**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**

**CAMPUS CURITIBA**

**GERÊNCIA DE PESQUISA E PÓS-GRADUAÇÃO**

**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E INFORMÁTICA INDUSTRIAL - CPGEI**

**ELIANE MARIA DE BORTOLI**

# MODELO COMPUTACIONAL DE PERCEPÇÃO DE CONTEXTOS DE ATIVIDADE PARA IDENTIFICAÇÃO DE COMUNIDADES

**DISSERTAÇÃO DE MESTRADO**

**CURITIBA**

**DEZEMBRO – 2006.**

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial

---

**DISSERTAÇÃO**

apresentada à UTFPR

para obtenção do grau de

**MESTRE EM CIÊNCIAS**

**por**

**ELIANE MARIA DE BORTOLI**

---

# MODELO COMPUTACIONAL DE PERCEPÇÃO DE CONTEXTOS DE ATIVIDADE PARA IDENTIFICAÇÃO DE COMUNIDADES

---

Banca Examinadora:

Presidente e Orientador:

| | |
|---|---|
| **PROF. DR. CESAR AUGUSTO TACLA** | **UTFPR** |

Examinadores:

| | |
|---|---|
| **PROF. DR. GUSTAVO GIMÉNEZ LUGO** | **UTFPR** |
| **PROF. DR. FABRÍCIO ENEMBRECK** | **PUC-PR** |
| **PROF. DR. JEAN MARCELO SIMÃO** | **UTFPR** |

Curitiba, dezembro de 2006.

**ELIANE MARIA DE BORTOLI**

# MODELO COMPUTACIONAL DE PERCEPÇÃO DE CONTEXTOS DE ATIVIDADE PARA IDENTIFICAÇÃO DE COMUNIDADES

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do grau de "Mestre em Ciências" – Área de Concentração: Informática Industrial.

Orientador: Prof. Dr. Cesar Augusto Tacla.

Curitiba
2006

Ficha catalográfica elaborada pela Biblioteca da UTFPR – Campus Curitiba

B739m Bortoli, Eliane Maria de

Modelo computacional de percepção de contextos de atividades para identificação de comunidades / Eliane Maria de Bortoli. Curitiba. UTFPR, 2006

VIII, 162 f. : il. ; 30 cm

Orientador: Prof. Dr. César Augusto Tacla

Dissertação: (Mestrado) – Universidade Tecnológica Federal do Paraná. Curso de Pós-Graduação em Engenharia Elétrica e Informática Industrial. Curitiba, 2006

Bibliografia: f. 97-103

1. Sistemas de informação. 2. Recuperação da informação. 3. Redes de computadores. 4. Ambientes virtuais. I. Tacla, César Augusto, orient. II. Universidade Tecnológica Federal do Paraná. Curso de Pós-Graduação em Engenharia Elétrica e Informática Industrial. III. Título.

CDD:658.4038

## AGRADECIMENTOS

À Deus, por ter me permitido passar por essa experiência, sempre iluminado meu caminho.

À minha família, especialmente à minha mãe, meu pai e meu irmão Marcos pelo auxílio, incentivo, compreensão e amor que me ofereceram durante todo esse período em que estive cursando o mestrado.

Ao meu namorado Peterson que sempre me incentivou, pela confiança e amor que me ofereceu durante toda essa fase da minha vida.

Às amigas Rúbia Eliza de Oliveira Schultz, Marisangela Pacheco Brites, Sílvia Letícia Zanmaria Palma, pela amizade, pelo apoio e pelos momentos de descontração proporcionados. A todas, desejo muita garra para que alcancem seus objetivos.

À amiga de longa data Beatriz Borsoi, pelo incentivo para que eu ingressasse no mestrado, pela força e confiança que sempre depositou em meu trabalho.

À todos aqueles que me incentivaram e me ajudaram, direta ou indiretamente, tais como os amigos Marlon Candido Guérios e Andrey Ricardo Pimentel. A eles meus sinceros agradecimentos, pois a ajuda prestada não apenas colaborou, mas foi necessária para a elaboração desse trabalho.

A todos os membros do Laboratório de Sistemas Inteligentes de Produção (LSIP), especialmente aos colegas Evangivaldo e Andrey, pela amizade e apoio que sempre me proporcionaram.

E por fim, alguém que não poderia deixar de listar, o meu orientador professor Dr. Cesar Augusto Tacla, cuja ajuda, dedicação e disponibilidade foram de enorme valia para a elaboração deste trabalho. Ele que me acompanhou a cada passo, ditando as direções e muitas das idéias que neste trabalho foram apresentadas. Por ter iniciado os estudos nessa área, juntamente com o professor Fabrício Enembreck, o que serviu de base para que esse trabalho pudesse ser desenvolvido.

**SUMÁRIO**

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## RESUMO

Através das redes de computadores é possível que colaboradores de uma organização que se encontram fisicamente dispersos compartilhem um mesmo ambiente de trabalho de forma virtual. Interações ocorrem nesses ambientes porque indivíduos realizam atividades que possuem pontos de intersecção, tais como relações pessoais e assuntos de interesse. Nesse caso, a percepção das atividades de outros indivíduos se torna essencial, pois o fluxo e naturalidade do trabalho são dificultados em função da distância e da impessoalidade inerentes aos ambientes virtuais. A fim de aumentar a percepção dos indivíduos que realizam grande parte de suas atividades em computadores conectados em uma rede, este trabalho apresenta um modelo conceitual de percepção que permite capturar e representar o contexto de atividade dos indivíduos e colocá-los em correlação para identificar comunidades. O modelo de percepção é baseado na análise de conteúdo dos artefatos textuais produzidos ou utilizados pelos indivíduos, permitindo-lhes perceberem outros que realizam ou realizaram atividades em contextos similares, proporcionando assim a identificação de comunidades potenciais. O modelo de percepção foi avaliado através de experimentos controlados baseados em dados coletados de uma amostra de docentes e discentes de duas instituições de ensino superior. No modelo experimental, foram utilizadas técnicas de recuperação da informação (cálculo do TF/IDF, remoção de *stop words*, *stemming*). O modelo proposto se distingue de trabalhos semelhantes, especialmente por propor a análise da evolução das comunidades ao longo do tempo, além de propor métricas para avaliar o desempenho de algoritmos de identificação de comunidades, denominadas desempenho local e global. Como resultados da experimentação, foram encontradas diferentes possibilidades de aplicação do modelo em função do desempenho local e o global, as quais são mais indicadas para certos tipos de aplicação: expansão de comunidades existentes, identificação de novas comunidades, identificação de comunidades existentes e identificação de equipes.



Palavras-chave: percepção. contexto de atividade. identificação de comunidades. recuperação de informação. sistemas colaborativos.

## *ABSTRACT*

Computer networks make possible for collaborators of an organization physically dispersed to share the same environment of work in a virtual way. Interactions happen in this kind of environment because individuals do activities having intersection points, such as personal relationships and subjects of interest. In this case, the perception of other individuals' activities becomes essential because the flow and naturalness of the work are hindered in function of the distance and of the inherent impersonality of virtual environments. In order to increase the individuals' perception that accomplish great part of their activities in networked computers, this work presents a conceptual model of perception that allows for capture and representing the contexts of the individuals' activities and to put them in correlation to identify communities. The perception model is based on contents of textual artifacts produced or used by the individuals, allowing them to notice others that do or have done activities in similar contexts, thus identifying potential communities. The perception model was evaluated through controlled experiments. Such experiments used data from a sample of teachers and students of two universities. In the experimental model, techniques of information recovery were used (TF/IDF, removal of stop words, stemming). The conceptual model distinguishes itself from similar works by taking into consideration the evolution of the communities along the time. Besides, it proposes two metrics to evaluate the performance of algorithms of identification of communities, called local and global performance. As the main result of the experimentation, different possibilities of application of the conceptual model were found as a function of the combination of the local and global performance values, which are more indicated for certain types of applications: expansion of existing communities, identification of new communities, identification of existing communities, and identification of teams.



Key-words: perception. activity context. communities' identification. information retrieval. collaborative systems.

## Capítulo I: INTRODUÇÃO

Este capítulo apresenta o contexto do problema de pesquisa a ser explorado, os objetivos gerais e específicos a serem atingidos durante o desenvolvimento deste trabalho, bem como as justificativas para a realização do mesmo. Por fim, é apresentada a estrutura deste documento.

### 1.1 Problemática de Pesquisa

É grande o número de pessoas que se utilizam do computador como ferramenta para a realização de suas atividades e que estão interconectadas através de redes de computadores, principalmente através da *World Wide Web* (WWW) ou simplesmente Web. A WEB tem sido bastante utilizada pelas pessoas como uma das principais fontes de informação para a realização de suas atividades, seja para fins de trabalhos acadêmicos ou profissionais e inúmeras são as fontes de informações disponíveis, como: periódicos, jornais e revistas *on-line*. A utilização do computador aliado às consultas na WEB em busca de subsídios para a execução de tarefas amplia os meios de criação de comunidades no âmbito de uma organização[1].

A utilização dos computadores em rede permite que colaboradores de uma organização, física ou geograficamente dispersos, compartilhem um mesmo ambiente de trabalho (virtual). Desta forma, os indivíduos, além das interações habituais do ambiente físico, tais como falar ao telefone, cruzar com um colega no corredor, no café, escutar a conversa de outros colegas na mesa ao lado ou “espiar” por cima dos ombros, experimentam um ambiente virtual onde podem se beneficiar de interações à distância mediadas por computador.

Há várias pré-condições e motivos para que uma interação ocorra. Como pré-condições têm-se a existência de um canal de comunicação e a afinidade entre os indivíduos. A afinidade pode ser suplantada pela relação hierárquica existente entre os mesmos. Por exemplo, numa organização mesmo que um indivíduo não tenha afinidade com seu superior, ambos obrigatoriamente devem interagir em um determinado momento para executar uma tarefa. Diversos outros motivos podem levar as pessoas a interagirem:

[1] Organizações são "unidades particulares formadas para atingir fins específicos, dirigidos por um poder que estabelece uma forma de autoridade que determina o *status* e o papel dos membros da organização". (RESTREPO e ANGULO, 1992).

trocar informações sobre um assunto completamente alheio (ex. lazer) ou que tenha alguma relação com a atividade da organização (ex. algum tema técnico). O fato é que indivíduos interagem não simplesmente porque existe um canal de comunicação entre eles ou porque dividem o mesmo espaço físico, mas por terem algo em comum - uma tarefa, um projeto, lazer ou tema de interesse - ou alguém em comum (ex. amigos). Resumidamente, indivíduos interagem porque realizam atividades que possuem pontos de intersecção, tais como, relações pessoais e assuntos de interesse, tanto em atividades presentes como passadas.

Num ambiente virtual a percepção é diminuída, ou seja, não se pode espiar por cima dos ombros, nem escutar conversas de corredor, observar se um colega está muito atarefado, estressado ou cansado (para decidir o melhor momento de interrompê-lo) e nem mesmo saber o que os colegas estão fazendo. Perceber é adquirir conhecimento por meio dos sentidos do que está acontecendo e do que as outras pessoas estão fazendo, mesmo sem se comunicar diretamente com elas (BRINCK e MCDANIEL, 1999).

Esta limitação da percepção em ambientes virtuais dificulta a criação de canais de comunicação entre indivíduos. Este ponto de vista é confirmado pela comunidade de CSCW (*Computer Supported Collaborative Work*) (GUTWIN e GREENBERG, 2002) que estuda formas de aumentar ou melhorar a percepção em sistemas colaborativos com o objetivo de aumentar a usabilidade[2] de tais sistemas.

Sistemas colaborativos devem prover elementos de percepção de forma a permitir a coordenação de tarefas cooperativas, principalmente onde a comunicação direta não ocorre, permitindo que indivíduos possam interpretar eventos, prever possíveis necessidades e transmitir informações de maneira organizada. Perceber as atividades dos outros indivíduos é essencial para o fluxo e naturalidade do trabalho e para diminuir as sensações de impessoalidade e distância, comuns nos ambientes virtuais. (FUKS & ASSIS, 2001). Dessa forma se fazem necessários meios de representação dessas atividades facilitando a percepção delas pelos indivíduos e conseqüentemente a criação de canais de comunicação.

[2] Usabilidade é a extensão na qual um produto pode ser usado por usuários específicos para alcançar objetivos específicos com efetividade, eficiência e satisfação em um contexto de uso específico (ISO, 2006).

## 1.2 Objetivos

Esse trabalho propõe definir um modelo conceitual computacional para identificar indivíduos que possuam interesses presentes e/ou passados em comum para fomentar a criação de comunidades. O modelo propõe a utilização dos produtos das atividades desenvolvidas pelos indivíduos, notadamente documentos eletrônicos, e a utilização destes para representar o contexto de atividades dos mesmos, colocando-os em correlação para identificar as comunidades potenciais.

Para alcançar o objetivo geral do trabalho e validar os resultados obtidos, as seguintes atividades se apresentam:

- realizar a implementação do modelo conceitual que permita identificar comunidades automaticamente;
- definir uma metodologia de experimentação que permita analisar os resultados produzidos pela implementação do modelo conceitual e, por conseqüência, validar o modelo conceitual quanto à possibilidade de realização.

## 1.3 Justificativas

A realização desse trabalho ocorrerá a partir da execução dos objetivos descritos acima, os quais se justificam pelos seguintes aspectos:

1. Há pouca preocupação na comunidade de CSCW em construir modelos conceituais computacionais de percepção. Existem várias implementações que podem ser consideradas instâncias de um único modelo conceitual. O fato de propor um modelo conceitual facilita o entendimento por parte dessa comunidade em relação aos sistemas existentes e aumenta as possibilidades de desenvolver novas ferramentas, isto é, ferramentas que se distinguem das existentes.
2. Os trabalhos existentes apresentam implementações que não se preocupam com as atividades passadas de cada usuário (Kamahara et al. (2005); Budzik et al. (2002); Vivacqua, Moreno e Souza (2005); Choo, Detlor e Tumbull (2000) e Stenmark (2001)).

3. Baseando-se nesse aspecto, este trabalho propõe a análise da evolução das comunidades ao longo do tempo.

4. A maior parte das implementações emprega técnicas que não levam em conta o conteúdo dos documentos produzidos/acessados durante as atividades dos usuários (Almeida e Almeida (2003; Reddy e Kitsuregawa (2002); Murata (2003) e Prinz, Kolvenbach e Klockner (2002)).

5. Freqüentemente, as métricas de avaliação de sistemas de identificação de comunidades são baseadas somente na recuperação de documentos sem levar em conta se os indivíduos por trás das coleções de documentos formariam uma comunidade. O presente trabalho também utiliza métricas adaptadas da recuperação de informação (precisão e cobertura) para avaliar a qualidade local (i.e. por membros recuperados) e global das comunidades (i.e. por comunidade recuperada), no entanto não se restringe a elas. A experimentação utiliza um grupo controlado, envolvendo docentes e discentes de duas diferentes instituições de ensino superior, com os quais é possível comparar os resultados obtidos pela técnica proposta com informações reais obtidas via contato direto com os usuários.

## 1.4 Benefícios Esperados

Os sistemas destinados a agrupar indivíduos, oferecendo facilidades de interação informal de maneira automática, tem um papel de extrema importância. Segundo Alarcón e Fuller (2002), estudos apontam que a maioria das conversas que ocorrem nas organizações não são formalmente planejadas ou agendadas e que elas acontecem devido à proximidade. Tais estudos mostram ainda que essas interações informais desempenham um papel central como suporte à obtenção de conhecimentos pelos colaboradores, seu entendimento, sua adaptação e aplicação a processos e procedimentos formais em seu ambiente de trabalho. Interações informais influenciam na solução de problemas e ajudam a construir uma comunidade de trabalho. Sendo assim, esse trabalho busca:

- representar um modelo conceitual de percepção capaz de identificar comunidades potenciais existentes a partir de uma população de indivíduos que executam atividades diferentes e desconhecidas;

- identificar métricas de avaliação eficazes e inovadores aplicadas a sistemas de identificação de comunidades, as quais possam ser aplicadas a outros sistemas dessa natureza;
- obter resultados passíveis de serem aplicados em organizações reais, os quais podem, por exemplo, possibilitar: a descoberta de competências pela identificação de comunidades antes desconhecidas, a criação de um ambiente de aprendizado através da união de indivíduos de comunidades pré-existentes, aumento da interação entre indivíduos que atuam em áreas comuns pela identificação de agrupamentos dos mesmos, que possivelmente já se encontram pré-estabelecidos.

Pode-se dizer que tudo isso vem de encontro ao conceito de comunidades de prática[3], que se refere à maneira como as pessoas trabalham em conjunto ou se associam a outras naturalmente. De acordo com Terra (2005), ao oferecer um ambiente de aprendizado forte, baseado em trocas de informação síncronas ou assíncronas, as comunidades de prática se tornam um conceito bastante atraente, tanto para os indivíduos como para as organizações. É no contexto de suas múltiplas comunidades de prática (formalizadas ou não) que o conhecimento organizacional se desenvolve.

## 1.5 Organização do Trabalho

No Capítulo 2 são apresentados os conceitos fundamentais - grupos, equipes, comunidades e redes sociais – visando compreender a relação entre eles e definir os limites do modelo conceitual proposto. O capítulo 3 descreve o conceito de percepção, suas características, requisitos, tipologia, e contexto de percepção. Também apresenta como as comunidades de pesquisa utilizam esse conceito, além dos sistemas computacionais existentes que visam a formação automática de comunidades. O capítulo 4 apresenta os modelos conceitual e experimental. O capítulo 5 apresenta a metodologia de testes, os experimentos e análise dos resultados obtidos e discussões sobre a identificação de comunidades, notificação aos indivíduos, evolução das comunidades ao longo do tempo e

[3] Segundo Wenger e Snyder (2000), comunidades de prática consistem em pessoas que estão ligadas informalmente, assim como contextualmente, por um interesse comum no aprendizado e, principalmente, na aplicação prática.

trabalhos futuros. Por fim, são apresentadas as considerações finais acerca do trabalho realizado.

## Capítulo II: GRUPOS, EQUIPES E COMUNIDADES

Neste capítulo exploram-se os conceitos de grupos, equipes, comunidades e redes sociais, todos advindos das ciências sociais. O objetivo é introduzir estes conceitos fundamentais, compreender a relação entre eles e definir os limites do modelo conceitual proposto dentro da teoria mais abrangente.

### 2.1 Distinção entre Grupos, Equipes e Comunidades

Os termos grupos, equipes e comunidades se fazem presentes quando o assunto é colaboração e cooperação[4] entre indivíduos, o que torna importante defini-los adequadamente, diferenciando-os um do outro.

De acordo com Schlichter, Koch e Xu (1998), há três tipos de agrupamento: grupos, equipes e comunidades. Em geral, membros de equipes conhecem uns aos outros e colaboram para atingir uma meta comum, enquanto membros de comunidades possuem apenas interesses e preferências comuns. As equipes são normalmente formadas a partir de uma decisão de gerenciamento, selecionando os membros de acordo com seu perfil, competências e potencial de contribuição para um objetivo específico da equipe. Normalmente as equipes são agrupamentos cuja interação é forte e onde os interesses da equipe prevalecem sobre os interesses pessoais dos membros.

Quanto às comunidades, os autores dizem que essas não possuem uma meta comum e desta maneira, a interação entre os membros da comunidade normalmente é livre. Na maioria dos casos os membros não conhecem uns aos outros e os interesses pessoais prevalecem sobre os interesses da comunidade. Dessa forma, pode-se utilizar o termo grupos para comunidades onde os membros podem ou não conhecer uns aos outros e não necessariamente cooperam. Como exemplos de grupos, podem ser citados grupos de amigos ou membros de um instituto de pesquisa. A Figura 1 apresenta uma pirâmide representando os três termos.

Muitas são as definições encontradas na literatura para o termo comunidade. De acordo com Mynatt et al. (1997), comunidades são consideradas como um “agrupamento

[4] O conceito de colaboração está relacionado com contribuição. Já a cooperação, além de atingir o significado de colaboração, envolve o trabalho coletivo visando alcançar um objetivo comum. BARROS (1994).

social que apresenta diversos níveis: relações compartilhadas no espaço, convenções sociais, um sentido de parceria limitada e um ritmo progressivo de interação social".

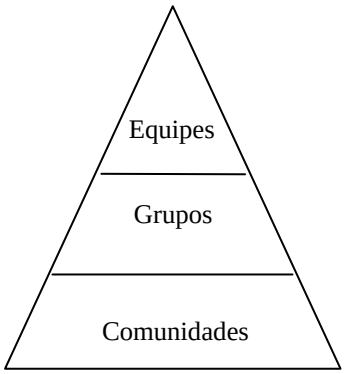



FIGURA 1 - Equipes, Grupos e Comunidades.

Já Beamish (1995), explica que o significado de comunidade gira em torno de dois sentidos mais comuns. O primeiro refere-se ao lugar físico, geográfico, como a vizinhança, a cidade ou o bairro. Dessa forma, indivíduos que habitam em um determinado lugar estabelecem relações entre si, devido à proximidade física, e vivem sob convenções comuns. O segundo sentido refere-se ao grupo social, de qualquer tamanho, que divide interesses comuns, sejam religiosos, sociais, profissionais, etc. Percebe-se então, que Beamish separa o conceito sob dois aspectos: *(i)* o do território como elemento principal na constituição da comunidade e *(ii)* o do interesse comum (e neste caso, o território comum não é mais condição para a existência das relações entre as pessoas) como núcleo da constituição da comunidade. Essa definição de comunidade proposta por Beamish (1995) pode ser reforçada pelo conceito de espaço social proposto por Bourdieu (1996), o qual afirma que quando se pensa em realidade social, pensa-se em um "conjunto de posições distintas e coexistentes, exteriores umas às outras, definidas umas em relação às outras por sua exterioridade mútua e por relações de proximidade, de vizinhança ou de distanciamento e, também, por relações de ordem, como acima, abaixo e entre".

## 2.2 Comunidades Virtuais

A Comunicação Mediada por Computador (CMC) é uma área de estudo considerada recente, afetando a sociedade, influenciando a vida das pessoas e a noção de comunidade. Por isso, muitos autores optaram por definir as novas comunidades, surgidas

no seio da CMC por “comunidades virtuais”. (RHEINGOLD, 1994; DONATH, 1999 e SMITH, 1999).

Ainda no século passado, a colaboração entre os pares já era reconhecida como uma ação imprescindível para a aprendizagem, porque expressa a heterogeneidade presente nos grupos e ajuda a desenvolver estratégias e habilidades de solução de problemas, em virtude do processo cognitivo implícito na interação e comunicação (VYGOTSKY,1998 *apud* HAETINGER, 2005). Neste sentido, segundo afirma Freire (1999) “ninguém aprende sozinho, ninguém ensina ninguém, todos aprendem em comunhão".

