dos indivíduos. Um exemplo é ilustrado na tabela 5. Ela mostra os valores obtidos, a partir do cálculo da similaridade, quando utilizado um vetor de 200 termos para representar os contextos de atividade dos indivíduos durante 6 meses. A tabela é lida linha a linha, assim, a linha U1 contém os valores de similaridade de contexto do indivíduo U1 com os demais. A identificação das comunidades é realizada manualmente pelo procedimento:

a) Cálculo do percentual relativo de similaridade do usuário i (linha) em relação aos demais usuários j (coluna) para todo i ≠ j (equação 6). O valor de similaridade do usuário i com o usuário j é divido pelo maior valor de similaridade do usuário i, excetuando a similaridade do usuário i em relação a ele mesmo. Os resultados obtidos para os valores listados na tabela 5 são apresentados na tabela 6.

$$percentual\_relativo_{i,j} = \frac{similaridade_{i,j}}{máximo(similaridade_i)} \quad (6)$$

b) A partir dos valores percentuais relativos, utiliza-se um limite de similaridade para identificar as comunidades.

Tabela 5 – Valores de similaridade para contextos de atividade (centróides) com 200 termos

Usuário	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12
U1	**271,136**	4,238	24,694	2,436	6,566	20,005	3,232	21,948	6,752	3,865	3,160	8,548
U2	4,238	**68,942**	3,751	1,151	1,569	27,410	1,344	2,367	1,377	1,012	1,614	1,459
U3	24,694	3,751	**198,995**	5,732	12,222	16,748	11,949	15,422	18,729	2,675	8,738	14,729
U4	2,436	1,151	5,732	**77,434**	6,594	8,703	11,606	2,817	4,382	0,796	4,281	16,981
U5	6,566	1,569	12,222	6,594	**812,228**	12,695	6,548	11,842	16,173	6,840	4,542	8,391
U6	20,005	27,410	16,748	8,703	12,695	**1583,708**	8,932	17,007	67,304	8,039	21,052	9,504
U7	3,232	1,344	11,949	11,606	6,548	8,932	**234,429**	6,091	22,042	1,283	11,266	7,044
U8	21,948	2,367	15,422	2,817	11,842	17,007	6,091	**260,074**	9,621	5,809	3,231	5,069
U9	6,752	1,377	18,729	4,382	16,173	67,304	22,042	9,621	**987,629**	2,475	15,278	9,071
U10	3,865	1,012	2,675	0,796	6,840	8,039	1,283	5,809	2,475	**27,573**	1,473	2,538
U11	3,160	1,614	8,738	4,281	4,542	21,052	11,266	3,231	15,278	1,473	**264,446**	3,105
U12	8,548	1,459	14,729	16,981	8,391	9,504	7,044	5,069	9,071	2,538	3,105	**186,883**

Tabela 6 – Valores relativos de similaridade (%)

Usuário	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12
U1		17,16	100,00	9,86	26,59	81,01	13,09	88,88	27,34	15,65	12,80	34,61
U2	15,46		13,68	4,20	5,72	100,00	4,90	8,64	5,03	3,69	5,89	5,32
U3	100,00	15,19		23,21	49,49	67,82	48,39	62,45	75,85	10,83	35,39	59,65
U4	14,34	6,78	33,75		38,83	51,25	68,35	16,59	25,81	4,69	25,21	100,00
U5	40,60	9,70	75,57	40,77		78,49	40,49	73,22	100,00	42,29	28,08	51,88
U6	29,72	40,73	24,88	12,93	18,86		13,27	25,27	100,00	11,94	31,28	14,12
U7	14,66	6,10	54,21	52,66	29,71	40,52		27,64	100,00	5,82	51,11	31,96
U8	100,00	10,78	70,27	12,83	53,95	77,49	27,75		43,83	26,47	14,72	23,10
U9	10,03	2,05	27,83	6,51	24,03	100,00	32,75	14,29		3,68	22,70	13,48
U10	48,08	12,59	33,28	9,90	85,08	100,00	15,96	72,26	0,00		18,32	31,57
U11	15,01	7,67	41,51	20,33	21,58	100,00	53,52	15,35	72,57	7,00		14,75
U12	50,34	8,59	86,74	100,00	49,41	55,97	41,48	29,85	53,42	14,95	18,28	