Além da colaboração, as ações e atitudes cooperativas podem ser identificadas na dinâmica das comunidades *on-line*. Tjara (2002 *apud* HAETINGER, 2005) relata que a cooperação requer a ocorrência dos seguintes elementos entre os participantes de um grupo: objetivos e valores comuns, trabalho coletivo, respeito mútuo, tolerância, interdependência, negociação constante, saber conviver e lidar com as diferenças e com a liderança situacional que é variável conforme a personificação dos ideais coletivos vigentes entre os membros. Nesse sentido, o autor afirma ainda que:

> As comunidades virtuais pressupõem cooperações, estimulam a participação, o desenvolvimento compartilhado. Elas negam a possibilidade de coerção. Estimulam a troca, respeitando sempre os valores preestabelecidos; prevalece a ação conjunta, estimulando ações espontâneas [...] os desequilíbrios são vistos como oportunidades de construção de novos conhecimentos, os ganhos são recíprocos. (TJARA, 2002).

De acordo com estudo realizado, Haetinger (2005) constata que a formação e a manutenção das comunidades virtuais ocorre principalmente em função da motivação dos participantes, da disponibilidade de cada um em colaborar e cooperar, do intercâmbio de suas criações e descobertas e do compromisso estabelecido entre eles. Para a formação, também são relevantes as afinidades, curiosidades ou os objetivos comuns e o desejo de compartilhar competências, habilidades e valores.

Em síntese, os seguintes benefícios podem ser trazidos pelas comunidades. (KOCH, GEORG e HILLEBRAND, 2002):

- Provêm um meio ou um canal para comunicação direta e para troca indireta de objetos de informação ou comentários em objetos dentro de um escopo comum (o espaço de informação) da comunidade. O canal de informação pode ser fortalecido com recursos que usam informações sobre os membros da

comunidade para fazer, de forma semi-automática, uma filtragem e personalização.

- Provêm percepções de outros membros e ajuda descobrir relações de parceria. Isso pode ajudar a encontrar parcerias de cooperação para interações diretas, descobertas de competências, etc.

Na opinião de Preece (2001), o termo comunidade virtual não é difícil de se entender, apesar de complicado de se definir, e significa coisas diferentes para pessoas diferentes. Esta dificuldade em se encontrar uma definição deve-se ao forte interesse multidisciplinar que o assunto inspira. De acordo com Souza e Preece (2004) "sociólogos tendem a focar em redes de relações sociais, etnógrafos nos papéis e atividades de pequenos grupos de pessoas, e os estudiosos da área tecnológica na estrutura do software subjacente que dá suporte à comunidade".

Dessa forma, muitas são as definições para comunidades virtuais apresentadas na literatura, as quais contêm diversos enfoques, entre eles: social, de negócios, de aprendizado e tecnológico.

Para Rheingold (1993), uma Comunidade Virtual é uma associação de indivíduos (os membros da comunidade, participantes ou usuários) que compartilham interesses entre si, conhecimento e objetivos, em um domínio temático específico, através da Internet.

LeFever (2003), define comunidade virtual como "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 perseguir seus interesses através do tempo". Para LeFever, comunidades *on-line* são primeiro "comunidades" e depois "*on-line*", no sentido de que a parte *on-line* é secundária em relação à comunidade do "mundo real" que utiliza a Internet.

Uma definição abrangente do termo comunidade virtual é apresentada por Preece (2000). Para o autor, comunidade virtual é "qualquer espaço social virtual onde as pessoas se reúnem para obter e oferecer informações ou apoio, para aprender ou para encontrar companhia", independente de localização geográfica (local, nacional ou internacional) e tamanho (pequeno ou grande).

Dessa forma, pode-se dizer que uma comunidade virtual se caracteriza como um agrupamento de pessoas que possuem algum tipo de relação. A comunidade virtual é um

elemento do *ciberespaço*, existente apenas enquanto as pessoas realizarem trocas e estabelecerem laços sociais. Este tipo de comunidade sustenta-se pela co-atuação de indivíduos que compartilham valores, interesses, metas e posturas de apoio mútuo, através de interações no universo *on-line*.

### 2.3 Classificação das Comunidades Virtuais

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. Por essa razão, se fez necessário classificar as comunidades virtuais de acordo com a sua finalidade, sendo um esforço de classificação apresentado a seguir (KOLLOCK, 1997):

- **de aprendizado:** como a comunidade dos estudantes e professores de um curso de graduação;
- **de prática:** como a formada pelos colaboradores de uma organização; ou
- **de lazer:** como a dos amantes da pesca, colecionadores de selos, etc.

Outra tentativa de classificação é proposta por Valtersson (2002), a qual afirma que, em função do seu objetivo, uma comunidade pode ser:

- **Virtual de relacionamentos**: Uma comunidade virtual construída em cima de relacionamentos promove tipos especiais de ligações entre pessoas, interconexões que resultam em uma harmonia peculiar, só encontrada em famílias ou grupos de amigos. Essas ligações podem ser baseadas em um interesse, objetivo ou problema comum, mas, a todo instante, a ênfase é no relacionamento construído entre os participantes. Pontos como compromisso, confiança e valores são inerentes a qualquer relacionamento que surge em uma comunidade.
- **Virtual de lugar**: Indivíduos desse tipo de comunidade usufruem o mesmo *habitat* ou local. Esse compartilhamento de um lugar com outros pode oferecer um senso de segurança e comodidade. O lugar não precisa ser físico, e em comunidades virtuais eles não o são. Pessoas de diferentes países se encontram

em um lugar virtual na Internet, por exemplo, porque estudaram em uma mesma universidade.

- **Virtual de conhecimentos**: Comunidades de conhecimentos ajudam as pessoas a encontrarem outras com os mesmos objetivos, os mesmos valores e as mesmas concepções sobre determinado assunto. Um exemplo seria o das pessoas que procuram decifrar um código secreto em algum texto. As duas principais características desse tipo de comunidade são o compartilhamento e as idéias; elas, no entanto, são expressas de uma forma impessoal e técnica. Outro exemplo claro de comunidades de conhecimentos são pesquisadores acadêmicos que se juntam para resolver algum problema científico.
- **Virtual de memória**: Uma comunidade virtual de memória é baseada em um passado compartilhado ou em algo histórico. Um exemplo forte desse tipo de comunidade é a rede que liga os sobreviventes de um Holocausto, na Internet. Sobreviventes e descendentes de sobreviventes podem conversar com pessoas que se sensibilizam com o assunto.
- **Virtual de necessidade**: Uma comunidade virtual de necessidade é baseada em algum fato ou acontecimento semelhante e que tenha algum fator que afete emocionalmente os participantes. Um exemplo desse tipo de comunidade é a rede que liga pessoas com câncer ou pessoas que perderam os filhos em acidentes rodoviários.

O desenvolvimento de comunidades virtuais de maneira eficaz e segura, independentemente de sua finalidade, requer que sejam atendidos alguns princípios. De acordo com Palazzo et al. (2001), há uma discussão sobre alguns princípios considerados fundamentais na criação e desenvolvimento de comunidades virtuais, tais princípios foram baseados em Mongoose (2001). Ainda segundo o autor, os seis primeiros princípios estão relacionados com as necessidades e expectativas individuais dos usuários, enquanto que os demais estão relacionados com a organização e estrutura necessária para a eficácia e viabilidade de uma comunidade virtual de forma geral. Os princípios de uma comunidade virtual segundo Palazzo et al. (2001) são:

1. **Propósito:** Toda comunidade virtual deve ter um mesmo objetivo. Para que todos busquem o mesmo propósito, este deve ser o tema ou foco principal da comunidade virtual.

2. **Identidade:** Os membros da comunidade virtual devem ser devidamente identificados, ou seja, todos devem poder se conhecer dentro da comunidade virtual.
3. **Comunicação:** A comunicação deve ser priorizada nas comunidades virtuais, mostrando que todos podem dar suas opiniões e ajudar a construir o conhecimento coletivo, partilhando idéias e expressando opiniões.
4. **Confiança:** Com o passar do tempo, os membros da comunidade passam a se conhecer e a confiar mais uns nos outros. Através de seus atos, manifestações e posições assumidas os membros da comunidade adquirem uma reputação perante os demais. A reputação é essencial para o estabelecimento de laços de confiança e complementa também o princípio da *Identidade*.
5. **Reputação:** Os membros da comunidade adquirem um *status* no grupo com base nas ações realizadas.
6. **Subgrupos:** A comunidade virtual pode ser dividida em subgrupos, que possuem algum interesse particular comum ligado ao objetivo central da comunidade virtual, para que seja possível estabelecer relacionamentos entre os membros do grupo para depois se unir à comunidade geral.
7. **Ambiente:** O ambiente deve ser agradável aos usuários e deve reunir as preferências e objetivos de seus usuários em um mesmo espaço em que todos possam compartilhar idéias, opiniões, sugestões e conteúdos, para que o aprendizado seja mútuo.
8. **Limites:** É necessário impor limites à comunidade virtual, de modo a saber sempre quem faz parte da comunidade e quem não faz, para que se possa ter um controle do sistema.
9. **Governo:** Nas comunidades virtuais o governo é na verdade uma forma de autogoverno. É importante que cada membro da comunidade se sinta responsável por seu desenvolvimento. Esta forma de administração é mais eficiente, escalável e simpática do que qualquer outra.
10. **Intercâmbio:** A troca de informações deve ocorrer sempre, facilitando a interação entre os usuários e aumentando o conhecimento, ao ponto em que

software e outros recursos são disponibilizados para que toda a comunidade tenha acesso.

11. **Expressão:** As pessoas que quiserem expressar suas preferências e opiniões devem sempre ter um local para isso de acesso livre do grupo, ou seja, todos devem poder se expressar livremente.
12. **História:** A história está ligada principalmente à documentação da comunidade. Deve-se permitir que pessoas que deixam ou passam a fazer parte da comunidade possam saber como ela está organizada, quem faz parte dela, quem já fez, e como tudo funciona. Em qualquer momento desejado, deve-se poder perceber como a comunidade está evoluindo e o que mudou.

Pode-se dizer que esse conjunto de princípios estabelece de forma coerente os principais relacionamentos e ações que sustentam a evolução das comunidades virtuais.

Na Figura 2 demonstram-se algumas propriedades de uma comunidade virtual (PALLOFF, 1999).

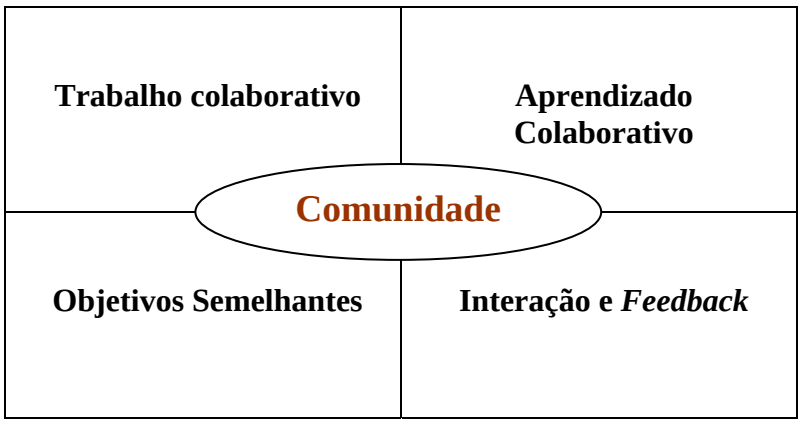



FIGURA 2 - Propriedades de uma comunidade virtual.

## 2.4 Comunidades de Prática

Conforme apresentado anteriormente existem variadas definições para comunidades virtuais, envolvendo diversos enfoques. Além disso, existem outros termos utilizados e que adotam uma descrição similar a essas definições. 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, 2004).

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. Apesar de ser considerada uma área de estudo recente, tem sido bastante explorada, sendo definida pelos estudiosos de diversas formas.

Para Wenger (2000) “uma comunidade de prática não é apenas uma agregação de pessoas definida por algumas características. O termo não é um sinônimo para grupo, equipe ou rede”. Diferente dos conceitos de grupo (as pessoas se conhecem, mas não necessariamente cooperam), equipe (as pessoas se conhecem e possuem uma meta comum) e comunidade (as pessoas não necessariamente se conhecem mas podem possuir objetivos em comum, prevalecendo os interesses pessoais), as comunidades de prática são formadas dentro das organizações por indivíduos que possuem razões ou intenções comuns, objetivando a colaboração, o aperfeiçoamento ou o aprendizado através da troca de experiências na execução de ações em comum, não levando em consideração apenas os interesses pessoais.

Pode-se dizer que uma comunidade de prática é um grupo de pessoas com o intuito de aprender, mas de acordo com Pór (2005), trata-se de uma comunidade que aprende. O autor diz que as comunidades “não surgem meramente trocando idéias ao redor de um refrigerador de água, compartilhando e beneficiando cada um dos membros com a habilidade dos outros, mas através da integração com os colegas para desenvolver conjuntamente melhores práticas”.

Com base nessas afirmações, pode-se dizer que 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. Dessa forma, a prática é tida sempre como uma prática social (WENGER, 2000).

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.

## 2.5 Identificação de Comunidades Virtuais

Comunidades geralmente são criadas por elementos que possuam algum interesse ou objetivo semelhante. Assim sendo, para a geração de uma comunidade virtual em um ambiente *on-line*, independente da organização envolvida, é necessário identificar um perfil do indivíduo envolvido nesse ambiente. Segundo KOCH (2000), que aponta uma das formas possíveis de realização, pode-se empregar uma técnica semi-automática que utiliza informações sobre os interesses do indivíduo e sobre o contexto no qual ele está inserido.

De acordo com Paliouras (2002), o trabalho de construção de comunidades virtuais na Internet é semelhante ao trabalho de exploração do uso da própria Internet pelo fato de que as comunidades são construídas com a coleta de dados dos usuários, durante sua interação com o sistema. O objetivo é identificar padrões comportamentais e de interesse na interação e basear os modelos da comunidade nesses padrões.

Segundo o autor, os estágios de descoberta dos dados da comunidade são os mesmos de qualquer outro processo de mineração de dados: coleta dos dados, pré-processamento dos dados, descoberta de padrões e pós-processamento dos padrões, descritos a seguir:

- **Coleta dos dados**: Durante este estágio, dados de vários locais são colhidos e suas estruturas e conteúdos são identificados. Os dados coletados na Internet podem aparecer de várias formas. Dependendo do tipo de acesso, a informação que foi coletada varia entre pesquisas, perfis e *logs* de navegação. Devido a essa variação, a transformação dos dados em um conjunto de objetos, mantendo a informação que é útil para a modelagem, é um trabalho de organização significativo.
- **Pré-Processamento dos dados**: Este é o estágio em que os dados são limpos de ruídos, suas inconsistências são resolvidas, sendo eles integrados e consolidados, para servirem de entrada para o próximo estágio de descoberta dos padrões. Esses dados podem ser provenientes de pesquisas, perfis e *logs* de navegação, entre outros. Mas, em todos os casos, o objetivo é o mesmo:

organizar as características que descrevem o comportamento dos usuários e revelar as similaridades entre as mesmas.

- **Descoberta de padrões**: Estando os dados na sua forma correta, padrões de interesse são descobertos com o uso de técnicas apropriadas, como agrupamento, classificação, descoberta de regras de associação, etc.

Para Paliouras (2002), esses três processos nessa ordem são necessários para a descoberta de perfis e a conseqüente geração de comunidades virtuais. Esse processo de descoberta de perfis pode ser realizado a partir da captura dos dados dos usuários que manipulam o sistema.

## 2.6 Redes Sociais

Uma rede social é constituída de nós (indivíduos) conectados por laços sociais (WATTS, 2003 apud RECUERO, 2005). WELLMAN (2001 apud RECUERO, 2005) mostra que as redes de computadores são redes sociais porque conectam pessoas e afirma que "quando as redes de comunicação mediadas por computador conectam pessoas, instituições e conhecimento, elas são redes de suporte social por computador."

Atualmente as práticas comunitárias *on-line* são cada vez mais comuns, na perspectiva da *Sociedade em Rede* (CASTELLS, 1999). Em se tratando de comunicação via computador, constituem um fenômeno historicamente recente, a ser amplamente explorado, visando compreender os elementos e fatores que favorecem a colaboração e cooperação em ambientes virtuais.

A colaboração e a cooperação manifestam a heterogeneidade presente nos grupos. Pessoas sistematicamente reunidas confrontam e reorganizam seus interesses e saberes a partir das trocas interpessoais, aproveitando a fonte de conhecimento que cada uma representa. Esse procedimento de socialização de informações, atitudes, comportamentos e inteligências variadas, gera uma intensa atividade cognitiva individual, a qual culmina com a aprendizagem (HAETINGER, 2005).

> A noção de partilha de conhecimentos está associada, também, a idéia de que todos os agentes envolvidos no sistema são potencialmente beneficiários e provedores de conhecimentos, e que cada um aprenderá com outros agentes que constituem o sistema e ajudará os outros a melhorar suas competências. Evidentemente, isso não significa que todos os agentes aproveitarão da mesma

> maneira as oportunidades de aprendizagem que lhes são oferecidas. (DEPOVER e MARCHAND, 2002 apud HAETINGER, 2005).

O homem, ao viver na coletividade, interage com seus semelhantes, avaliando e atualizando continuamente seus conhecimentos. As relações estabelecidas entre grupos heterogêneos de pessoas, cada um com seu ponto de vista, fazem com que cada integrante de um grupo converta seus pensamentos e construa novos conhecimentos. As redes sociais vêm de encontro a essa filosofia, onde há a valorização dos elos informais e das relações, em detrimento das estruturas hierárquicas.

O estudo das redes coloca assim em evidência um dado da atual realidade social e que precisa ser mais bem explorado, de que os indivíduos, dotados de uma variedade de recursos e capacidades, organizam e executam suas atividades, baseados em socializações e mobilizações provocadas pelo próprio desenvolvimento das redes. Pode-se dizer que pequenas decisões (micro) são influenciadas pela coletividade (macro), tendo a rede como intermediária.

As redes de conexões tem facilitado o trabalho das pessoas há muito tempo, sem que as pessoas as tivessem percebido como uma ferramenta organizacional. De acordo com Lipnack e Stamps (1992), “o que é novo no trabalho em redes de conexões é sua promessa como uma forma global de organização com raízes na participação individual. Uma forma que reconhece a independência enquanto apóia a interdependência. O trabalho em redes de conexões pode conduzir a uma perspectiva global baseada na experiência pessoal”.

Segundo Marteleto (2001), desde os estudos clássicos de redes sociais até os mais recentes, concorda-se que não existe uma “teoria de redes sociais” e que o conceito pode ser empregado com diversas teorias sociais, necessitando de dados empíricos complementares, além da identificação dos elos e relações entre indivíduos. A análise de redes pode ser aplicada no estudo de diferentes situações e questões sociais.

Sociólogos chamam a potencial conexão de indivíduos e comunidades de capital social. (COLEMAN, 1988). O termo se refere à posição dos atores na rede e consiste da habilidade de atrair os recursos contidos pelos membros da rede.

De acordo com Marteleto (2004), o capital social não deve ser confundido com o capital humano, nem com infra-estrutura. Segundo a autora, “o capital humano engloba as habilidades e conhecimentos dos indivíduos que, em conjunto com outras características

pessoais e o esforço despendido, aumentam as possibilidades de produção e de bem-estar pessoal, social e econômico. A infra-estrutura refere-se ao conjunto fundamental de instalações e meios para que a produção se realize e se distribua".

O capital social é definido por Marteleto (2004), como "as normas, valores, instituições e relacionamentos compartilhados que permitem a cooperação dentro ou entre os diferentes grupos sociais. Dessa forma, são dependentes da interação entre, pelo menos, dois indivíduos". Por essa razão, fica clara a existência de uma rede envolvendo o conceito de capital social. Reforçando esse conceito, Bourdieu (2000) e Coleman (1994) ao discutirem o capital social o analisam como indivíduos inseridos em uma rede de relações sociais que podem se beneficiar de sua posição ou gerar externalidades positivas para outros agentes. Ampliaram esta conceituação ao incluir relações verticais, caracterizadas por relações hierárquicas e distribuição desigual de poder.

Portanto, capital social envolve o conjunto de recursos que um indivíduo ou grupo pode obter a partir de sua posição em uma rede de relações sociais estáveis. Corresponde ao elemento que mantém a união ou harmonia das sociedades e está baseado na confiança entre as pessoas e na rede de relacionamentos entre elas e os grupos sociais que formam as comunidades.

Dessa forma, rede social é uma relação moral de confiança de um grupo de agentes individuais que têm em comum normas ou valores além daqueles necessários às transações habituais de mercado. As normas e os valores abrangidos nesta definição podem ir da simples norma de reciprocidade entre dois amigos até os complexos sistemas de valores criados por religiões organizadas. (FUKUYAMA, 2002).

Ainda em relação ao capital social, pode-se dizer que esse pode ter uma influência essencial sobre a vida das pessoas, afetando certos aspectos como, por exemplo, a procura por um trabalho e o potencial de um indivíduo em ser promovido. O capital social pode ser atribuído para um indivíduo ou para uma grande comunidade, tal como uma organização. As propriedades que constituem o capital social de um indivíduo incluem aspectos como publicidade, influência, carisma, credibilidade, *status*, posição, autoridade, centralidade, visibilidade, confiabilidade, reconhecimento e popularidade. (NIINIVAARA, 2004).

Diante desse panorama é possível dizer que atualmente o trabalho informal em rede é uma forma de organização humana presente na vida cotidiana e nos mais diferentes

níveis de estrutura das instituições modernas, sendo que a geração de redes sociais e a conseqüente aquisição de capital social dependem de fatores culturais, políticos e sociais.

Tendo em vista a importância e a complexidade da geração das conexões que surgem, surge também uma percepção natural da necessidade do suporte tecnológico, em especial de melhores técnicas computacionais, para gerenciá-las de forma eficaz. Resnick e Varian (1997), usam o termo "capital sócio-tecnológico" para se referir a recursos produtivos que são inerentes ao modelo das relações sociais que são representadas e mantidas com o suporte das tecnologias da informação e comunicação.

A partir do ponto de vista do capital sócio-tecnológico, um sistema deve habilitar os usuários para, entre outros aspectos (NIINIVAARA, 2004): compartilhar e trocar recursos, criar conexões, formar grupos de interesse, encontrar outras pessoas e experiências, utilizar o conhecimento dos outros e coordenar o processo de comunicação e ações interdependentes.

A relação existente entre as redes sociais e a formação de comunidades está na formação do capital social, especialmente das relações sociais que permitem a cooperação entre indivíduos de diferentes grupos sociais. Levando em consideração que uma rede social consiste de uma rede informal, onde cada indivíduo constrói e compartilha seu capital social, fica favorecida a identificação de comunidades, nas quais os indivíduos possuam elementos de seu capital social que sejam comuns a outros indivíduos, culminando na cooperação entre os mesmos. Dentro dessa visão, Marteleto (2004) comenta que muitos são os tipos de redes que podem ser formadas, dependendo das características intrínsecas a elas, tais como: diversidade dos participantes, institucionalização de normas de decisão, objetivos gerais ou específicos, tamanho e área geográfica, etc.

## 2.7 Posicionamento e Discussão

O trabalho em questão é direcionado à identificação de comunidades. Para tanto, será adotado o conceito de comunidades virtuais proposto por Preece (2001): "qualquer espaço social virtual onde as pessoas se reúnem para obter e oferecer informações ou apoio, para aprender ou para encontrar companhia", independente de localização geográfica (local, nacional ou internacional) e tamanho (pequena ou grande). Complementando a conceituação anterior, pode-se dizer que a abordagem está

especialmente relacionada à afirmação de Kock, Georg e Hillebrand (2002), os quais definem que as comunidades provêm percepção de outros membros e ajudam a descobrir relações de parceria, o que pode ajudar a realizar parcerias de cooperação para interações diretas, descobertas de competências, entre outros benefícios.

Apesar de não ser restrito apenas ao ambiente organizacional – onde serão realizados os experimentos necessários para a validação da proposta - esse trabalho também relaciona-se ao conceito de comunidades de prática defendido por Efios (2004) que a define como um conjunto de pessoas com objetivos e interesses em comum, 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. Sendo assim, pode-se dizer que comunidades de prática se diferenciam do conceito de comunidades propriamente dito, principalmente pelo fato de proporcionarem um engajamento mais forte entre os indivíduos, os quais compartilham interesses, experiências e propósitos de uma prática em comum. O conceito de comunidade de prática se assemelha muito ao conceito de equipe, com a diferença de que em uma equipe os indivíduos necessariamente conhecem uns aos outros e cooperam para uma meta comum e em uma comunidade de prática não é necessário que os indivíduos se conheçam, havendo objetivos comuns, mas não uma produção final em comum.

Não serão considerados explicitamente os conceitos de redes sociais apresentados anteriormente, mesmo sabendo que uma comunidade é intrinsecamente uma rede social. Nesse momento, de um ponto de vista puramente tecnológico, pretende-se apenas identificar indivíduos potenciais para a formação de comunidades, não importando definir o tipo e o grau de relação existente entre os indivíduos pertencentes às comunidades.

Como o trabalho será restrito basicamente à identificação de indivíduos que tenham interesses em comum, visando oferecer uma maneira facilitada de agrupá-los em comunidades afins, alguns dos princípios das comunidades virtuais definidos por Palazzo et al. (2001) não serão utilizados na abordagem proposta, ou seja, nesse momento não haverá preocupação com o fato de que todos os membros tenham os mesmos propósitos ou com o *status* que possuem no grupo, reputação e confiança. Estas três últimas propriedades podem ser tratadas futuramente evoluindo a abordagem proposta para uma que utilize redes sociais.

Neste capítulo foram estudados os conceitos fundamentais de comunidades, sendo possível perceber que um dos componentes essenciais para a formação de uma comunidade é a existência de pessoas com interesses ou objetivos em comum. Outras características são importantes também, tais como afinidade, desejo de compartilhar competências, habilidades e, respeito mútuo, tolerância, valores comuns e interdependência (HAETINGER, 2005; TJARA, 2002), no entanto, características que representam aspectos subjetivos da personalidade humana, tais como tolerância, desejo de compartilhar e respeito mútuo, são dificilmente representáveis num sistema computacional. Por isso, a abordagem proposta será fundamentada na identificação de indivíduos que possuam interesses em comum, sendo que será construída uma representação computacional destes interesses a partir dos vestígios (artefatos textuais) das operações realizadas no computador (acesso a WEB, acesso a documentos locais, *e-mail*). Assim, a base dessa abordagem está na coleta destes vestígios, na representação dos interesses e no estabelecimento de correlação entre estes interesses. Isto é, serão oferecidos meios aos usuários de uma rede de computadores de uma ou mais organizações – que podem ser externas - de perceberem os interesses de outros usuários.

## Capítulo III: PERCEPÇÃO

Este capítulo apresenta o conceito de percepção, suas características, requisitos e tipologia, bem como definições a respeito do contexto de percepção. Diversas comunidades de pesquisa trabalham neste tema, notadamente, a de *Computer Supported Collaborative Work* (CSCW), de Aprendizado de Máquina (*Information Retrieval* - IR) e de Sistemas Multi-agentes.

A comunidade de aprendizado de máquina utiliza a identificação de comunidades como campo de estudo para medir o desempenho ou, simplesmente, como aplicação de suas técnicas, sendo que as técnicas de IR estão incluídas. Neste caso, a identificação de comunidades não é, na maioria dos casos, vista como um problema de percepção, mas sim como uma tarefa de classificação de indivíduos.

Mitchell (1997) refere-se ao aprendizado de máquina como sendo a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência. De acordo com Mitchell (1997), um programa aprende a partir da experiência **E**, em relação a uma classe de tarefas **T**, com medida de desempenho **P**, se seu desempenho em **T**, medido por **P**, melhora com **E.** Assim, no exemplo de identificação de comunidades, tem-se:

- **Tarefa T**: classificar indivíduos como membros/não-membros de uma comunidade.
- **Medida de Desempenho P**: porcentagem de indivíduos classificados corretamente.
- **Experiência de Treinamento E**: uma base de dados histórica em que os indivíduos conhecidos são previamente classificados como membros ou não-membros da comunidade.

A comunidade de CSCW utiliza técnicas de aprendizado de máquina como ferramenta para implementar mecanismos de percepção, tratando o problema de percepção de forma mais abrangente que a comunidade de aprendizado de máquina.

Mecanismos de percepção são de natureza distribuída, e é nesse aspecto que ocorre um envolvimento mais forte da comunidade de sistemas multi-agentes, aplicando métodos de coordenação e comunicação necessários a um sistema como esse.

A comunidade de CSCW considera que as informações de percepção são necessárias à coordenação das atividades de uma equipe e ao aumento da usabilidade de sistemas computacionais que implementam ambientes virtuais de trabalho cooperativo.

Este capítulo apresenta o conceito de percepção advindo da comunidade de CSCW e apresenta técnicas e sistemas computacionais que implementam mecanismos de percepção. No final do capítulo, a abordagem proposta é posicionada em relação às apresentadas.

## 3.1 Percepção e Contexto

Quando duas ou mais pessoas interagem para realizar algo em conjunto alguns problemas são observados. A falta de organização das informações utilizadas e do conhecimento construído durante as atividades e a falta de conhecimento do contexto relativo às atividades dos colegas e do grupo podem gerar inconsistências e contradições em se tratando de trabalho em grupo, comprometendo a qualidade dos trabalhos e a eficiência do grupo e, inclusive, impedir que objetivos previamente definidos sejam atingidos. Um contexto consiste em "um ponto importante em sistemas cooperativos, estendendo-se para não somente o conteúdo das contribuições individuais, mas também o seu significado para o grupo como um todo e seu objetivo". (BORGES, CAVALCANTI e CAMPOS, 1995).

Ainda, segundo Dey e Abowd (2000), um contexto pode ser definido como "alguma informação que pode ser usada para caracterizar a situação de uma entidade. Uma entidade é a pessoa, lugar ou objeto que é considerado relevante entre os usuários e as aplicações, incluindo o próprio usuário e a própria aplicação".

Sendo assim, se faz necessária a percepção de informações que representem a situação de cada um dos membros de uma equipe de trabalho e das ações realizadas pelos demais, possibilitando a realização de ações pró-ativas por parte desses membros. Em se tratando de comunidades virtuais, a percepção possui papel fundamental na identificação de pessoas com interesses em comum ou que possuam conhecimentos ou competências relevantes para um determinado indivíduo.

Usuários que trabalham juntos precisam de informações adequadas sobre o seu ambiente cooperativo, tais como presença de outros membros e atividades e

compartilhamento de artefatos (GROSS e PRINZ, 2004). Mecanismos de percepção em ambientes virtuais trazem estas informações, facilitando o trabalho e a interação entre membros de comunidades de maneira cooperativa.

Dourish e Belotti (1992) identificam percepção como “um entendimento das atividades dos outros que provê um contexto para suas próprias atividades”. Cabe ressaltar que essa percepção não está condicionada apenas às atividades realizadas, mas também ao conhecimento e às competências de outros indivíduos que venham a contribuir para a realização de determinada atividade.

De acordo com Steinfield, Jang e Pfaff (1999), a percepção propõe que os membros de uma equipe de trabalho tenham consciência das atividades de seus colegas, o que os encoraja a trabalhar de maneira cooperativa, compartilhando seus conhecimentos e reduzindo o re-trabalho.

De forma geral, é possível afirmar que a percepção consiste em permitir que todos os membros de uma equipe ou comunidade tenham uma visão única de todas as atividades que ocorrem em seu ambiente de trabalho, dos recursos e conhecimentos da equipe e dos elementos que a compõe. Proporciona a visão do que aconteceu, do que acontece e o que poderá acontecer dentro das atividades da equipe. Além disso, possibilita que todos os membros de uma equipe saibam qual o significado de suas ações, para os demais membros, para a equipe e para o trabalho como um todo. Isso pode ser entendido como sendo contexto de trabalho. Sendo assim, torna-se evidente que sistemas computacionais de suporte ao trabalho colaborativo ou à formação/identificação de comunidades considerem contextos de atividades e mecanismos de percepção como elementos necessários.

## 3.2 Mecanismos de Percepção

Neste trabalho adota-se a seguinte definição para mecanismos de percepção: são os elementos de *software* desenvolvidos a partir de tecnologias que permitem capturar, representar e difundir contextos de atividades atuais e passadas dos indivíduos a fim de possibilitar a comunicação, colaboração e a cooperação através da utilização das informações de contexto proporcionadas pelo mecanismo.

Schilit et al. (1994) aponta que para modelar contextos as seguintes questões devem ser respondidas: onde você está, quem você é e quais são os recursos próximos. Dey (2000) determinou os seguintes parâmetros para modelagem de contextos: localização e identificação dos indivíduos envolvidos no contexto, tempo e ambiente (ou atividade). Normalmente, estes modelos funcionam apenas para um único indivíduo e não suportam cooperação e contextos compartilhados.

Diversos mecanismos de suporte à percepção foram analisados, nos quais foram identificados elementos importantes do espaço de trabalho para suportar a percepção. Tais elementos podem ser organizados e apresentados de forma mais abrangente conforme a seguir (GUTWIN e GREENBERG, 2002):

- **O que**: refere-se a quais informações devem ser fornecidas aos usuários;
- **Quando**: refere-se a quando ocorrem os eventos geradores das informações de percepção e quando se dá a apresentação destas informações;
- **Onde**: refere-se a onde as informações são geradas e apresentadas;
- **Como**: refere-se a como as informações são apresentadas aos usuários e como é sua interface;
- **Quem**: refere-se a quem está atuando e quem está atento no momento;
- **Quanto**: qual é a quantidade ideal de informações que deve ser apresentada ao usuário, a fim de prover a ele percepção sobre o seu grupo de trabalho e suas atividades.

Além disso, segundo Gross e Prinz (2004), para que um espaço de trabalho baseado em percepção seja considerado eficaz, são requeridos os seguintes requisitos:

- Um modelo preciso com uma correspondência muito próxima da realidade;
- Um mapeamento claro dos eventos e situações reais partindo do modelo;
- Modelagem simples e fácil adaptação do modelo em caso de mudanças a partir da realidade;
- Nenhum ou pouco esforço adicional do lado dos usuários para capturar e presenciar/perceber a informação.

Embora se trate de um conjunto genérico de requerimentos, eles possuem importância específica para modelos de percepção. Partindo do princípio de que a

percepção consiste em detectar mudanças rápidas dos usuários em suas situações de trabalho, é essencial que o modelo seja realístico com a característica de ser facilmente reconfigurável para retratar mudanças de cenários. (HEATH et al., 2002).

As características apresentadas acima constituem aspectos vitais para o provimento de percepção em ferramentas de *groupware* síncronas e assíncronas. Ambientes síncronos se caracterizam pela necessidade dos membros de uma equipe trabalharem simultaneamente – a interação síncrona descreve a situação em que mais de um usuário acessa concorrentemente a dados compartilhados (MARIANI, 1997) como em videoconferências.

Em um ambiente assíncrono há um intervalo de tempo entre a atuação de um usuário e a percepção da ação realizada ou em curso por seus colegas – interação assíncrona ocorre quando os usuários compartilham um objeto por um longo período de tempo (MARIANI, 1997), não requerendo que os usuários trabalhem simultaneamente para que o objetivo seja atingido. Sendo assim, esses mecanismos diferem também quanto à necessidade de percepção.

Gutwin e Greenberg (1995) apresentam uma definição mais específica da percepção em equipes, argumentando que as reuniões de trabalho em tempo real permitem que usuários dispersos geograficamente colaborem de forma síncrona em um espaço virtual compartilhado, mas que eles necessitam de uma comunicação mais eficaz e percepção das interações face-a-face. Os autores definem a percepção síncrona de grupo como:

> O mais recente conhecimento das atividades de outras pessoas que é requerido por um indivíduo para coordenar e completar sua parte de uma tarefa grupal. A percepção de grupo é mantida pela preservação das informações de rotina tais como: a localização de outros participantes no espaço compartilhado (onde eles estão trabalhando?), suas ações (o que eles estão fazendo?), o histórico das interações (o que eles já fizeram?), e suas intenções (o que eles farão na seqüência?).

A definição permite identificar quatro diferentes tipos de percepção relevantes quando se trata de aplicações de suporte para espaços virtuais compartilhados, como equipes de trabalho ou comunidades virtuais, conforme descrito abaixo (GUTWIN et al., 1996) *apud* (GROSS, STARY e TOTTER, 2005).

- **Informal:** a percepção informal é o conhecimento de quem está ao redor, o que estas pessoas estão fazendo, e o que eles irão fazer. A percepção informal é um pré-requisito para a interação espontânea.
- **Social:** percepção social trata da disponibilidade de diferentes tipos de informações, tais como interesse e atenção ou estado emocional de uma relação social. Isso é freqüentemente percebido de uma maneira não verbal através do retorno de um canal de comunicação secundário e estímulos não verbais como contato visual, expressão facial e linguagem corporal.
- **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, GIBBS e REIN, 1991).
- **Espaço de trabalho:** inclui conhecimentos sobre o espaço de trabalho em geral – informações sobre outras interações dos participantes com o espaço compartilhado e os artefatos nele contidos.

A Figura 3 (GUTWIN et al., 1996) ilustra esses tipos de percepção, os quais se sobrepõem e são interdependentes.

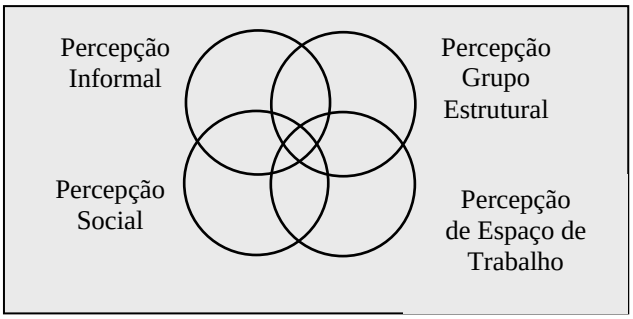



FIGURA 3 - Tipos de percepção em grupos de trabalho.

### 3.3 Contextos de Percepção

Gross e Prinz (2004) definem um contexto dentro de uma configuração cooperativa como as condições inter-relacionadas em que algumas entidades existem (ex. usuário, grupo, artefato) e coisas ocorrem (uma ação realizada por um humano ou máquina).

Ainda de acordo com os autores, um contexto de percepção pode emergir em várias dimensões: contextos geográficos e localizações (construções, recintos, escritórios); contextos organizacionais (departamentos ou projetos); contextos pessoais e sociais (pessoas, família, amigos próximos); contexto tecnológico (usuários de tecnologias específicas, como por exemplo, programadores em linguagem Java); contextos de tarefas ou ações (usuários que realizam ações ou tarefas similares com ferramentas similares) e assim por diante.

Um contexto pode ter os seguintes atributos com valores correspondentes (GROSS e PRINZ, 2004; AGOSTINI, et al. 1996).

- Cada contexto de percepção possui um nome único;
- O administrador de um contexto é a pessoa que configura e gerencia o contexto;
- Membros de um contexto são todos os usuários que trabalham em um contexto e que conseqüentemente produzem eventos através de suas ações;
- Os meios para que eventos possam ser produzidos, são tanto eletrônicos (ex. um espaço de trabalho compartilhado) ou áreas físicas (ex. salas de encontro);
- Os artefatos de um contexto são todos os objetos sobre os quais usuários podem operar;
- Cada contexto é associado com vários indivíduos e aplicações cooperativas (ex. editores de texto, ambientes de programação, aplicações de *groupware*);
- Eventos produzidos em um contexto são descritos por seus tipos;
- Cada membro de um contexto de percepção pode ter o direito de produzir eventos, alterar eventos ou tipos de eventos, e decidir como eles querem que os eventos sejam apresentados, o que pode ser oferecido por uma lista de controle de acesso (LCA). Para cada espaço compartilhado, seus membros podem definir sua própria política de privacidade.
- Cada contexto de percepção possui diversas conexões para seu ambiente e para outros contextos (ex. dois projetos com contextos semelhantes ou membros comuns). Contextos grandes consistindo de muitos membros e

muitos artefatos compartilhados devem ser desdobrados sobre várias dimensões e devem ser organizados em sub-contextos.

A descrição de um contexto não requer a especificação de valores para todos os atributos (GROSS e PRINZ, 2004). Por exemplo, um contexto pode ser criado e alguns atributos como localizações ou aplicações serem especificados mais tarde; ou um contexto pode não ter nenhuma localização ou nenhuma aplicação. Ainda, é possível que os atributos de um contexto podem ser gerados automaticamente. Por exemplo, se um contexto consistir em um espaço de trabalho compartilhado, a lista dos membros e dos artefatos do contexto pode ser dinamicamente obtida da informação sobre o espaço de trabalho compartilhado.

### 3.4 Sistemas Baseados em Percepção para a Formação de Comunidades

As abordagens mais usuais para a identificação de comunidades estão relacionadas às técnicas de aprendizagem de máquina (ex. algoritmos de agrupamento, árvores de decisão), bem como de recuperação de informação (ex. extração de palavras-chaves, análise de *links*). A maioria dos motores de busca realizam tanto análise de *links* como de conteúdo para incrementar a qualidade dos resultados das buscas. De forma individual ou aliada às abordagens citadas, se encontram as técnicas de percepção que visam tornar o usuário ciente do contexto onde atua, ou seja, conhecedor de quem está ao seu redor, que atividades desempenham e que artefatos utilizam entre outras informações relevantes quando se trata de cooperação e colaboração em ambientes compartilhados. 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:

- **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.
- **Análise da localização**: 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.

- **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. Trata-se de um mecanismo de suporte à formação de comunidades, mas fazem parte delas apenas indivíduos que possuírem o sistema instalado, como por exemplo, uma barra de navegação agregada ao *browser*.

A Tabela 1 apresenta alguns dos sistemas mais recentes e outros clássicos propostos por diversos pesquisadores visando a formação de comunidades. Esses sistemas são descritos pela forma de construção das representações de contexto, tipo de percepção utilizado (social, grupo-estrutural, informal, espaço de trabalho), se realizam análise temporal[5] dos contextos e por uma breve descrição da funcionalidade principal.

Tabela 1 – Classificação de ferramentas de percepção para identificação de comunidades.

| ***Análise de Conteúdo*** | | | |
|---|---|---|---|
| **Sistemas** | **Tipo de Percepção** | **Análise temporal dos contextos** | **Funcionalidade principal** |
| **JENA** Dias, Welfer e D' Ornellas, 2004 | Informal e Espaço de trabalho | Não | Desenvolver comunidades virtuais de pesquisa científica a fim de prover o compartilhamento de informações |
| **Cyclades** Choo, Detlor e Tumbull (2000); Stenmark (2001) | Informal e Espaço de trabalho | Não | Oferecer um espaço virtual para comunicação e colaboração |
| **Recomendações Oportunistas** Kamahara et al. (2005) | Informal | Não | Recomendação de programas de televisão que apresenta recomendações oportunistas |
| **I2I** Budzik et al. (2002) | Informal e Espaço de trabalho | Não | Oferecer comunicação oportunista entre os usuários. |
| **CUMBIA** Vivacqua, Moreno e Souza (2005) | Informal e Espaço de trabalho | Não | Detectar oportunidades para colaboração de forma dinâmica. |
| ***Análise da Localização*** | | | |
| **Hyperlink-Induced Topic Search (HITS)** Gibson, Kleinberg e Raghavan (1998) | Informal | Não | Analisar e identificar comunidades com base em topologias de *links* (referência em um documento de hipertexto para outro documento) acessados pelos usuários. |
| **Descoberta de Comunidades Baseadas no** | Informal | Não | Descobrir comunidades de interesse em *WEB Services*, com base no comportamento de acesso dos usuários |

[5] A análise temporal dos contextos permite o estudo da evolução (alterações) das comunidades ao longo do tempo.

| | | | |
|---|---|---|---|
| **Comportamento do Usuário** Almeida e Almeida (2003) | | | |
| **Comunidades através de Grafos Bipartidos** Reddy e Kitsuregawa (2002) | Informal | Não | Extrair e relacionar estruturas de *links* a partir de uma extensa coleção de páginas WEB a fim de formar comunidades. |
| **Estrutura de comunidades da WEB baseadas em motores de busca** Murata (2003) | Informal | Não | Proporcionar a visualização de comunidades na WEB através da formação de grafos, visando facilitar o acesso do usuário a páginas relacionadas a seus interesses. |
| ***Por Adesão*** | | | |
| **Peepholes** Greenberg (1996) | Informal | Não | Oferecer suporte a comunidades distribuídas visando a interação casual. |
| **AROMA** Pedersen e Sokoler (1997) | Social | Não | Explorar imagens para transmitir um sentido de presença remota a fim de construir comunidades nas quais os membros são conhecidos. |
| **Portholes** Dourish e Bly (1992) | Informal e Social | Não | Suportar a percepção compartilhada através de imagens, auxiliando na construção de comunidades. É possível visualizar colegas remotos assim como aqueles que estão presentes localmente. |
| **CyberWindow** Liechti, Mark e Ichikawa (1998) | Social | Não | Capturar atividades ocorridas na WEB e notificar ao usuário sobre os eventos que ocorram no espaço virtual a que ele estiver conectado. |
| ***Community Toolbar*** Prinz, Kolvenbach e Klockner (2002) | Grupo-estrutural e Espaço de trabalho | Não | Oferecer suporte aos membros das comunidades virtuais para que possam criar e compartilhar informações, estruturar e classificar as informações compartilhadas, e estarem cientes do andamento das atividades dentro das comunidades. |
| ***Social WEB Cockpit*** Grather e Prinz (2001) | Informal e Social | Não | Proporcionar navegação na WEB por meio de uma ferramenta de cooperação, para o suporte efetivo (notificação de páginas WEB interessantes, construção colaborativa de conhecimentos da comunidade e o desenvolvimento de um vocabulário comunitário) e auto-organização das comunidades virtuais. |

## 3.5 Discussão

O modelo proposto (descrito no capítulo V) visa a formação de comunidades baseadas na similaridade dos contextos das atividades de usuários distribuídos geograficamente. Para isso, utiliza a análise de conteúdo para representar o contexto dos indivíduos envolvendo a percepção informal e espaço de trabalho, além de realizar a análise temporal dos contextos. A informação contextual provém do conteúdo de artefatos

textuais (ex. arquivos, páginas WEB) que tenham sido utilizados ou produzidos em suas atividades de trabalho, e portanto o modelo classifica-se na categoria de análise de conteúdo. Tais artefatos podem ser denominados itens de interesse dos usuários.