Com base no exemplo dado na tabela 6, as seguintes comunidades foram identificadas quando aplicado um limite de similaridade de 40% aos valores: C1(U3, U5, U12), C2(U1, U3, U8), C3(U4, U12), C4(U6, U9), C5(U2, U6), C6(U3, U5, U8), C7(U4, U7), C8(U3, U7), C9(U5, U10), C10(U3, U5, U12) e C11(U7, U11). Considerando a identificação prévia de comunidades entre os usuários, pode-se afirmar que foi possível identificar as comunidades previamente definidas, constatando a eficiência do modelo.

8. CONSIDERAÇÕES FINAIS

Esse trabalho tratou de aspectos relacionados à identificação de comunidades através da análise do conteúdo dos artefatos textuais manipulados pelos usuários de computadores interconectados por uma rede de comunicação. Dessa forma, foi proposto um modelo conceitual que pode ser aplicado como um mecanismo de percepção para participantes de grupos, acerca das atividades dos seus colegas e, também, na identificação de comunidades. Um dos elementos que distingue o modelo conceitual apresentado de trabalhos similares é a análise da evolução das comunidades ao longo do tempo. Além disso, foi apresentada uma implementação parcial do modelo proposto, enfatizando uma técnica baseada em processamento estatístico de textos para construir e comparar contextos de atividades.

Os trabalhos futuros incluirão experimentos mais aprofundados, utilizando-se de uma população maior e mais heterogênea. Está em desenvolvimento um protótipo do sistema que funcione de maneira distribuída em uma arquitetura peer-to-peer, visando eliminar algumas das fases manuais do modelo experimental. O protótipo inclui a definição de um protocolo de notificação aos indivíduos que apresentam contextos similares. Na seqüência ocorrerá o desenvolvimento de uma representação gráfica de visualização das comunidades potenciais e a análise temporal das comunidades, o que poderá trazer informações importantes sobre o capital intelectual das organizações pela identificação de pessoas-chave, possibilitando apontar novas direções para os investimentos.

Os resultados do modelo permitem afirmar que ele gera benefícios como disseminar o conhecimento existente nas organizações, agilizar os processos, descobrir competências desconhecidas, formar comunidades de prática que favoreçam a criação de um ambiente de aprendizado, compartilhar conhecimentos e realizar o mapeamento do capital intelectual.

9. REFERÊNCIAS BIBLIOGRÁFICAS

BUDZIK, J.; BRADSHAW, S. **"Community of practice"?** Disponível em http://www.co-il.com/coil/knowledge-garden/cop/definitions.shtml. Acessado em 09/08/2008.

SHANKAR, S.; KARYPIS, G. A feature weight adjustment algorithm for document categorization, **KDD-2000 Workshop on Text Mining**, Boston, USA, August 2000.

STORCK, J.; HILL, P. A. Knowledge diffusion through strategic communities. **Sloan Management Review**, v. 41, n. 2, p. 63-74, Winter 2000.

TACLA, C. A.; ENEMBRECK, F. An awareness mechanism for enhancing cooperation in design teams. **9th International Conference on Computer Supported Cooperative Work in Design**. 2005. pp. 920-925.

TUOMI, I. Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organization memory. **Journal of Management Information Systems**, v. 16, n. 3, p. 103-117, Winter, 1999.

VIVACQUA, A.; MORENO, M.; SOUZA, J. Cumbia: an agent framework to detect opportunities for collaboration. **9th International Conference on Computer Supported Cooperative Work in Design**. 2005. p.417-422.

WENGER, E. **Communities of practice and social learning systems**. Organization, v. 7, n. 2, p. 225-256, 2000.

WENPIN, T. Social capital, strategic relatedness and the formation of intraorganizational linkages. **Strategic Management Journal**, v. 21, n. 9, Set. 2000.