Trata-se de um serviço de percepção distribuído que se utiliza de itens de interesse dos usuários ao invés de sons e imagens providas por eles ou de seus locais de trabalho, conforme apresentado por Dourish e Bly (1992), Liechti, Mark e Ichikawa (1998) e Pedersen e Sokoler (1997).

Diferente da proposta de Budzik et al. (2002) que se baseia no conteúdo de um único documento manipulado pelo usuário em um ambiente distribuído, e estabelece um contexto compartilhado pela visitação do mesmo local na WEB, o modelo proposto pretende considerar a percepção de um subconjunto dos documentos (artefatos textuais) utilizados e produzidos durante as atividades dos usuários. Além disso, o modelo proposto visa a percepção de contextos de atividades passadas e presentes, de forma a proporcionar interação entre os indivíduos tanto de forma síncrona como assíncrona no tempo.

A proposta desse trabalho se assemelha muito ao CUMBIA de Vivacqua, Moreno e Souza (2005), que visa a colaboração entre um grupo de usuários de maneira dinâmica, considerando documentos criados ou acessados recentemente, páginas da WEB e ações realizadas pelos usuários. Todos esses elementos representam o perfil atual de cada usuário que é comunicado aos demais por meio de agentes[6]. O que difere o trabalho proposto em relação ao CUMBIA é que o objetivo deste último é a formação de comunidades sem levar em consideração o contexto de atividades passadas dos usuários.

Como o contexto de cada usuário será constituído pelos itens de interesse dos usuários extraídos a partir de suas atividades de trabalho e não depende exclusivamente de visitação a um local da WEB por exemplo, o modelo proposto não requer uma ferramenta associada ao *browser*, conforme apresentado no *Community Toolbar* por Prinz, Kolvenbach e Klockner (2002).

Budzik et al. (2002) comenta que estudos deixam claro que atualmente muitos locais na WEB possuem conteúdos idênticos ou muito similares, o que levantou questões sobre a eficácia das técnicas de agrupamento baseadas exclusivamente em URL. Além disso, os conteúdos de uma URL podem ser alterados diariamente ou ainda podem se tratar

[6] E um programa de computador que pode operar autonomamente e efetuar tarefas singulares sem a direta supervisão humana (HOFFMAN e NOVAK, 1996).

de páginas dinâmicas. Eis a importância dos métodos de computação de similaridade baseados no conteúdo dos contextos dos usuários. Dessa forma, o trabalho em questão se propõe a realizar uma busca textual em um subconjunto dos documentos utilizados pelo usuário.

A identificação de comunidades é viabilizada graças à capacidade de percepção de contextos similares. Dessa forma, a partir dos contextos de cada um dos usuários é possível realizar o cálculo de similaridade entre eles e em seguida aplicar um algoritmo de agrupamento baseado em similaridades que permita a tais usuários ingressar automaticamente em comunidades onde possam compartilhar seus interesses e colaborar entre si, sem a exigência de esforço por parte deles. Esse agrupamento é flexível, pois permite a um usuário fazer parte de mais de uma comunidade, desde que sua similaridade com os demais usuários se encontre dentro de um limite pré-definido.

Outra distinção observada da abordagem proposta em relação às demais é o interesse em identificar comunidades de forma contínua. Um indivíduo provavelmente tem potencial para participar de várias comunidades. Com seus interesses sendo identificados ao longo do tempo é possível identificar interesses passados e presentes e relacioná-los aos interesses de outros identificando comunidades de maneira contínua. Dessa forma é possível fazer uma análise temporal, o estudo da evolução (ou da modificação) das comunidades ao longo do tempo, o que apresenta contribuições relevantes às organizações.

Em se tratando do tipo de percepção utilizado, pode-se dizer que o modelo proposto proporciona uma mescla de percepção informal e de espaço de trabalho conforme ilustrado na Figura 4. 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.

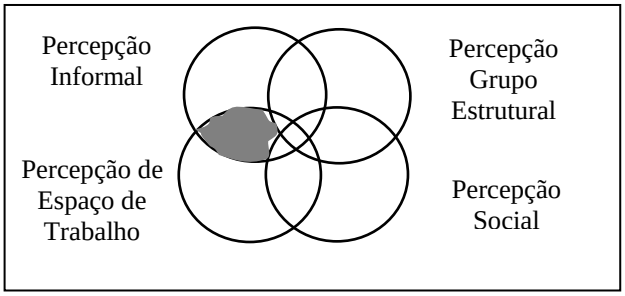


FIGURA 4 - Tipos de percepção aplicados ao modelo proposto.

## CAPÍTULO IV: MODELO PROPOSTO

Neste capítulo, é proposto um modelo de percepção. Este modelo é uma síntese do entendimento adquirido sobre o problema de percepção (capítulo II) e dos sistemas que tratam das percepções informal e espaço de trabalho (capítulo III) aplicado ao problema de identificação de comunidades.

### 4.1 Modelo Conceitual

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 contexto da mesma. A identificação dos membros de uma comunidade ocorre em função do cálculo da similaridade dos contextos das suas atividades passadas e presentes (Figura 5). Em função do grau de similaridade, um indivíduo pode ser identificado como participante de nenhuma ou de diversas comunidades.

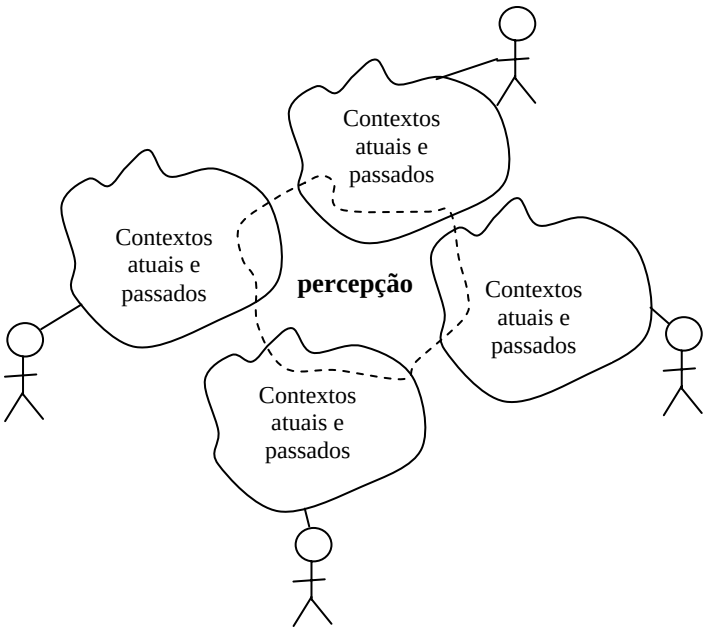



FIGURA 5 - Contexto como conceito fundamental da percepção.

Atividade assume aqui um sentido genérico, tais como atividades profissionais, de lazer, de informação e outras. Uma atividade engloba várias tarefas que se desdobram em um conjunto de ações “informáticas” (ex. abrir um documento, pesquisar na WEB,

conversar com um colega (*e-mail*, *chat* – desde que essa informação seja salva) para atingir um objetivo (ex. obter informação, aprender algo, construir um artefato).

A atividade está sempre associada a um motivo (transformar um objeto). Ela é composta por um conjunto de ações que são elaboradas conscientemente. Cada ação visa atingir uma meta que só pode ser compreendida dentro da estrutura da atividade.

A atividade do trabalho é o elemento central, organizador e estruturante dos componentes da situação do trabalho. Para Montmollin (1984), atividade é o que se faz realmente, enquanto a tarefa indica o que deve ser feito. Assim sendo, a atividade sugere o modo que o sujeito encontra para realizar as ordens.

De acordo com Franco (2001), enquanto a tarefa consiste naquilo que deve ser realizado e que meios estão disponíveis para esta realização, a atividade significa o que realmente é realizado pelo trabalhador com os meios disponíveis (ex. aprender algo, construir um artefato). É o trabalho real, enquanto a tarefa é o trabalho formal.

Kobielus (1997) introduz uma dimensão humana afirmando que se trata de uma "unidade de trabalho executada por um único responsável, que tem condições determinadas de início e fim".

Em se tratando de um processo[7], pode-se dizer que todo processo é composto de várias atividades que possuem papéis e responsabilidades diferenciados. As atividades se subdividem em procedimentos, que devem indicar (MOECKEL, 2000): início e término de uma atividade; o que dá início à atividade; como a atividade deve ser executada; quais ferramentas serão utilizadas para realização da atividade. O autor afirma que os procedimentos são compostos por conjuntos de tarefas, chamados passos, que são a menor parte realizável de uma tarefa.

Dessa forma, atividade assume aqui um sentido genérico, tais como atividades profissionais, de lazer, de informação e outras. Um indivíduo realiza diferentes atividades ao longo do tempo cada uma delas em um contexto específico. A mesma atividade realizada em momentos diferentes pode apresentar contextos diferentes. Por exemplo, a atividade "manter-se informado sobre política" pode ser realizada em contextos diferentes. Em um dia, pode-se acessar as páginas de política do jornal A porque houve uma notícia

[7] Processo é uma série de atividades que consomem recursos e produzem um bem ou serviço (HRONEC, 1994).

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:

- Data, horário, local geográfico e local físico (GROSS e PRINZ, 2004).
- Os artefatos manipulados, tais como: documentos, desenhos, imagens e outros produtos resultantes da utilização de aplicativos (GROSS e PRINZ, 2004).
- As ferramentas utilizadas para manipular os artefatos, tais como, softwares aplicativos.
- As pessoas envolvidas na atividade, seus estados emocionais, a sub-rede social destas pessoas com as relações interpessoais, informações de reputação e confiança.
- Fatos e eventos que influenciam a execução da atividade: durante a atividade de redação de um relatório técnico por exemplo, um *e-mail* do gerente chega adiando a entrega do mesmo, assim a atividade pode temporariamente ser suspensa (GROSS e PRINZ, 2004).
- Seqüência de ações executadas (operações informáticas, tais como: consultar páginas WEB, abrir, criar ou editar um documento texto).

#### 4.1.2 Identificação de comunidades

Uma comunidade passa a ser entendida como a entidade formada pelos indivíduos que possuem contextos similares. Considerando que o contexto de trabalho de cada usuário possui diversos elementos que são constantemente modificados pelas suas ações, nota-se a necessidade de atualizar o cálculo da similaridade (conforme especificado na seção 4.2.2) entre os contextos a intervalos regulares.

Em relação à similaridade de contextos, dois indivíduos ainda que realizem exatamente a mesma atividade, dificilmente terão contextos iguais tal a variedade de ações que podem realizar. Portanto, a percepção de contextos deve buscar contextos similares (não exatamente iguais) e deve ser assíncrona (as atividades não precisam ocorrer ao mesmo tempo).

A freqüência do cálculo de similaridade depende de vários fatores, entre eles: recursos computacionais e largura de banda de comunicação disponíveis, além das exigências da aplicação. Poucas aplicações necessitam de cálculos em tempo real para identificar comunidades, a unidade temporal é normalmente medida em dias.

### 4.1.3 Evolução das Comunidades

A análise temporal das comunidades pode revelar informações importantes sobre o capital intelectual da organização apontando pessoas-chave para o seu desenvolvimento e direções para investir em comunidades que atuam em áreas de conhecimento importantes para a organização, bem como o aproveitamento (compartilhamento) do conhecimento organizacional, muitas vezes subutilizado.

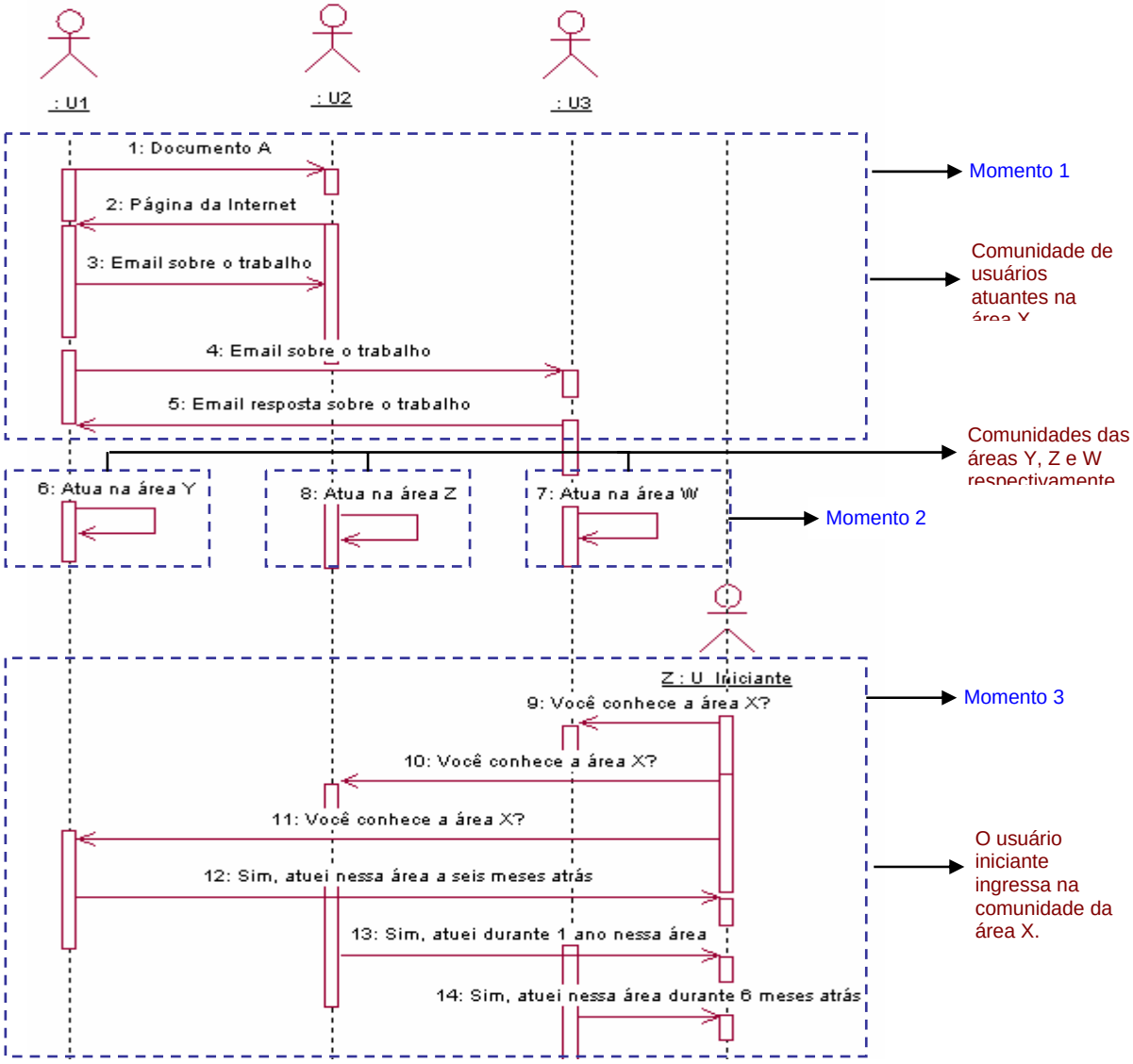



Figura 6 – Exemplo de Diagrama temporal de contextos de atividade.

Considere a situação onde um novo colaborador de uma organização inicia seus trabalhos em uma área onde não há outra pessoa trabalhando no momento. As pessoas que já trabalharam nessa área se encontram em outros locais fisicamente distantes, mas continuam detendo o conhecimento da área. Essas pessoas podem auxiliar o colaborador iniciante, mas para isso ele teria que saber quem se interessa ou se interessou pela sua área. Nesse momento é que são utilizados os contextos de atividades passadas dos usuários, visando o compartilhamento do conhecimento que adquiriram em sua função anterior. A Figura 6 ilustra esse exemplo.

### 4.1.4 Diagrama funcional

De acordo com Paliouras (2002), o trabalho de construção de comunidades virtuais na Internet 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 e que são propostos nesse modelo conceitual. Tais estágios são apresentados no diagrama em blocos da Figura 7, os quais podem ser divididos em dois processos principais: captura do contexto de atividades e identificação de comunidades.

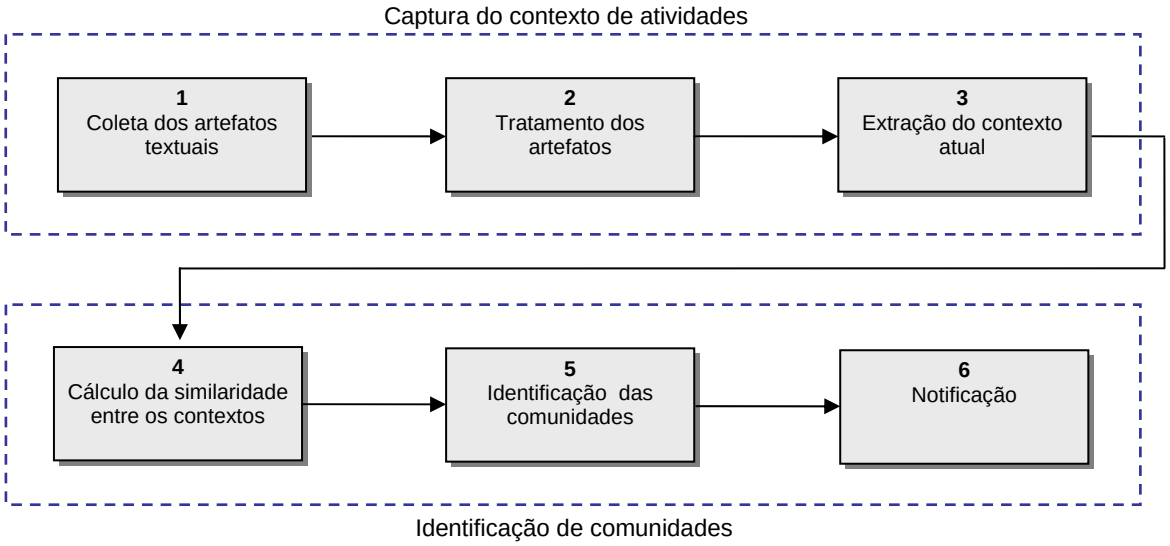



Figura 7 – Estágios necessários à identificação de comunidades segundo Paliouras (2002).

1- Coleta dos artefatos textuais: percepção das atividades dos usuários a fim de extrair uma coleção de artefatos utilizados por eles em suas atividades diárias, o que pode ser realizado de forma manual ou automatizada.

2- Tratamento: seleção dos artefatos representativos da atividade em execução e conversão dos artefatos para o formato adequado para a realização da análise de seu conteúdo pelo algoritmo de extração do contexto.

3- 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 respectivo.

4- Cálculo da similaridade entre os contextos de atividade dos usuários: realizado com base no contexto de atividades gerado no estágio 3. Considerando os sistemas apresentados no capítulo anterior, 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*).

5- Identificação das comunidades: obtida confrontando os valores de similaridade obtidos entre os usuários a fim de encontrar intersecção entre os valores de similaridade com base em um limite de similaridade pré-estabelecido.

6- Notificação: aviso aos usuários sobre as alterações ocorridas nas comunidades (ingresso de novos membros, saída de um membro da comunidade, permissão para inserir um membro em uma comunidade).

#### 4.1.4.1 Processo de Identificação de Comunidades e Notificação

A partir do contexto de atividade de cada usuário, torna-se necessário computar as similaridades entre os contextos pertencentes ao sistema e manter seus usuários cientes de quem possui contextos similares. Dessa forma, o processo de notificação deve consistir de: (1) computação para identificar os usuários cujos contextos de atividade possuem maior similaridade; (2) identificação de comunidades; (3) um protocolo para transferir (notificar) a similaridade entre os usuários.

O momento e a freqüência de notificação influenciam significativamente no desempenho do algoritmo, portanto a notificação não deve ocorrer muito freqüentemente,

pois isso poderia comprometer o desempenho do sistema. Sendo assim, o processo de notificação proposto deve ocorrer da seguinte forma:

1. Para cada um dos usuários pertencentes ao sistema, é coletada uma coleção de artefatos textuais utilizados em suas atividades. O processo de notificação está condicionado a operação de coleta, a qual é realizada a cada período de tempo pré-estabelecido pelo sistema (ex. semanalmente, mensalmente), pois os usuários podem ter seus interesses alterados no decorrer desse período. Conforme apresentado no modelo proposto, com base na coleção de artefatos textuais, é gerado o contexto de atividades para cada usuário. Portanto, a cada período de tempo pré-estabelecido é preciso verificar se houve alteração no contexto de atividades dos usuários - o que é feito através do cálculo de similaridade – para então realizar o processo de notificação.
2. Cada usuário ingressante no sistema, após a geração de seu contexto de atividades, notificará os demais usuários enviando uma mensagem contendo o seu contexto, perguntando quem possui um contexto similar ao seu e caso seja similar, se o usuário permite seu ingresso na comunidade correspondente ao contexto.
3. Cada um dos demais usuários, de posse de seu contexto e de uma lista de comunidades às quais pertence, realizará o cálculo de similaridade entre o contexto enviado pelo novo usuário e o seu contexto. Caso um usuário do sistema seja similar ao novo usuário (ambos estejam dentro do limite de similaridade pré-estabelecido pelo sistema) notificará o novo usuário respondendo afirmativamente a ele sobre a similaridade de seus contextos e se permite ou não seu ingresso na comunidade respectiva. Além disso, quais outros usuários também são similares a ele, ou seja, a qual comunidade deve pertencer, conforme o exemplo da Figura 8. Nesse exemplo, os usuários U1 e U2 formam uma comunidade relativa ao contexto de atividades sobre Informática – denominada C1 - e o usuário U3 ingressa no sistema contendo o mesmo contexto de atividades. Então ele envia uma mensagem para U1 e U2 informando o seu contexto (CAU3) e perguntando se são similares a ele, em seguida ambos os usuários enviam uma resposta afirmativa juntamente com os demais usuários pertencentes à comunidade relativa ao contexto, além de

incluir o novo membro na comunidade correspondente. Dessa forma, U3 ingressa na comunidade e mantém armazenados os membros da comunidade à qual pertence. Esse procedimento é benéfico, pois se um dos usuários deixar de fazer parte da comunidade por motivo de falha do sistema, os demais usuários possuem a informação necessária para que a comunidade continue existindo.

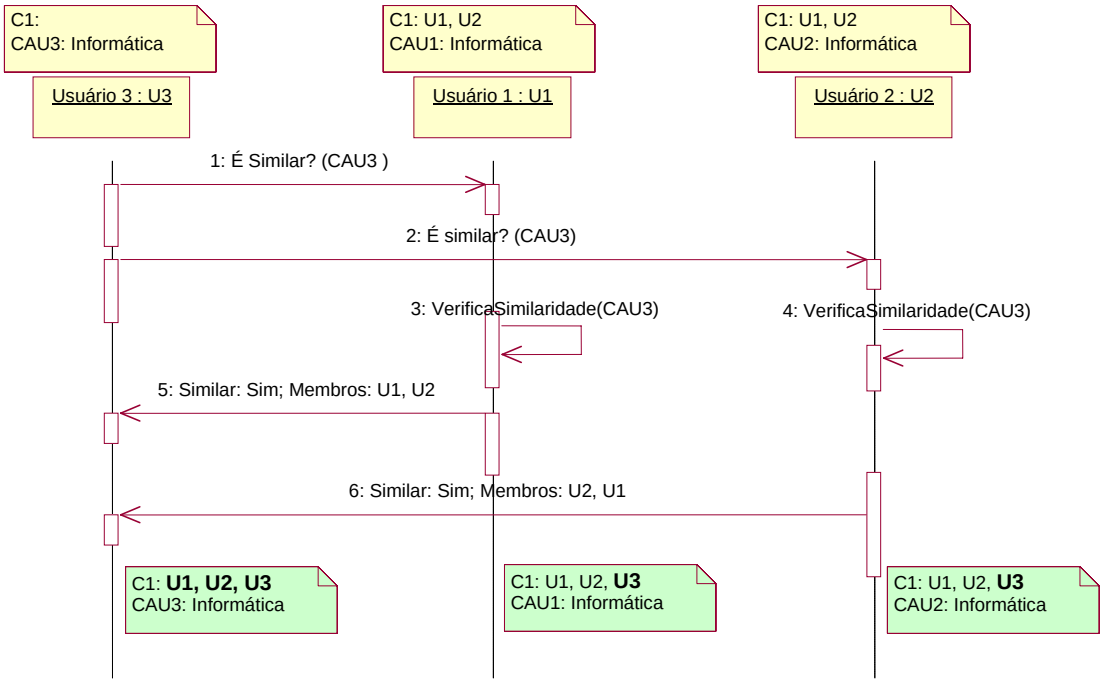



FIGURA 8 – Exemplo do processo de notificação para um novo usuário do sistema.

4. Quando um usuário tem seu contexto de atividades alterado (o que é verificado após o cálculo de similaridade, realizado periodicamente), recebe um período de expiração, e passa fazer parte de outra comunidade, repetindo a execução do Passo 3. Os usuários pertencentes à comunidade relacionada ao contexto antigo são notificados do fato. Até que não ocorra o vencimento do período de expiração, o usuário é mantido na comunidade relativa ao seu contexto passado - além de fazer parte da comunidade correspondente ao contexto atual - e continua a fazer parte do cálculo de similaridade cada vez que um novo usuário, relacionado ao contexto antigo, ingressa no sistema. Nesse caso, um usuário pode fazer parte de duas ou mais comunidades ao mesmo tempo.

Conforme descrito acima, as comunidades seriam identificadas de acordo com a área de interesse identificada em seu contexto. Um aspecto importante a ser considerado

nesse caso é como determinar a área representada por um contexto de atividades, a qual irá denominar uma comunidade. No modelo proposto não há uma representação do tema da comunidade. Uma comunidade é representada de maneira distribuída onde cada participante sabe que pertence a uma comunidade formada pelos indivíduos x1, x2, ..., xn. O problema é como atribuir um nome às comunidades. Suponha o seguinte: U1 forma comunidade com U2 e U3 porque possuem atividades similares em relação ao tema t1. O usuário U1 muda de atividade, atuando em algo relativo ao tema t2. O mesmo ocorre com U2 e U3. Então, U1, U2 e U3 formam uma nova comunidade, mas agora em relação ao tema t2. Nesse caso, a questão é: Como distinguir estas duas comunidades? Uma possível solução é a utilização de um espaço de nomes que garanta que duas comunidades com mesmos participantes, mas referentes a diferentes temas não terão o mesmo nome. Outra abordagem seria constituir um nome para a comunidade (ex. U1 15112006GMT 12h00’05”025 , tal que <id iniciador> <data> <hora>).

Nota-se que a complexidade computacional de um processo conforme descrito acima seria linear em relação ao número de mensagens. O aumento destas é linear em função do número de usuários (*n*):

$$\textit{Número de mensagens}(n) = 2 * n - 2, \text{ para } n > 1.$$

Em relação à execução da função *VerificaSimilaridade*, a que exige mais memória RAM[8] e tempo de processador, o número de execuções depende do número de usuários e de contextos, ou seja, do número de comunidades a que pertencem cada usuário. Assim, se os *n* usuários participam de *c* comunidades em média tem-se *Número de Execuções(c,n) = c * n*. Este número poderia ser reduzido, pois a comparação do novo usuário com as comunidade já existentes se repetem em cada um dos usuários. Por exemplo, se a comunidade c é formada por u2, u3 e u8, a comparação será feita três vezes (por u2, u3 e u8).

### 4.2 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 da WEB. O contexto é composto

[8] *Random Access Memory*

pelos dados básicos (data, hora) e pelos *ats* criados ou acessados pelos usuários em suas atividades diárias.

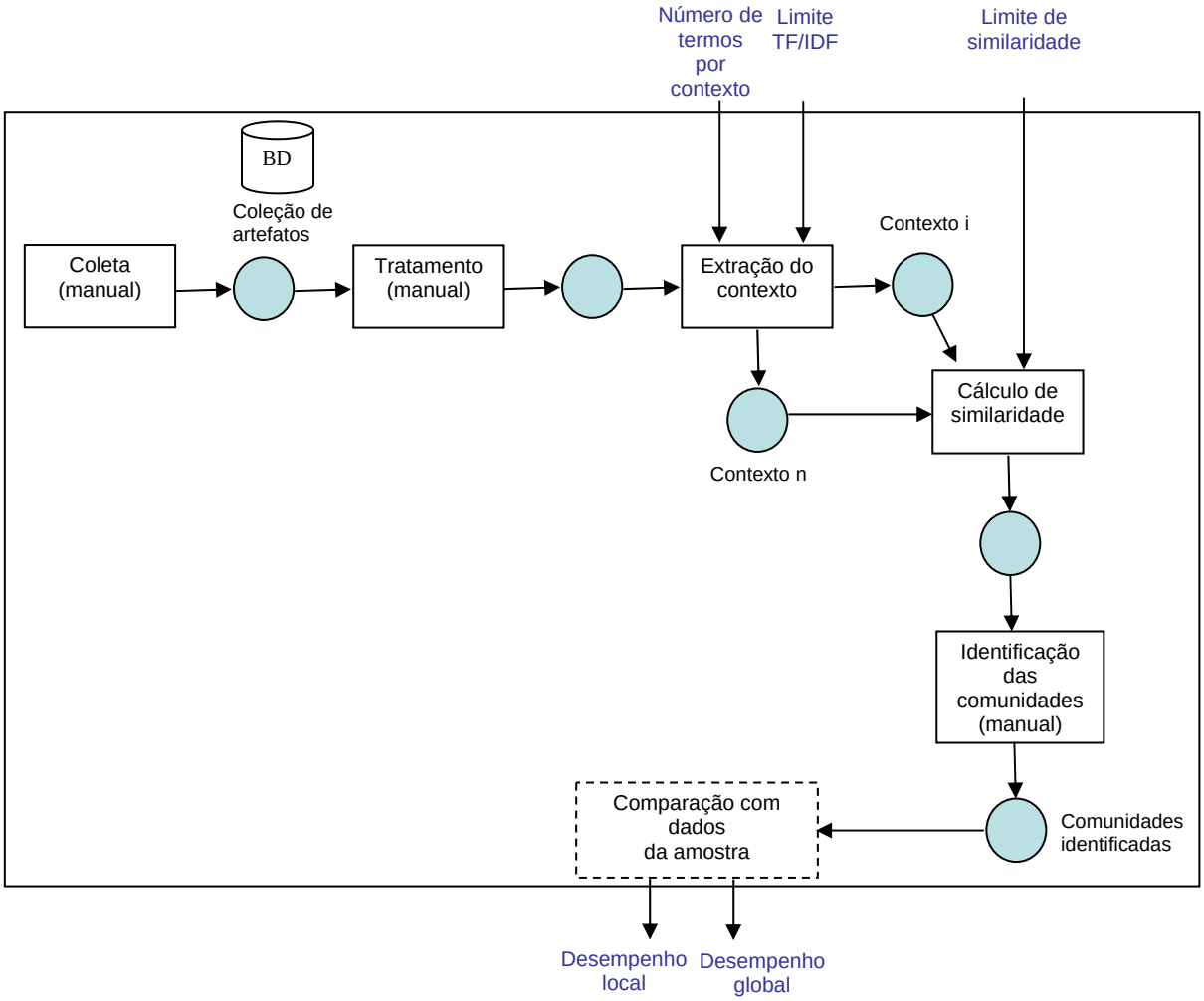



FIGURA 9 – Modelo experimental.

O mecanismo de percepção é constituído por dois processos principais: a construção do contexto de atividade de um usuário – que captura *at*s e extrai termos representativos dos conteúdos – e o processo de identificação de comunidades. A Figura 9 apresenta os elementos do modelo experimental explicados em seguida:

1) Variáveis independentes: número de termos no vetor de contexto e limite de similaridade para considerar dois usuários como participantes de uma mesma comunidade. Há outras variáveis independentes de importância secundária, denominadas variáveis moderadoras: tamanho da coleção de artefatos, limite TF/IDF para incluir um termo no contexto de atividades e vetor de *stop words* utilizado pelo minerador de textos.

2) Variáveis dependentes: traduzem o resultado do modelo, são elas: desempenho local e desempenho global.

3) 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, estabelecer um número de documentos e uma tolerância.

4) 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 em suas atividades de trabalho durante o período de 6 (seis) meses – Janeiro a Junho de 2006. Um *at* típico é um documento eletrônico textual (.txt, .doc, .pdf, .htm, .html) contendo um número ilimitado de páginas na língua portuguesa. Cada usuário classificou os artefatos textuais listados para cada mês de acordo com sua utilização 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. Cada pessoa recebeu uma descrição dos procedimentos a serem seguidos para a realização da coleta de forma adequada, de acordo com o Apêndice 1.

5) 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.

6) 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 para um contexto foi aplicado um limite de 0.1 para os valores de TF/IDF. Cada um dos termos foi armazenado em um vetor – denominado vetor de contexto – o qual foi classificado em ordem decrescente de acordo com os tamanhos de vetores a serem utilizados nos experimentos (200, 500, 1000, 5000). Esse procedimento foi realizado para cada um dos indivíduos pertencentes à amostra.

As seções seguintes detalham os processos de extração do contexto de atividades, cálculo de similaridades – baseados na abordagem de Tacla e Enembreck (2005) -, identificação de comunidades e cálculo de desempenho do algoritmo.

### 4.2.1 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. $TF_i$ designa, por sua vez, a freqüência do i-ésimo termo em um *at* particular. A fórmula TF/IDF é dada pela equação 1.

$$TFIDF(i) = TF_i \times \log\left( {|\ AT\ |} \Big/ {DF_i} \right) \qquad (1)$$

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

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

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.

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

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 vetores de representação dos mesmos. Cada vez que um contexto (o vetor médio dado pela equação

3) sofre alterações, implica alteração dos valores de similaridade, gerando a necessidade de atualização dos vetores de representação de contexto.

Tabela 2 - 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 |

A Tabela 2 ilustra a representação dos contextos de atividades para três usuários ($c_1$, $c_2$ e $c_3$). 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. Estes valores são calculados com base na equação 3.

### 4.2.2 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). Mas para realizar o cálculo da similaridade, primeiramente é preciso normalizar os contextos de atividade para comparar e descobrir que termos melhor discriminam esses contextos. Um termo que é comum, ou seja, é importante para muitos usuários (ex. “projeto”) não seria um bom discriminador.

#### 4.2.2.1 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:

- *{c1, c2, ...., cm}* como sendo um conjunto de contextos de atividade computados de acordo com a equação (3);

- $T_i$ é o vetor derivado a partir da relevância do termo i em todos os contextos - $T_i = \{c_{1i}, c_{2i}, \ldots, c_{mi}\}$;
- $T'_i$ 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 $T_i$).
- o poder de discriminação do termo *i* – denominado *pi* – é dado pela equação 4.

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

Para cada termo *i*, $p_i$ é igual ao somatório dos quadrados dos elementos do vetor $T'_i$. O valor de $p_i$ está sempre no intervalo [1/m, 1]. $p_i$ 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 $p_i$ ocorre quando apenas um contexto de atividade possui o termo *i*.

Tabela 3 - 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 2, o cálculo do *pi* é ilustrado na última coluna da Tabela 3, para cada um dos termos de acordo com a equação 4. Os vetores $T'_i$, necessários para o cálculo de $p_i$, também estão ilustrados na Tabela 3 (os elementos da coluna $p_i$ 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.

#### 4.2.2.2 Similaridade

Para quantificar a similaridade entre dois contextos de atividade *c1* 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 2 e 3, a Tabela 4 mostra estes vetores tomando-se como vetor base o contexto do usuário 1 ($c_1$).

Tabela 4 - 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 5 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 segundo lugar, $c_1$ com $c_2$ e, finalmente, $c_3$, que não tem nenhum termo comum com $c_1$ e apresenta similaridade zero.

Tabela 5 - 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** |

O cálculo de similaridade pode ser visto como parte de um algoritmo de classificação capaz de manipular diversas classes não conhecidas à *priori* (algoritmo não supervisionado). A justificativa para isso é que a aplicação colaborativa pretendida consiste de diversos usuários distribuídos física ou geograficamente, cada usuário representando uma classe (contexto de atividade) no sistema. Os usuários com contextos similares são os exemplares a serem classificados. Em tal aplicação a técnica de classificação deve ser flexível para suportar inserção e remoção de usuários.

### 4.2.3 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 6, que 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 para um determinado período de tempo (6 meses). A tabela é lida linha a linha, assim a linha U1 contém os valores de similaridade de contexto do indivíduo U1 com todos os demais.

A identificação das comunidades é realizada manualmente de acordo com o procedimento:

- 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 6 são apresentados na Tabela 7.

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

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

Tabela 6 - 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 7 – 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, na Tabela 8, são apresentadas as comunidades identificadas quando aplicado um limite de similaridade de 40% aos valores da Tabela 6.

Tabela 8 - Comunidades Identificadas pelo Algoritmo para um Limite de Similaridade de 40%

| **Comunidades identificadas** | |
|---|---|
| **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 |
| **C11** | U7, U11 |

### 4.2.4 Cálculo do desempenho

O desempenho do modelo experimental (seção 4.2) é calculado adaptando-se duas medidas de recuperação da informação: precisão e cobertura (*recall*) (SALTON e McGILL, 1983). Precisão avalia se somente documentos relevantes foram recuperados, ou seja, é dada como a fração entre os documentos relevantes recuperados sobre a quantidade de documentos recuperados (percentagem dos itens recuperados que são relevantes). Cobertura avalia quantos dos documentos relevantes existentes foram recuperados, ou seja, é a fração entre documentos relevantes recuperados e a quantidade de documentos relevantes (porcentagem dos itens relevantes que foi recuperada). Logicamente, é necessário um conhecimento prévio ou estabelecer critérios para determinar os documentos relevantes.

Neste trabalho, não se trata de recuperar documentos e, sim, possíveis comunidades previamente identificadas de acordo com os interesses de cada indivíduo em cada um dos meses envolvidos na coleta dos dados. O desempenho do modelo é calculado segundo duas medidas:

1. desempenho local: número de membros recuperados por comunidade pré-identificada;
2. desempenho global: número de comunidades recuperadas em relação às pré-identificadas.

#### 4.2.4.1 Cálculo de desempenho local

O desempenho local mede o grau de qualidade da composição de uma comunidade identificada, sendo que **comunidade identificada** é qualquer agrupamento de pelo menos dois indivíduos trazido pelo algoritmo. O desempenho local é tanto melhor quanto menos membros extras forem trazidos em relação às comunidades pré-identificadas. São utilizadas duas medidas no cálculo do desempenho local, cobertura e precisão.

Dada uma comunidade pré-identificada *cp*, a cobertura de uma comunidade identificada *c*, no mês *m*, utilizando-se um limite de similaridade *l* é dada pela divisão do número de membros recuperados pelo número de membros de *cp* no mês *m* (equação 7). **Membros recuperados** é o número de indivíduos que estão em *c* **e** *cp*.

$$cobertura(cp,c,m,l) = \frac{membros\_recuperados(cp,c,m,l)}{membros\_comunidade(cp,m)} \qquad (7)$$

Uma comunidade identificada que apresenta cobertura de 100% em relação a uma comunidade pré-identificada é denominada **comunidade recuperada.** Logo, uma comunidade recuperada é uma comunidade pré-identificada com, possivelmente, alguns membros extras.

Para cada comunidade recuperada *c* calcula-se a precisão em relação à comunidade pré-identificada *cp.* A precisão é a divisão dos membros recuperados pela soma dos membros recuperados no mês *m* para o limite de similaridade *l*. **Membros extras** são aqueles que figuram somente em *c.*

$$precisão(cp,c,m,l) = \frac{membros\_recuperados(cp,c,m,l)}{membros\_recuperados(cp,c,m,l)+membros\_extras(cp,c,m,l)} \qquad (8)$$

Considerando o perfil dos usuários envolvidos no exemplo da seção 4.2.3, para um mês qualquer são pré-identificadas as comunidades apresentadas na Tabela 9.

Tabela 9 - Comunidades Pré-identificadas

| **Comunidades Pré-identificadas** | |
|---|---|
| **C1** | U3, U12 |
| **C2** | U1, U3, U8 |
| **C3** | U4, U12 |
| **C4** | U6, U9 |
| **C5** | U2, U6 |

O **desempenho local** do algoritmo é dado pela média da precisão das comunidades recuperadas (equação 9) e **somente para elas**. Esse cálculo é efetuado para um dado mês *m*, um limite de similaridade *l*, um conjunto *C* de comunidades recuperadas tal que $c \in C$ e um conjunto *CP* de comunidades pré-identificadas tal que $cp \in CP$. O fato de considerar uma cobertura de 100% como condição para que uma comunidade seja considerada recuperada faz com que o valor de cobertura não seja utilizado no cálculo de desempenho local.

$$desempLocal(l,m) = \frac{\sum_{cp} \sum_{c} precisao(cp,c,l,m)}{|C|} \qquad (9)$$

A Tabela 10 apresenta os valores das medidas de precisão e cobertura para cada comunidade recuperada. Nota-se que para a comunidade **C1** foi recuperado um membro adicional (U5), o que fez com que o valor de precisão fosse reduzido e, conseqüentemente, o desempenho local para o mês em questão.

Tabela 10 – Cálculo do desempenho local para um mês qualquer e limite de similaridade de 40%

| **Comunidades** | **Pré-Identificadas** | **Recuperadas** | ***Precisão* (%)** | ***Cobertura* (%)** |
|---|---|---|---|---|
| **C1** | U3, U12 | U3, **U5**, U12 | 66,66 | 100,00 |
| **C2** | U1, U3, U8 | U1, U3, U8 | 100,00 | 100,00 |
| **C3** | U4, U12 | U4, U12 | 100,00 | 100,00 |
| **C4** | U6, U9 | U6, U9 | 100,00 | 100,00 |
| **C5** | U2, U6 | U2, U6 | 100,00 | 100,00 |
| | | **Desempenho local** | **93,33** | |
| | | **Desvio Padrão** | **13,34** | |

O **desempenho local total** é calculado pela média aritmética dos desempenhos locais de todos os meses por limite de similaridade. A equação 10 indica a fórmula de cálculo para um limite *l* e um conjunto de meses *M*.

$$desempLocalTotal(l) = \frac{\sum_{m} desempLocal(l,m)}{|M|} \qquad (10)$$

#### 4.2.4.2 Cálculo de desempenho global

O desempenho global mede a taxa de recuperação das comunidades como um todo. Para o cálculo do desempenho global foram consideradas a precisão e a cobertura em cada um dos meses envolvidos na amostra (Janeiro a Junho de 2006). Com base nesses valores é calculada a cobertura e precisão globais (equações 11 e 12) para cada um dos limites de similaridade estabelecidos (*l*) e para cada um dos meses (*m*):

$$precisãoGlobal(l,m) = \frac{comunidades\_recuperadas(l,m)}{comunidades\_recuperadas(l,m)+comunidades\_novas(l,m)} \qquad (11)$$

$$coberturaGlobal(l,m) = \frac{comunidades\_recuperadas(l,m)}{comunidades\_pré(m)} \qquad (12)$$

tal que,

- *comunidades_pré (m)* designa o número de comunidades pré-identificadas no mês *m*;
- *comunidades_recuperadas(l, m):* designa o número de comunidades recuperadas no mês *m* para o limite de similaridade *l*. Uma comunidade pré-identificada *cp* é considerada **recuperada** se existe pelo menos uma comunidade recuperada *c* tal que a *cobertura*(*cp, c, m, l*) é igual a 100%. Por exemplo, dadas uma comunidade pré-identificada *cp* = {U1, U2} e duas comunidades recuperadas c1 = {U1, U2, U3} e c2 = {U1, U2, U8} então c1 e c2 apresentam cobertura igual a 100% em relação à *cp*. Logo, incrementa-se um à variável *comunidades_recuperadas*;
- *comunidades_novas(l, m):* designa o número de comunidades novas no mês *m* para um limite de similaridade *l*. Considera-se **comunidade nova** toda comunidade identificada que possui cobertura menor do que 100% em relação a cada uma das comunidades pré-identificadas. De outra forma, uma comunidade nova:
  - possui pelo menos dois membros comuns em relação a uma comunidade pré-definida, mas com cobertura menor do que 100% **ou**
  - não tem membros de nenhuma das comunidades pré-identificadas, mas apresenta similaridade maior ou igual ao limite *l* entre seus indivíduos.

Ao observar a Tabela 8 tem-se um total de 11 comunidades identificadas pelo algoritmo, das quais 5 são recuperadas (comparando-se com as comunidades pré-identificadas da Tabela 9) e 6, novas. Dessa forma, ao aplicar as equações 11 e 12 a tais números, obtêm-se os valores de precisão e cobertura globais conforme apresentados na Tabela 11.

Tabela 11 – Precisão, cobertura e desempenho globais

| ***Precisão* (%)** | ***Cobertura* (%)** | **Desempenho (%)** |
|---|---|---|
| 45% | 100% | 62% |

O **desempenho global** para cada mês é dado pela medida de desempenho *F-measure*, chamada de *média harmônica* entre a precisão e a cobertura (VAN RIJSERGEN, 1979). A vantagem de usar a média harmônica em relação à média aritmética é que ambas as medidas precisam ser altas para a média harmônica ser alta. O cálculo do desempenho global é realizado através da equação 13. O valor obtido para o exemplo em questão é apresentado na Tabela 11.

$$desempGlobal(l,m) = \frac{2 * precisãoGlobal(l,m) * coberturaGlobal(l,m)}{precisãoGlobal(l,m) * coberturaGlobal(l,m)} \qquad (13)$$

O **desempenho global total** (equação 14) é calculado pela média aritmética dos desempenhos globais dos meses *m* pertencentes ao conjunto *M* dado o limite de similaridade *l*.

$$desempGlobalTotal(l) = \frac{\sum_m desempGlobal(l,m)}{|M|} \qquad (14)$$

# Capítulo V: EXPERIMENTAÇÕES E RESULTADOS

Neste capítulo são apresentados os procedimentos metodológicos utilizados para a realização dos experimentos de avaliação do algoritmo assim como os resultados obtidos seguidos da discussão acerca dos mesmos. O objetivo da experimentação é sugerir valores adequados às variáveis independentes do modelo experimental a fim de identificar comunidades com valores de precisão e cobertura satisfatórios quando comparadas a comunidades previamente conhecidas.

## 5.1 Metodologia

Utilizou-se de uma experimentação controlada que cobre seis meses de atividade de uma amostra da população de discentes e docentes de duas diferentes instituições (UTFPR e FADEP). As comunidades potenciais existentes na amostra foram determinadas com base nas áreas de atuação dos indivíduos selecionados para permitir uma análise comparativa entre os resultados produzidos pelo algoritmo e os conhecidos previamente.

As seções seguintes detalham a metodologia de experimentação, apresentando as características da população e amostra, implementação do protótipo para a realização dos experimentos e análise dos resultados.

### 5.1.1 População e Amostra

Foram coletados dados – de acordo com a seção 4.2 - de 12 (doze) pessoas (U1, U2, U3, U4, U5, U6, U7, U8, U9, U10, U11, U12) pertencentes ao meio acadêmico (docentes e discentes), atuantes em diferentes áreas (conforme a Tabela 12) e pertencentes a duas diferentes instituições de ensino – UTFPR e FADEP.

Foram selecionados indivíduos atuantes preferencialmente na área de computação e engenharia, visando facilitar sua classificação em comunidades (com a ajuda deles). Também, foram incluídos indivíduos de outras áreas (ciências biológicas, ciências sociais aplicadas, ciências agrárias) a fim de medir a capacidade do algoritmo em distinguir os contextos de atividade desses usuários. Essa diversificação das áreas envolvidas na coleta visa validar o algoritmo de forma mais eficaz, pois caso fossem selecionados indivíduos atuantes somente na área de computação, o algoritmo poderia ser considerado preciso ao

retornar uma grande comunidade de computação, mas em uma situação envolvendo indivíduos pertencentes a diferentes áreas, poderia não retornar o resultado adequado.

De acordo com a Tabela de Áreas do Conhecimento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Classificação dos Sistemas de Computação da *Association for Computing Machinery* (ACM), os usuários foram classificados de acordo com seus interesses conforme a taxonomia apresentada na Tabela 12. A classificação do CNPq foi utilizada pelo fato de contemplar todas as áreas do conhecimento, assim foi possível classificar os usuários de áreas diferentes de computação e engenharia. Decidiu-se pela utilização da classificação da ACM para refinar a classificação do CNPq, pois esta última é genérica e poderia levar, por exemplo, a uma única grande comunidade de informática. Destaca-se que a classificação abaixo foi definida de acordo com os interesses específicos informados pelos próprios indivíduos, ou seja, desdobrando cada uma das grandes áreas em suas respectivas subáreas, conforme a área de interesse informada.

Tabela 12 - Árvore Taxonômica das Áreas de Interesse dos Usuários

| |
|---|
| **Ciências Exatas** |
| ▪ **Organização de Sistemas de Computadores** |
| ❍ Geral |
| ⇨ **U11** (professor) |
| ❍ Computadores – Redes de Comunicação |
| • *Geral* |
| ⇨ **U4** (professor) |
| ❍ Desenvolvimento e Arquitetura de Redes |
| ⇨ **U6** (mestrado) |
| ❍ Sistemas Distribuídos |
| ⇨ **U2** (mestrado) |
| ⇨ **U4** (professor) |
| ⇨ **U12** (professor) |
| ▪ **Software** |
| ❍ Engenharia de Software |
| • *Requisitos/Especificações* |
| - Metodologia |
| - Estruturada |
| ⇨ **U3** (professor) |
| - Orientada a Objetos |

⇨ **U7** (professor)

⇨ **U12** (professor)

- *Interoperabilidade*

- Sistemas multi-agentes

⇨ **U4** (mestrado)

- *Trabalho Cooperativo Suportado por Computador (CSCW)*

⇨ **U12** (orientação de mestrado)

- *Linguagem de Programação*

- Java

⇨ **U2** (mestrado)

- **Sistemas de Informação**
  - Armazenamento e Recuperação da Informação
    - *Sistemas e Software*

- Sistemas de Percepção

⇨ **U3** (mestrado)

⇨ **U12** (orientação de mestrado)

  - Aplicações de Sistemas de Informação
    - *Geral*

- Gestão Empresarial

⇨ **U9** (atividade profissional)

    - *Automação de Escritório*

- Sistemas de *Groupware* (comunidades)

⇨ **U3** (mestrado)

⇨ **U12** (orientação de mestrado)

  - Interfaces e Apresentação da Informação
    - *Interface Homem-máquina*

⇨ **U11** (professor)

  - Aplicações de Comunicações

- Sistemas para Ambiente Internet

⇨ **U11** (professor)

- **Metodologias de Computação**
  - Inteligência Artificial

- Métodos e Formalismos de Representação do Conhecimento

- Algoritmos Genéticos

⇨ **U9** (mestrado)

- Representação (procedural e baseada em regras)

- Nebulosa (*fuzzy*)

- ⇨ **U3** (mestrado)
- ⇨ **U6** (mestrado)
- ⇨ **U9** (mestrado)

❍ Processamento de Textos e Documentos

- WEB Semântica
  - ⇨ **U11** (mestrado)

- Ontologias
  - ⇨ **U12** (orientação de mestrado)

■ **Aplicações Computacionais**

❍ Ciências Médicas e da Vida

- Biomédica
  - ⇨ **U2** (mestrado)

■ **Computação de Ambientes**

❍ Computadores e Educação

- Informática na Educação
  - ⇨ **U3** (professor)
  - ⇨ **U11** (mestrado)

❍ Computação Pessoal

- Informática Básica
  - ⇨ **U3** (professor)
  - ⇨ **U4** (professor)
  - ⇨ **U7** (professor)
  - ⇨ **U9** (mestrado e atividade profissional)

**Ciências Biológicas**

■ **Ecologia**
- ⇨ **U10** (mestrado)

■ **Fisiologia**
- ⇨ **U10** (mestrado)

■ **Parasitologia**
- ⇨ **U10** (professor)

**Ciências Sociais Aplicadas**

■ **Administração**

❍ Marketing
- ⇨ **U1** (professor)
- ⇨ **U3** (professor)

- Administração Estratégica
  - ⇨ **U8** (professor)
- Organização, Sistemas e Métodos
  - ⇨ **U8** (professor)
- Geral
  - ⇨ **U6** (atividade profissional)

- **Economia**
  - ⇨ **U1** (professor)

- **Ciência da Informação**
  - Redes Sociais
    - ⇨ **U12** (orientação de mestrado)

**Ciências Agrárias**

- **Agronomia**
  - ⇨ **U5** (acadêmico)
  - ⇨ **U10** (professor e mestrando)
- **Fertilidade do Solo e Adubação**
  - ⇨ **U5** (acadêmico)
- **Parasitologia Agrícola**
  - ⇨ **U10** (professor)
- **Produção e Beneficiamento de Sementes**
  - ⇨ **U5** (acadêmico)
  - ⇨ **U10** (professor)
- **Fisiologia de Plantas Cultivadas**
  - ⇨ **U10** (mestrado)
- **Silvicultura**
  - ⇨ **U5** (acadêmico)
  - ⇨ **U10** (mestrado)
- **Zootecnia**
  - ⇨ **U5** (acadêmico)
  - ⇨ **U10** (professor)
- **Piscicultura**
  - ⇨ **U5** (acadêmico)
  - ⇨ **U10** (professor)
- **Fruticultura**
  - ⇨ **U5** (acadêmico)
  - ⇨ **U10** (professor)
- **Olericultura**

- ⇨ **U5** (professor)
- ⇨ **U10** (acadêmico)

### 5.1.2 Perfil Mensal da Amostra

A seguir são apresentadas as características da amostra em cada um dos seis meses, em função da variação no número de documentos coletados por mês (Tabela 13) e a explicação para que isso tenha ocorrido. De acordo com a identificação das atividades realizadas, foi possível perceber que em alguns meses ocorreram algumas atividades esporádicas, o que influenciou no número de documentos e na formação das comunidades.

Além do volume de documentos, para cada mês são apresentadas as comunidades previamente identificadas, levando em consideração os interesses normais dos indivíduos – conforme a Tabela 14 – e os interesses esporádicos, como por exemplo, a realização de um concurso. Cabe salientar que a classificação apresentada na Tabela 12 foi definida considerando os interesses específicos, ou seja, considerando as subáreas e não as grandes áreas de classificação. A partir do conhecimento prévio das áreas de interesse e das atividades realizadas pelos indivíduos, foi possível fazer uma análise do seu perfil, e com isso identificar comunidades em potencial para cada um dos meses. Dessa forma, a Tabela 14 apresenta as comunidades previamente identificadas por mês.

Tabela 13 - Número de documentos x Usuário – todos os meses

| **Usuários** | **Janeiro** | **Fevereiro** | **Março** | **Abril** | **Maio** | **Junho** |
|---|---|---|---|---|---|---|
| **U1** | 55 | 59 | 134 | 24 | 36 | 64 |
| **U2** | 6 | 8 | 20 | 38 | 125 | 34 |
| **U3** | 40 | 59 | 33 | 143 | 67 | 59 |
| **U4** | 4 | 25 | 14 | 21 | 148 | 29 |
| **U5** | 6 | 6 | 7 | 57 | 10 | 44 |
| **U6** | 13 | 15 | 10 | 7 | 7 | 8 |
| **U7** | 14 | 23 | 53 | 9 | 55 | 25 |
| **U8** | 32 | 7 | 12 | 20 | 18 | 16 |
| **U9** | 9 | 14 | 12 | 5 | 55 | 28 |
| **U10** | 8 | 8 | 25 | 667 | 70 | 93 |
| **U11** | 13 | 100 | 131 | 67 | 39 | 330 |
| **U12** | 3 | 17 | 18 | 23 | 37 | 22 |
| ***Média*** | **16,91667** | **28,41667** | **39,08333** | **90,08333** | **55,58333** | **62,66667** |
| ***Desvio Padrão*** | **15,78743** | **27,91194** | **43,46159** | **177,77348** | **41,38530** | **87,46255** |

O número de documentos obtidos de cada usuário para os meses de Janeiro a Junho de 2006 é apresentado na Tabela 13, na qual percebe-se uma grande variação no volume de documentos de um mês para outro e de um usuário para outro. Enquanto alguns deles não disponibilizaram mais do que 3 (três) documentos, outros apresentaram até 667 documentos em um dos meses.

Tabela 14 - Comunidades Previamente Identificadas – todos os meses

| | **Janeiro** | **Fevereiro** | **Março** | **Abril** | **Maio** | **Junho** |
|---|---|---|---|---|---|---|
| **C1** | U3, U12 *Análise de sistemas* | U1, U3 *Marketing e administração* | U1, U3, U8 *Administração* | U1, U3 *Marketing* | U1, U3 *Marketing* | U1, U3 *Marketing* |
| **C2** | U1, U3, U8 *Administração* | U4, U12 *Sistemas multiagentes* | U3, U12 *Percepção e Comunidades* | U3, U9 *Fuzzy* | U4, U12 *Sistemas distribuídos* | U1, U8 *Ciências sociais – economia* |
| **C3** | U4, U12 *Redes e SD* | U3, U7 *Informática básica* | U3, U7, U9 *Informática básica* | U5, U6, U10 *Agricultura* | U3, U7, U12 *Análise de sistemas* | U7, U11 *Informática na educação* |
| **C4** | U6, U9 *Fuzzy* | U1, U9 *Gestão empresarial, vendas* | U7, U11 *Informática na educação* | U3, U12 *Análise de sistemas* | U4, U6 *Redes de computadores* | U3, U12 *Percepção e identificação de comunidades* |
| **C5** | U2, U6 *Atividades administrativas* | U6, U9 *Redes RBF, redes de sensores* | U8, U10 A*tividades da coordenação* | U1, U10 *Políticas agrícolas e desenv. sustentável* | U5, U10 *Agrárias* | U5, U10 *Agrárias* |
| **C6** | | | | U1, U9 *Gestão de vendas e economia* | U3, U9 *Concurso* | U3, U4, U9, U11, U12 *Concurso* |
| **C7** | | | | U6, U12 *Redes e SD* | U3, U4, U9, U11, U12 *Concurso* | U6, U11 *Redes, segurança* |
| **C8** | | | | | U1, U8 *Atividades da coordenação* | |
| **C9** | | | | | U6, U12 *Redes e SD* | |

Analisando as atividades realizadas pelos usuários em cada um dos meses notou-se que os motivos que influenciam na variação do número de documentos entre os usuários e de um mês para o outro são diversos. Por tratar-se de uma amostra composta pelo público docente e discente, a existência de poucos documentos deve-se ao período de férias, como é possível observar no mês de Janeiro (Tabela 13), quando realizam atividades extras, que muitas vezes não se relacionam com suas atividades reais, ou seja, de estudo ou trabalho. Já a ocorrência de muitos documentos está relacionada à efetivação de atividades freqüentes ou esporádicas. Foram detectadas algumas atividades esporádicas que têm

influência direta no tamanho da coleção de documentos, tais como: reorganização de arquivos em pastas, realização de um concurso, atividades de coordenação, orientação de estágios, pesquisas para a elaboração de trabalhos de disciplinas de mestrado ou para montar o curso de uma determinada disciplina. Sendo assim, observa-se que a partir do mês de Fevereiro ocorre um aumento considerável no número médio de documentos, conforme justificado acima.

### 5.1.3 Experimentos

Para a implementação do modelo proposto foi utilizada a linguagem de programação Java, a qual oferece portabilidade, flexibilidade e uma grande quantidade de recursos de programação, especialmente para mineração de textos. Sendo assim, foi utilizada a API Lucene 1.4.3 da Apache Jakarta, que além de outros recursos, oferece suporte à indexação, remoção de *stop words* e *stemming*. Apesar de Lucene disponibilizar seu próprio *stemming*, foi utilizada uma API adicional denominada *Stemmer_Port*, para realizar o *stemming* na língua portuguesa. O código das principais funções implementadas é apresentado no Apêndice 6.

Após a implementação foram realizados diversos testes com dados reais, coletados conforme apresentado nas seções anteriores, a fim de avaliar o potencial prático do modelo proposto.

Para isso, foram determinadas as variáveis independentes, conforme descrito na seção 4.2, para as quais foram atribuídos diferentes valores visando observar seu efeito na formação das comunidades. Cabe salientar que foi aplicado o limite mínimo de 0,1 aos valores de TF/IDF para que um termo fosse considerado parte do contexto de atividade.

Os valores utilizados para as variáveis independentes foram os seguintes:

- tamanho dos vetores de representação dos contextos de atividade: 200, 500, 1000 e 5000 termos;
- limites de similaridade para considerar que dois ou mais usuários fazem parte de uma comunidade: 20, 30, 40, 50, 60, 70 e 80%;

Para avaliar a formação das comunidades ao longo do tempo, foram realizados experimentos com os valores apresentados acima e de acordo com a seção 4.2.3, para cada um dos seis meses (Janeiro a Junho de 2006), observando o número de comunidades

identificadas pelo algoritmo e a composição das mesmas - conforme definido na Tabela 14.

A formação das comunidades foi observada a partir dos valores de similaridade retornados pelo algoritmo tomando um usuário como referência e comparando-o aos demais, a fim de encontrar valores maiores ou iguais ao limite de similaridade em utilização. Esse procedimento foi repetido para cada um dos usuários.

Para cada um dos experimentos foram geradas as medidas de desempenho local e global, conforme a seção 4.2.4, para identificar valores adequados às variáveis independentes apresentadas acima, de forma a obter um nível de qualidade aceitável para o algoritmo.

## 5.2 Análise dos Resultados

Foram realizados os experimentos apresentados na seção 5.1.3 para os diferentes tamanhos dos vetores de representação dos contextos de atividade dos usuários (200, 500, 1.000, 5.000 termos) quando utilizados diferentes limites de similaridade (20, 30, 40, 50, 60, 70, 80%). Para todos os casos foram gerados os valores de desempenho local e global para cada um dos meses, assim como o desempenho total para ambos os casos. Também foram gerados os gráficos contendo as curvas de precisão, cobertura e desempenho para cada um dos tamanhos em todos os meses. Todas esses dados são apresentados nos Apêndices 2, 3, 4 e 5.

A seguir será apresentada a análise dos resultados tanto para o desempenho local quanto para o global. Por fim serão apresentadas as conclusões a partir das análises realizadas.

### 5.2.1 Análise de desempenho local

A análise de desempenho local visa avaliar a qualidade da constituição interna das comunidades. Conforme definido nas seções anteriores, uma comunidade é considerada recuperada somente quando todos seus membros forem recuperados (cobertura 100%). A análise de desempenho local somente avalia a qualidade das comunidades recuperadas.

Assim, é possível perceber na Tabela 15, que tanto para vetores menores (200 e 500 termos) quanto para vetores maiores (1.000 e 5.000 termos), os melhores valores de desempenho ocorrem aplicando o limite de similaridade de 80%. Nota-se ainda, que para todos os limites de similaridade o desempenho tende a cair quanto maior for o tamanho dos vetores. Esse fato pode ser observado no Gráfico 1, cujos picos de desempenho para cada tamanho de vetor atingem 100% ou ficam muito próximos.

Tabela 15 – Desempenho Local Total – Número de Termos x Limite de Similaridade (%)

| Número de Termos | 20% | 30% | 40% | 50% | 60% | 70% | 80% |
|---|---|---|---|---|---|---|---|
| **200** | 51,02% | 66,82% | 82,24% | 90,51% | 93,75% | 93,75% | **100,00%** |
| **500** | 36,42% | 49,20% | 69,75% | 78,56% | 91,72% | 97,33% | **100,00%** |
| **1000** | 33,46% | 44,19% | 64,98% | 85,42% | 86,94% | 90,97% | **100,00%** |
| **5000** | 32,35% | 44,14% | 59,96% | 78,16% | 87,62% | 94,44% | **98,15%** |

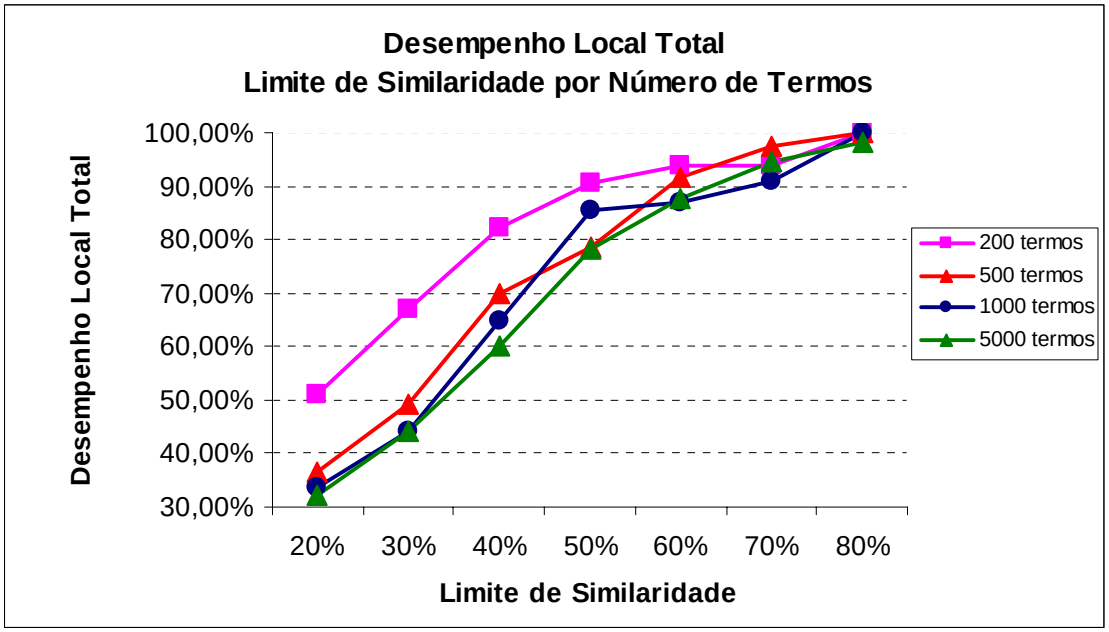



Gráfico 1 – Desempenho Local de todos os meses.

*Influência do limite de similaridade no desempenho local*

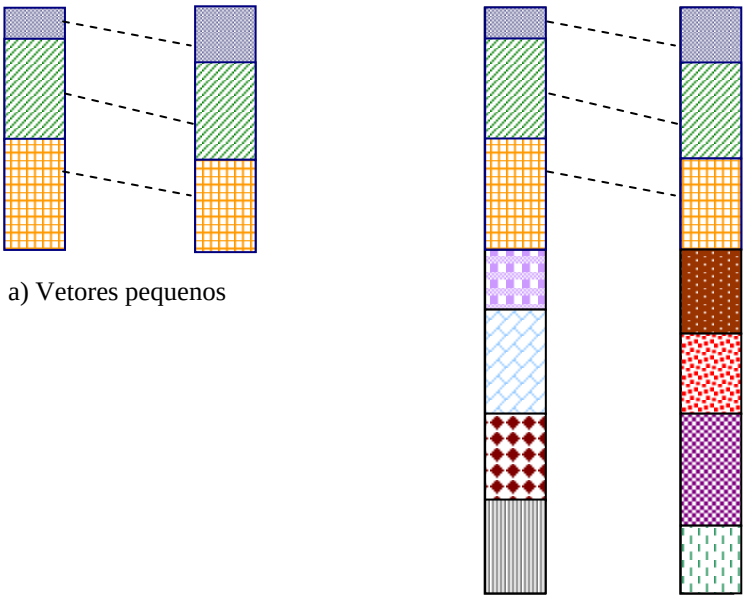
Analisando as curvas do Gráfico 1, pode-se constatar que conforme os limites de similaridade aumentam, ocorre um aumento gradativo do desempenho até que chegue ao seu ponto mais elevado. Esse fato se explica, pois ao se utilizar limites de similaridade menores, são recuperados muitos membros extras, ou seja, são adicionados muitos ruídos, o que gera uma redução dos valores de precisão – e conseqüentemente do desempenho - até que seja atingido o maior valor de precisão, ou seja, o pico de desempenho. Isso ocorre independentemente do tamanho do vetor.

Aumentando-se o limite de similaridade, o número de membros extras diminui, podendo-se constatar um aumento gradativo das curvas de desempenho até atingir os picos.

*Influência do tamanho do vetor no desempenho local*

Quanto maior o tamanho do vetor maior a probabilidade de encontrar usuários similares (independente do nível de similaridade), pois os contextos englobam um número maior de áreas de interesse. Então com vetores maiores, é maior a probabilidade de haver membros extras nas comunidades o que causa uma diminuição da precisão (desempenho). Inversamente, quanto menor o tamanho dos vetores, maiores são os índices de desempenho. Assim, os maiores desempenhos foram obtidos utilizando-se vetores menores. Vetores pequenos cobrem poucas áreas de interesse. Assim, usuários que possuem áreas de interesses bem definidas possuem altos índices de similaridade entre si.

A Figura 10 ilustra esta relação de tamanho de vetor e similaridade. Na parte *a* da figura, mostra-se o caso onde vetores pequenos são utilizados para representar dois contextos de atividades muito similares, visto que os termos mais representativos de cada um são semelhantes (áreas com mesmas hachuras possuem grande similaridade entre seus termos).



a) Vetores pequenos

b) Vetores grandes

FIGURA 10 – Áreas de interesse representadas por vetores de contexto pequenos (a) e grandes (b). Áreas com hachuras iguais em vetores distintos representam termos com elevada similaridade.

À medida que são utilizados vetores de tamanhos maiores, um maior número de áreas são representadas pelos vetores de contexto (Figura 10b), o que pode fazer com que as áreas identificadas como mais representativas em contextos menores passem a ser menos representativas em contextos maiores, ou seja, a similaridade relativa entre os contextos diminui. Isto pode fazer com que a similaridade diminua a ponto de não atingir o limite de similaridade para colocá-los numa mesma comunidade.

Quanto maior a quantidade de indivíduos envolvidos na identificação de uma comunidade, maior a diversidade de áreas de interesse e, conseqüentemente, menores os índices de similaridade à medida que crescem os vetores (dificilmente indivíduos têm interesse em áreas exatamente iguais).

### 5.2.2 Análise de desempenho global

Na análise global, onde é considerado o número de comunidades recuperadas e não mais a constituição interna das comunidades, é possível visualizar no Gráfico 2 que os maiores valores de desempenho ocorrem ao aplicar o limite de similaridade de 20% para vetores de 500, 1.000 e 5.000. Para vetores de 200 termos o maior desempenho ocorreu aplicando-se o limite de 30%. As medidas de desempenho obtidas podem ser observadas na Tabela 16 e visualizadas no Gráfico 2.

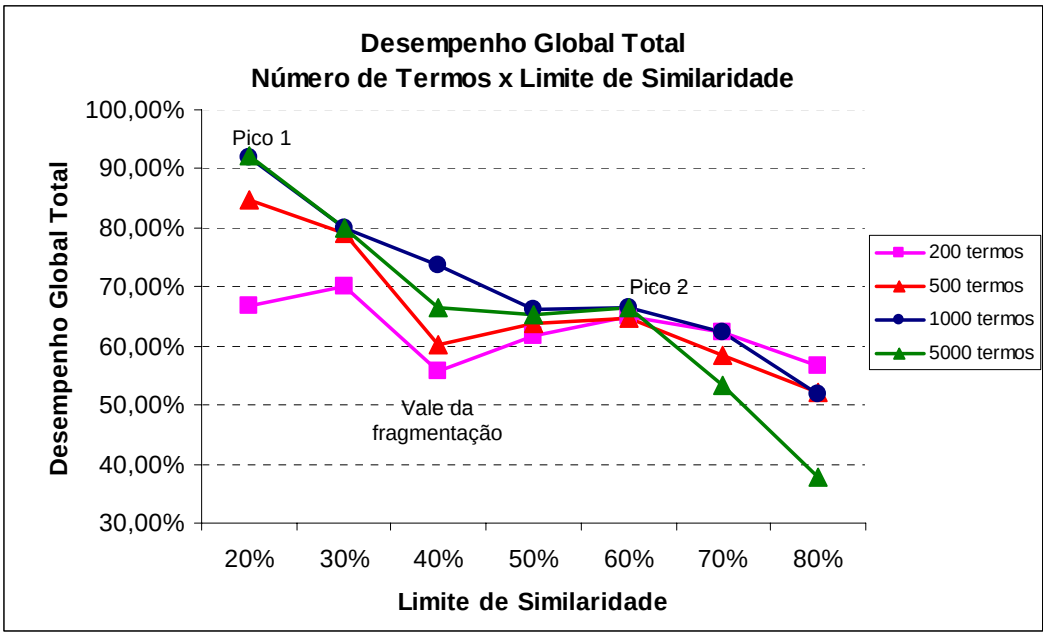



Gráfico 2 – Desempenho Global Total – Número de Termos x Limite de Similaridade.

Tabela 16 – Desempenho Global Total – Número de Termos x Limite de Similaridade (%)

| Número de Termos | 20% | 30% | 40% | 50% | 60% | 70% | 80% |
|---|---|---|---|---|---|---|---|
| **200** | 66,79% | **69,94%** | 55,75% | 61,86% | 65,14% | 62,16% | 56,70% |
| **500** | **84,61%** | 79,11% | 60,21% | 63,76% | 64,56% | 58,50% | 52,26% |
| **1.000** | **92,02%** | 79,94% | 73,65% | 66,28% | 66,55% | 62,32% | 51,96% |
| **5.000** | **92,24%** | 79,89% | 66,35% | 65,24% | 66,50% | 53,37% | 37,80% |

*Influência do limite de similaridade no desempenho global*

Observa-se no Gráfico 2 que, independente do tamanho dos vetores de contexto, com o aumento do limite de similaridade, há uma diminuição no desempenho global. De forma mais detalhada, todas as curvas apresentam a seqüência pico, vale e pico. Para explicar este fato, apresentam-se no Gráfico 3 as curvas de desempenho, precisão e cobertura para 500 termos em função do limite de similaridade. Os vetores de 200, 1.000 e 5.000 (Apêndices 2, 3, 4 e 5) apresentam curvas similares.

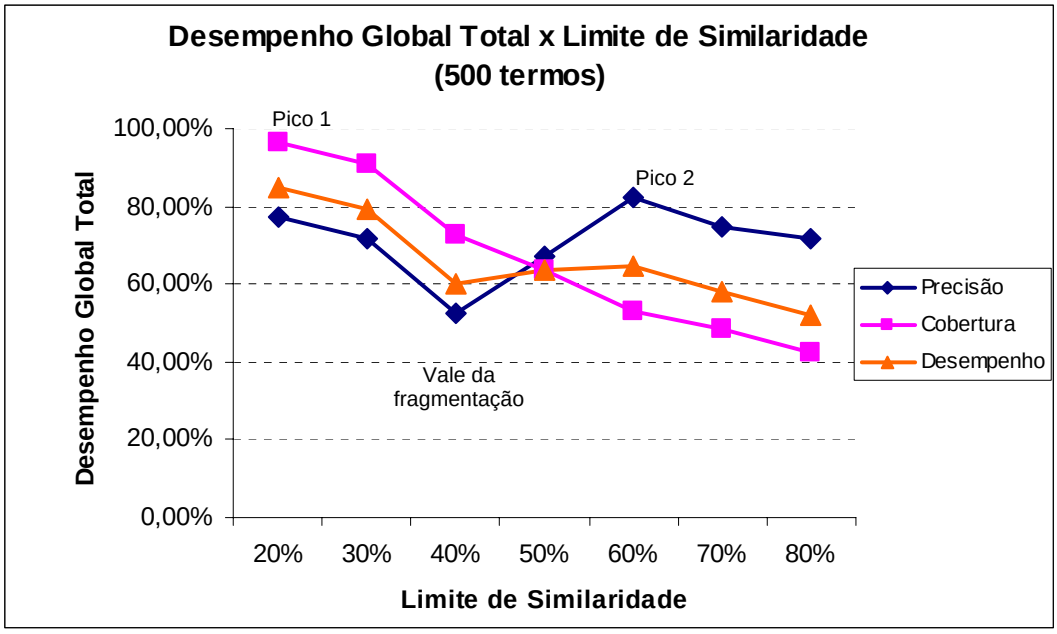



Gráfico 3 - Precisão e cobertura em função do limite de similaridade.

Inicialmente (até o vale do Gráfico 3) os valores de precisão e cobertura são altos e, por conseqüência, o desempenho global. Com limites de similaridade baixos (20 e 30% no Gráfico 4), o número de comunidades recuperadas é elevado e o número de comunidades novas é pequeno. O número pequeno de comunidades novas é devido à tendência de agrupar um elevado número de indivíduos nestas comunidades. Logicamente, com valores de precisão e cobertura altos tem-se um desempenho global alto.

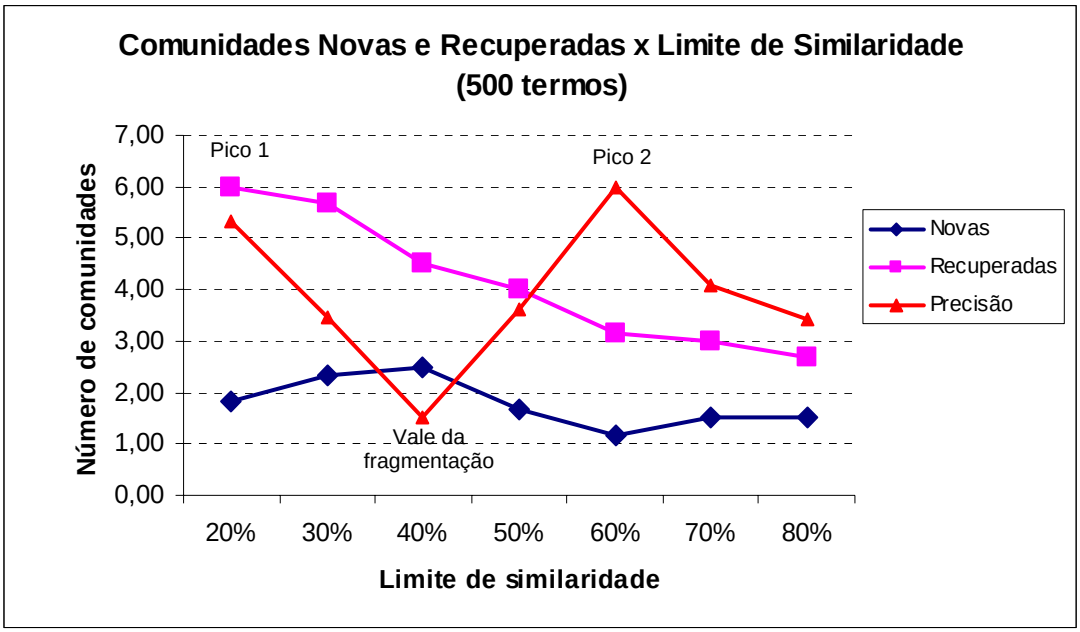



Gráfico 4 – Precisão (normalizada) em função do número de comunidades novas e recuperadas.

No vale (limite de similaridade 40%, Gráfico 3) ocorre um aumento do número de comunidades novas (Gráfico 4), pois elas tendem a se fragmentar à medida que o limite de similaridade aumenta. Em relação às comunidades recuperadas, a tendência é sempre de queda, pois à medida que o limite de similaridade aumenta, a probabilidade dos indivíduos serem similares entre si diminui. Isto ocasiona uma diminuição da precisão e, por conseqüência, do desempenho global.

Na parte ascendente até o segundo pico (Gráfico 4), o número de comunidades novas se reduz, tendendo a zero. As comunidades recuperadas continuam com a tendência de queda, mas como o número de comunidades novas é baixo, mesmo havendo poucas recuperadas, essas fazem com que a precisão seja elevada. Com a precisão elevada, o desempenho global aumenta, daí a existência do segundo pico.

Na parte descendente depois do segundo pico (Gráfico 3), ocorre uma queda gradativa tanto da precisão quanto da cobertura em função dos altos limites de similaridade que diminuem o número de comunidades recuperadas e novas.

*Influência do tamanho do vetor no desempenho global*

No Gráfico 2, observa-se que vetores maiores apresentam melhor desempenho global até o segundo pico (60%). Isso se explica, pois vetores menores apresentam valores

de similaridade maiores do que os apresentados pelos vetores maiores (1.000 e 5.000 termos). Assim, um maior número de comunidades novas é identificado para vetores de 200 e 500 termos (o que pode ser verificado no Gráfico 5), produzindo valores menores de precisão global e, por conseqüência, valores menores de desempenho global.

Após o segundo pico, a precisão apresenta uma queda mais acentuada para os vetores de 1.000 e 5.000 termos em relação aos de 200 e 500 termos, ocasionando uma inversão no desempenho global a partir do limite de similaridade de 70%. Depois do vale da fragmentação (Gráfico 3), o número de comunidades novas se reduz para todos os tamanhos de vetores (Gráfico 5), mas a redução é mais acentuada para vetores menores – 200 e 500 – o que também influencia no fato da precisão ser mais alta para esses vetores. Isso ocorre, pois o número de comunidades novas fica muito próximo de zero. Porém, o número de comunidades recuperadas é maior (ao aplicar limites de similaridade altos – 70 e 80%) para os vetores de 200 e 500 termos em função de apresentarem maiores valores de similaridade quando comparados entre si.

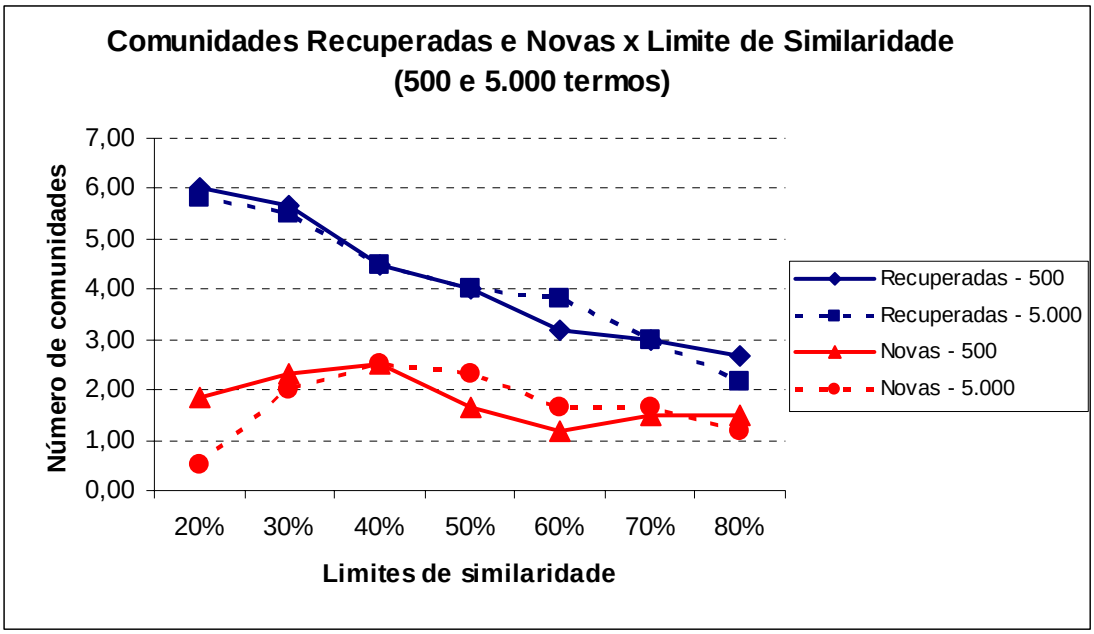



Gráfico 5 – Comunidades Recuperadas e Novas x Limite de Similaridade para vetores de 500 e 5000 termos.

### 5.2.3 Desempenho local x global

As diferentes combinações encontradas entre o desempenho local e o global são mais indicadas para certos tipos de aplicação, tais como, identificar comunidades existentes ou fomentar a criação de comunidades novas. Tomando-se a curva do

desempenho local e global como base, é observada a existência de quatro regiões bem características, tanto para vetores pequenos como grandes, como ilustram os Gráficos 6 e 7 respectivamente.

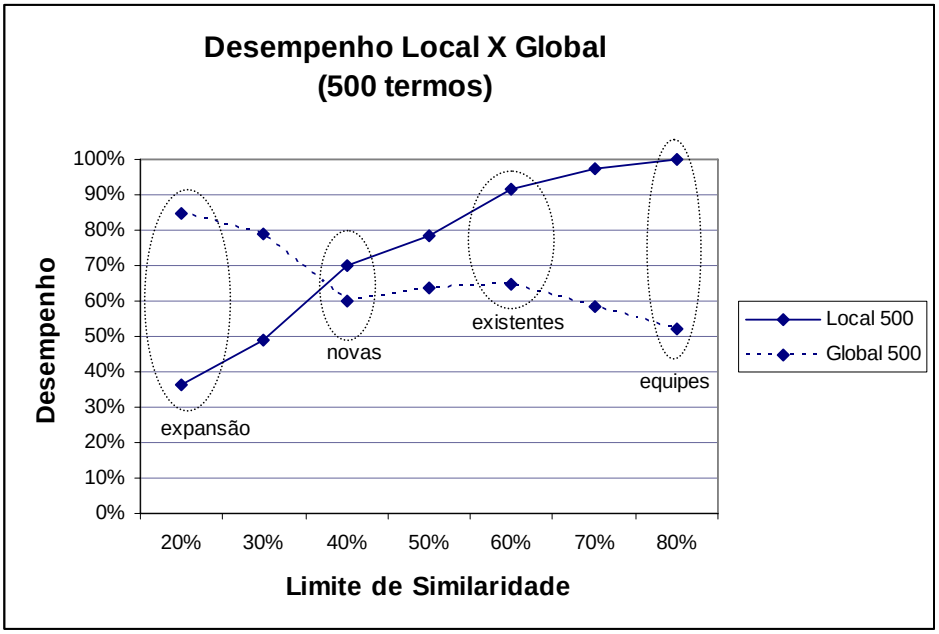



Gráfico 6 - Desempenho local x Global para vetores de 500 termos.

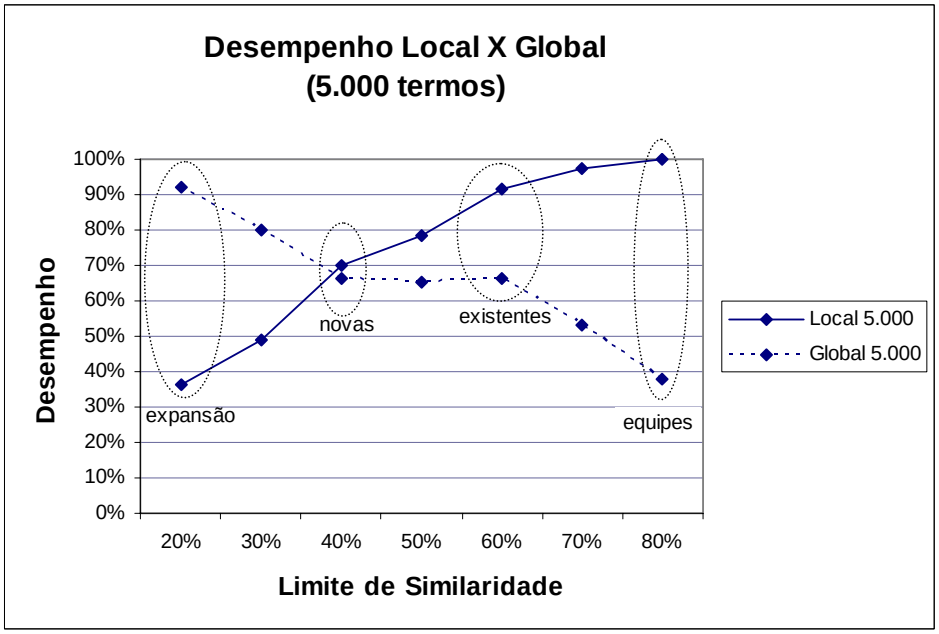



Gráfico 7 - Desempenho local x Global para vetores de 5.000 termos.

Um desempenho local baixo significa uma baixa qualidade na constituição interna das comunidades. Já um desempenho global baixo indica que poucas comunidades pré-identificadas foram recuperadas ou que um elevado número de comunidades novas foram

encontradas. Assim, analisando as regiões *expansão, novas, existentes* e *equipes* apresentadas nos Gráficos 6 e 7, pode-se sugerir as seguintes aplicações em função do limite de similaridade escolhido:

- **Expansão (20% e 30%)**: Significa expansão pelo aumento do número de membros das comunidades existentes. Nesse caso, não importa quais indivíduos venham a fazer parte de cada uma delas. Essa situação é recomendada quando se pretende aumentar as comunidades existentes, a fim de facilitar a criação de um ambiente de aprendizado, havendo o desenvolvimento de novas habilidades profissionais, o que pode ser conseguido, por exemplo, através da transferência de melhores práticas de trabalho e de conhecimentos entre os indivíduos. Neste caso, vetores maiores produzem melhores resultados, pois apresentam menor desempenho local e maior desempenho global (graças ao grande número de comunidades recuperadas), ou seja, identifica as comunidades pré-existentes com membros extras. Por exemplo, a comunidade C3 para o mês de Janeiro (Tabela 14) formada pelos usuários U4 e U12, recuperou além desses, os usuários U3, U5, U6 e U7.
- **Novas (40% e 50%)**: região onde é grande o número de comunidades novas identificadas. Quando se deseja identificar comunidades cujas áreas envolvidas não são conhecidas, ou seja, onde até então não se sabe da existência das áreas de interesse identificadas e nem mesmo quem são os indivíduos que se interessam por elas. Isso contribui para a identificação de competências dos indivíduos de uma organização. Nesta região, recomenda-se a utilização de vetores pequenos (200 e 500 termos), pois produzem maior quantidade de comunidades novas (menor desempenho global) que vetores grandes. Por exemplo, para o mês de Janeiro, foram identificadas 3 (três) novas comunidades além das apresentadas na Tabela 14, são elas: (U5, U10); (U3, U7, U11) e (U3, U7, U9).
- **Existentes (60% e 70%)**: esta região representa a identificação de comunidades existentes com um bom grau de precisão de seus indivíduos. Isso pode ser aplicado quando se quer aumentar a interação entre indivíduos que atuam na mesma área a fim de agilizar o fluxo de informações, bem como

agilizar os processos de trabalho individuais ou coletivos (ex. entre equipes de trabalho ou setores de uma organização). Para identificar comunidades existentes, é preciso um bom desempenho global com boa precisão (desempenho local). Neste caso, os vetores pequenos (200 e 500 termos) apresentam melhores resultados. Por exemplo, as comunidades recuperadas para o mês de Janeiro (C2, C3 e C4 da Tabela 14) continham apenas os membros originais, não foram recuperados membros adicionais.

- **Equipe (80%)**: a região representa a identificação de poucas comunidades com elevada precisão (desempenho local). Muito provavelmente, as comunidades identificadas se aproximarão das equipes existentes na organização. Este é um caso específico do anterior, portanto vetores pequenos (200 e 500 termos) são recomendados. Por exemplo, a comunidade C3, formada pelos usuários U4 e U12 formam uma equipe que trabalha com Sistemas distribuídos e Redes.

A Tabela 17 resume estas conclusões:

Tabela 17 – Tipos de aplicação em função do tamanho dos vetores , limites de similaridade e desempenhos.

| ***Tamanho do vetor*** | ***Limite de similaridade*** | ***Desempenho Local*** | ***Desempenho Global*** | **Tipo** |
|---|---|---|---|---|
| 1.000 ou 5.000 | 20%, 30% | BAIXO | ALTO | Expansão de comunidades existentes. |
| 200 ou 500 | 40%, 50% | MÉDIO | MÉDIO | Descoberta de novas comunidades. |
| | 60%, 70% | ALTO | BAIXO | Identificação de comunidades existentes. |
| | 80% | ALTO | BAIXO | Identificação de equipes. |

## 5.3 Análise da Evolução das Comunidades

Ao analisar a formação de comunidades pelos membros da amostra (Tabela 14) , bem como a identificação das comunidades pelo algoritmo para cada um dos meses, foi possível perceber alterações significativas, tanto nas comunidades formadas, quanto nos membros pertencentes a cada uma das comunidades. Esse fato deve-se à dinamicidade das atividades dos usuários, que em um momento atuam mais intensamente em uma determinada área, deixando-a em segundo plano em outro momento, ou ainda trabalham

em um assunto em função de uma necessidade esporádica (ex. disciplina de mestrado, estudo para um concurso) e não voltam a se interessar pela área no futuro.

Conforme descrito no modelo proposto, o estudo da evolução das comunidades ao longo do tempo possui diversas aplicações nas organizações, tais como:

- **Descoberta de lideranças**: uma comunidade que permanece estável em relação aos seus membros apresenta, muito provavelmente, um membro condutor/líder. O estudo das relações interpessoais em uma comunidade pode ser aprofundado pelo emprego de ferramentas de análise de redes sociais. A identificação da comunidade é apenas o passo inicial.
- **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. Cabe a esta propiciar meios para a comunidade se desenvolver.
- **Mapeamento do capital intelectual:** relacionado ao item anterior, 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. Uma análise mais detalhada pode determinar as áreas que necessitam/merecem investimento e as pessoas-chave para desenvolver competências que serão necessárias à organização.
- **Aproveitamento (compartilhamento) do conhecimento organizacional:** através do armazenamento de contextos de atividades passadas dos usuários, é possível que um usuário que trabalhou em uma determinada área, mesmo deixando de atuar nela, possa contribuir com os demais usuários, pois continuará detendo o conhecimento do assunto.

A fim de comprovar as alterações ocorridas nas comunidades formadas ao longo dos meses, foi tomada como exemplo a evolução das comunidades mês a mês envolvendo os usuários da amostra quando aplicado o limite de similaridade de 40% a vetores de 1.000 termos. Para esse caso, as variações ocorridas na formação das comunidades podem ser constatadas nas Tabelas 18, 19, 20, 21, 22 e 23.

Tabela 18 – Comunidades Pré-identificadas e Novas para Vetores de 1000 termos (40%) - Janeiro

| **Comunidades Pré-identificadas** | **Membros** |
|---|---|
| **C1** | U3, U12 |
| **C2** | U1, U3, U8 |
| **C3** | U4, U12 |
| **C4** | U6, U9 |
| **C5** | U2, U6 |
| **Comunidades Novas** | **Membros** |
| **C1** | U1,U3, U7 |
| **C2** | U3, U11 |
| **C3** | U3, U7, U9 |

Tabela 19 – Comunidades Pré-identificadas e Novas para Vetores de 1000 termos (40%) - Fevereiro

| **Comunidades Pré-identificadas** | **Membros** |
|---|---|
| **C1** | U1, U3 |
| **C2** | U4, U12 |
| **C3** | U3, U7 |
| **C4** | U1, U9 |
| **C5** | U6, U9 |
| **Comunidades Novas** | **Membros** |
| **C1** | U7, U11 |

Tabela 20 – Comunidades Pré-identificadas e Novas para Vetores de 1000 termos (40%) - Março

| **Comunidades Pré-identificadas** | **Membros** |
|---|---|
| **C1** | U1, U3, U8 |
| **C2** | U3, U12 |
| **C3** | U3, U7, U9 |
| **C4** | U7, U11 |
| **C5** | U8, U10 |
| **Comunidades Novas** | **Membros** |
| **C1** | U2, U9 |
| **C2** | U5, U10 |

Pode-se observar que algumas comunidades permanecem inalteradas durante um período de tempo, deixam de existir e retornam em outros períodos. Um exemplo é a comunidade formada pelos usuários U4 e U12, que foi identificada nos meses de Janeiro e Fevereiro, deixou de existir em Março, retornou em Maio e novamente deixou de existir em Junho.

Observando o mesmo caso, nota-se que no mês de Abril, os usuários U4 e U12 retornaram em uma nova comunidade, juntamente com o usuário U8, tal comunidade deixa

de existir em Maio e Junho. No mês de Junho, os tais usuários formam uma terceira comunidade, agora constituída pelos usuários U3, U4, U9, U11 e U12.

Tabela 21 – Comunidades Pré-identificadas e Novas para Vetores de 1.000 termos (40%) - Abril

| **Comunidades Pré-identificadas** | **Membros** |
|---|---|
| **C1** | U1, U3 |
| **C2** | U3, U9 |
| **C3** | U5, U6, U10 |
| **C4** | U3, U12 |
| **C5** | U1, U10 |
| **C6** | U1, U9 |
| **C7** | U6, U12 |
| **Comunidades Novas** | **Membros** |
| **C1** | U2, U4, U7 |
| **C2** | U4, U8, U11 |
| **C3** | U4, U8, U12 |

Tabela 22 – Comunidades Pré-identificadas e Novas para Vetores de 1.000 termos (40%) - Maio

| **Comunidades Pré-identificadas** | **Membros** |
|---|---|
| **C1** | U1, U3 |
| **C2** | U4, U12 |
| **C3** | U3, U7, U12 |
| **C4** | U4, U6 |
| **C5** | U5, U10 |
| **C6** | U3, U9 |
| **C7** | U3, U4, U9, U11, U12 |
| **C8** | U1, U8 |
| **C9** | U6, U12 |
| **Comunidades Novas** | **Membros** |
| **C1** | U2, U7, U9, U11 |
| **C2** | U2, U8, U10 |
| **C3** | U2, U8, U11 |

Tabela 23 – Comunidades Pré-identificadas e Novas para Vetores de 1.000 termos (40%) - Junho

| **Comunidades Pré-identificadas** | **Membros** |
|---|---|
| **C1** | U1, U3 |
| **C2** | U1, U8 |
| **C3** | U7, U11 |
| **C4** | U3, U12 |
| **C5** | U5, U10 |
| **C6** | U3, U4, U9, U11, U12 |
| **C7** | U6, U11 |
| **Comunidades Novas** | **Membros** |
| **C1** | U4, U6 |
| **C2** | U9, U10 |

Com base nas considerações efetuadas acima, pode-se dizer que situações semelhantes ocorrem também com outras comunidades. Isso indica que existe uma grande dinamicidade na formação das comunidades, gerando alterações nas já existentes ou mesmo novas comunidades ao longo do tempo (mês a mês para a amostra em questão). Cabe salientar que o motivo da geração dos agrupamentos pode ser diferente de um período para outro, o que se deve à possível mudança de interesses dos usuários dentro de um período de tempo.

## Capítulo VI: Conclusão

Além do modelo conceitual de percepção proposto, o presente trabalho apresentou uma experimentação do mesmo, enfatizando uma técnica baseada em processamento estatístico de textos para construir e comparar contextos de atividades.

A meta do modelo proposto é identificar comunidades automaticamente ao longo do tempo por meio da análise do contexto de atividades dos usuários do sistema. Trata-se de um serviço de percepção do conteúdo dos artefatos textuais (uma coleção de artefatos) utilizados pelos usuários a cada período de tempo a fim de extrair o seu contexto de atividades, o qual , na experimentação, foi representado por um vetor de termos relevantes. Sendo assim, é calculada a similaridade entre os contextos de atividades dos usuários, que podem estar distribuídos geograficamente, a fim de identificar comunidades.

Conforme apresentado na seção 4.1.4, os estágios que o constituem o processo de identificação de comunidades potenciais são os seguintes: coleta de artefatos textuais, tratamento dos artefatos, extração do contexto atual, cálculo da similaridade entre os contextos, identificação das comunidades e notificação.

Durante o desenvolvimento do trabalho, foi possível perceber algumas dificuldades em relação a alguns dos estágios citados acima. Pode-se dizer que os estágios de coleta e tratamento dos artefatos textuais representam uma grande parcela do trabalho total. Nesse trabalho, a coleta foi realizada manualmente (a partir das ferramentas de pesquisa do Windows, conforme apresentado no Apêndice 1), assim como a conversão dos artefatos para o formato de texto. Durante esse processo algumas situações críticas ocorreram, as quais devem ser tratadas no caso do desenvolvimento de um sistema completamente automatizado:

- Resistência dos usuários em revelar todo o conteúdo utilizado em suas atividades, pois temem dar a entender que não estão trabalhando com seriedade, ou ainda querem manter sigilo sobre o assunto ou atividade em que estão atuando;
- Alguns dos artefatos textuais armazenados pelos usuários podem não ter sido utilizados efetivamente na realização de suas atividades, sendo considerados como ruídos, pois podem não revelar seu(s) real(is) interesse(s);

- Podem surgir atividades esporádicas no decorrer do período que podem ser significativas no momento da análise, mas que na verdade não representam o verdadeiro interesse do indivíduo (ex. estudar para um concurso);
- Ocorrência de alguns fatos que influenciam no tamanho da coleção de artefatos de um usuário, como por exemplo, a organização de arquivos ou a consulta a artefatos (páginas da WEB e outros) para buscar algo que possa ser realmente útil à sua atividade. Isso gera a coleta de artefatos repetidos ou que não estão relacionados às atividades efetivas dos usuários, proporcionando uma grande variação no número de artefatos entre os usuários;
- O processo de seleção e conversão manual dos artefatos para o formato de texto é bastante moroso.

Algumas dessas situações podem ser amenizadas com o oferecimento de um alto nível de flexibilidade ao usuário no momento da coleta, ou seja, se o usuário tiver a liberdade de habilitar ou não a coleta de certos artefatos considerando certos parâmetros como horário ou atividade em que a coleta seria permitida. Isso evitaria o problema da privacidade, sigilo e até a inclusão de artefatos que podem não representar os reais interesses dos usuários. Ainda em relação à inclusão de artefatos textuais que não sejam relevantes o suficiente para serem incluídos na geração do contexto de atividades de um usuário, um mecanismo de coleta precisaria considerar aspectos como o tempo e a freqüência de acesso a um artefato, para então considerá-lo relevante ao contexto de fato.

Apesar do estágio de coleta possuir um papel indispensável para a qualidade dos resultados produzidos pelo algoritmo, o estágio de extração do contexto atual, no qual são extraídos os contextos de atividades dos usuários, pode ser considerado um dos mais complexos e gera algumas questões a serem pensadas quando da sua implementação:

- A necessidade de definir um limiar mínimo para os valores de TF/IDF, pois em função do volume da coleção de artefatos textuais, o número de termos extraídos é bastante grande, resultando em muitos ruídos, ou seja, termos considerados irrelevantes para a representação do contexto de atividades de um usuário (mesmo realizando a extração das *stop words*). O modelo em questão aplicou um limiar mínimo de 0.1, ou seja, somente os termos cujo valor de TF/IDF estava acima desse limiar foram incluídos na análise.

- Ao analisar os termos extraídos para a formação do contexto de atividades, notam-se algumas questões críticas específicas, para as quais torna-se necessário encontrar possíveis soluções:
    - Muitas palavras possuem o mesmo significado;
    - Existem palavras compostas que só tem sentido para um determinado assunto se forem pronunciadas juntas (ex. vantagem competitiva);
    - Existem palavras compostas separadas por um hífen (ex. pré-processamento, semi-automático, ponta-direita), que ao aplicar o *stemming* perdem o significado, pois é retornado apenas o seu radical.
- No *stemming* são definidos os radicais para que os termos possam ser reduzidos à sua provável raiz. É preciso verificar e analisar os radicais definidos no *stemming* para evitar que as algumas palavras não passem a serem tratadas como sendo iguais, mas que na verdade possuem significados diferentes. Por exemplo, o radical "petro" pode significar "petrografia", assim como pode significar "petrolífero". Tais termos podem pertencer a áreas de interesse diferentes.
- Outro fato a ser considerado é que um conjunto de termos diferentes pode representar o mesmo contexto, como por exemplo: laranja, pêra, morango e abacate. Todos esses termos se referem a frutas.

Após a extração do contexto atual foi realizado o cálculo da similaridade entre os contextos utilizando-se o algoritmo implementado conforme apresentado na seção 4.2. Em seguida efetuou-se a identificação de comunidades, o qual se deu de forma manual, gerando portanto uma certa morosidade. Nesse caso é sugerida a automatização desse processo.

A partir da identificação das comunidades partiu-se para a validação do modelo proposto. Conforme especificado no modelo experimental, foram consideradas duas variáveis independentes com vistas à validação do modelo, são elas: tamanho do vetor de contextos e limite de similaridade aplicado para a identificação de comunidades. Essas

duas variáveis possuem influência direta nas variáveis dependentes: precisão, cobertura e desempenho final do algoritmo.

Destaca-se aqui uma contribuição importante desse trabalho quando propõe uma forma adequada para avaliar o desempenho de algoritmos aplicados à identificação de comunidades. Portanto, foram realizadas as análises de desempenho local - a partir dos membros recuperados por comunidade - e global – a partir do número de comunidades recuperadas. Para ambos os casos, foram realizados experimentos com diferentes valores para as variáveis independentes. Para a amostra utilizada, foram identificadas quatro combinações entre o desempenho local e global que são indicadas para:

a) identificar comunidades existentes e aumentá-las com membros extras (limite de similaridade pequeno e vetor pequeno);

b) identificar comunidades novas (limite de similaridade médio e vetor pequeno);

c) identificar comunidades existentes com precisão na definição dos seus membros (limite de similaridade grande e vetor pequeno);

d) identificar equipes (limite de similaridade muito grande e vetor pequeno).

Quanto ao processo de notificação foi apresentada nesse trabalho, a qual gerou as seguintes questões críticas:

- Se as comunidades devem ser identificadas de acordo com a área de interesse representada por seu contexto, é preciso definir uma forma de confrontar o contexto representativo de uma comunidade com o contexto de um usuário, para então alocá-lo em uma determinada comunidade
- Como nomear de forma única uma comunidade.
- Conforme afirmado no decorrer do trabalho, o momento e a freqüência de notificação influenciam significativamente no desempenho do algoritmo. Dessa forma, a notificação não deve ocorrer muito freqüentemente para não comprometer o desempenho do sistema. Essa freqüência deve ser determinada em função do tamanho do sistema (número de usuários), quanto maior o sistema menor deve ser a freqüência de notificação, pois esse procedimento implica na execução dos estágios 3, 4, 5 e 6, apresentados na

seção 4.1.4, podendo gerar um custo elevado em termos de tempo de computação.

Outra contribuição desse trabalho foi propor a análise da evolução das comunidades ao longo do tempo, o que possui inúmeros benefícios, conforme comentados na seção 5.6. Uma implicação desse processo diz respeito a qual período de tempo considerar para essa análise (ex: uma quinzena, um mês, um bimestre). Pode-se dizer que a análise da evolução das comunidades ao longo do tempo está ligada ao processo de notificação, o qual contempla as alterações ocorridas dentro de um determinado período em cada uma das comunidades existentes, ou mesmo a geração de novas comunidades. A partir do processo de notificação é possível gerar e armazenar um histórico dos interesses (passados e presentes) de cada usuário, visando saber em um determinado momento quais são os conhecimentos adquiridos por esse usuário e como ele pode contribuir para com os demais.

A formação das comunidades possui a característica de ser dinâmica, assim como a geração dos conteúdos utilizados pelos usuários – artefatos textuais criados, alterados ou pesquisados e páginas da WEB. Dessa forma, a técnica proposta pode ser considerada flexível – independe do número de artefatos textuais atribuídos a cada usuário, propõe a seleção dos artefatos a serem disponibilizados à análise pelos próprios usuários e o cálculo periódico da similaridade para a atualização das comunidades existentes ou a criação de novas comunidades, bem como a notificação dos usuários. Pode-se dizer que o modelo experimental para a amostra estudada contempla a alocação dos usuários de forma adequada às comunidades existentes ou novas mesmo em meio a esse contexto sujeito a constantes mudanças.

Em relação aos benefícios da aplicação da abordagem proposta nas organizações, tem-se: agilizar o fluxo de informações interna ou externamente às organizações, agilizar os processos de trabalho individuais ou coletivos, através do compartilhamento de interesses, identificar competências inerentes a uma comunidade pertencente a uma organização, facilitar a criação de um ambiente de aprendizado, havendo o desenvolvimento de novas habilidades profissionais - o que pode ser conseguido, por exemplo, através da transferência de melhores práticas de trabalho entre os indivíduos.

Por fim, pode-se dizer que esse trabalho propõe um modelo de percepção – envolvendo a percepção informal e espaço de trabalho - diferenciado para a identificação

de comunidades de indivíduos em ambientes interconectados. Conforme apresentado, esse modelo objetiva melhorar alguns aspectos observados no modelo apresentado em Budzik et al. (2002), bem como em Vivacqua, Moreno e Souza (2005), resultando em uma técnica possível de ser aplicada em organizações reais.

## 6.1 Trabalhos Futuros

Considerando os resultados obtidos e analisando as questões relacionadas na seção anterior, são percebidas diversas situações que requerem um estudo mais aprofundado e específico, o que gera diversas possibilidades de trabalhos futuros, tais como:

1. O desenvolvimento de um sistema para a coleta automática dos artefatos textuais utilizados nas atividades dos usuários  Esse algoritmo deve considerar os seguintes aspectos:

   a) As atividades do usuário não são conhecidas à *priori*, i.e. não há um modelo de atividades. Logo, deve-se encontrar uma forma de identificar um contexto associado a uma atividade durante a realização de uma série de atividades sendo que não há indicação por parte do usuário do início e do término da atividade.

   b) Como determinar quais contextos são suficientemente importantes para que o indivíduo seja considerado como candidato a participar de uma comunidade. Há contextos de atividades esporádicas que devem ser ignorados. Há contextos de atividades freqüentes que também devem ser ignorados porque o indivíduo não deseja divulgá-lo ou participar de uma comunidade que discuta tal assunto.

2. Desenvolvimento do sistema para que funcione de maneira distribuída em uma arquitetura *peer-to-peer* (envolvendo desde o estágio de coleta até a notificação dos usuários).

3. Realização de um estudo das relações inter-pessoais em uma comunidade, o que pode ser aprofundado pelo emprego de ferramentas de análise de redes sociais. A identificação da comunidade é apenas o passo inicial.

4. Aplicação da lógica *fuzzy* a partir de métodos de agrupamento *fuzzy*, sobre os índices de similaridade de cada usuário em relação aos demais, visando extrair um classificador *fuzzy* baseado em algoritmos de agrupamento como por exemplo *Fuzzy C-Means* (FCM). Isso poderia facilitar a alocação de novos usuários no sistema.

5. Definir métodos adequados para a análise da evolução das comunidades ao longo do tempo.

6. Definição e aplicação de um algoritmo que proporcione uma forma gráfica de visualização das comunidades potenciais.

7. Desenvolvimento de uma arquitetura baseada em metadados, que a exemplo do sistema JENA (DIAS, WELFER e D' ORNELLAS, 2004), possa permitir a descrição de informações de forma não ambígua ou redundante, através do uso de um vocabulário de metadados específico para cada comunidade criada no sistema. Através desse vocabulário, pode-se associar um significado ao conteúdo manipulado pelos usuários (representado pelo seu contexto de atividades), formando uma base de dados com descrições semânticas de todas as áreas envolvidas nas comunidades. Tudo isso pode estar aliado à utilização de agentes inteligentes que poderão compreender o significado do conteúdo que está sendo produzido pela comunidade. Dessa forma seria possível determinar a área representada por um determinado contexto de atividades e assim denominar uma comunidade. Isso poderia proporcionar a alocação mais precisa de um usuário em uma comunidade cuja área seja realmente compatível com seus interesses.

## REFERÊNCIAS


AGOSTINI, A.; MICHELIS, G.d.; GRASSO, M.; PRINZ, W. e SYRI, A. **Contexts, Work Processes, and Workspaces. Computer Supported Cooperative Work: The Journal of Collaborative Computing**, vol. 5, no. 2–3, pp. 223–250, 1996.

ALARCÓN, R. e FULLER, D. **Intelligent Awareness in Support of Collaborative Virtual Work Groups**. In: Haake, J. M. and Pino, J. A. (Eds.) CRIWG 2002, LNCS 2440, pp. 147 – 167, Springer-Verlag, 2002.

ALMEIDA, F. B. e ALMEIDA, V., **Design and Evaluation of a User-based Community Discovery Technique**. In *Proceedings of the 4th International Conference on Internet Computing*, pp. 17–23, 2003.

BARROS, L. A. **Suporte a ambientes distribuídos para aprendizagem cooperativa**. Rio de Janeiro: UFRJ, 1994.

BEAMISH, Anne. *Commuties on-line: A Study of Community – Based Computer Networks.* Tese de Mestrado em Planejamento de Cidades. Instituto de Tecnologia de Massachusetts – Estados Unidos. 1995. Disponível em http://albertimit.edu/arch/4.207/anneb/thesis/toc.html

BORGES, M. R. S.; CAVALCANTI, M. C. R. e CAMPOS, M. L. M. **Suporte por Computador ao Trabalho Cooperativo**. In: XV Congresso da Sociedade Brasileira de Computação. XV Jornada de Atualização em Informática. JAI'95. Anais. Canela, RS. 1995.

BOURDIEU, P. **Les Estructures sociales de l'économie**. Paris: Seuil, 2000.

BOURDIEU, P. **Espaço social e espaço simbólico**. In: . Razões práticas - sobre a teoria da ação. Campinas: Papirus, 1996, p. 13-33.

BRINCK, T. e MCDANIEL, S. E. ***Awareness* in Colaborative Systems**, *Workshop Report*, SIGCHI Bulletin, 1999.

BUDZIK, J.; BRADSHAW, S.; FU, X. e HAMMOND, K. J., **Clustering for Opportunistic Communication**. *WWW 2002*, Honolulu, Hawaii, USA. ACM, 2002.

CASTELLS, M. **A Sociedade em rede**. São Paulo: Paz e Terra, 1999. 510p.

CHOO, C.W., DETLOR, B. e TUMBULL, D. **WEB Work: Information Seeking and Knowledge Work on the World Wide WEB**. Kluwer Academic Publishers, Dortrecht, NL, 2000.

COLEMAN, J. **A Rational Choice Perspective in Economic Sociology *in*** – SMELSER, Neil e SWEDBERG, Richard (eds). - **The Handbook of Economic Sociology** - Princeton University Press e Russel Sage Foundation – Princeton, New York, 1994.

COLEMAN, J., **Social Capital in the Creation of Human Capital**. American Journal of Socio*log*y, 94 supplement, pp. 95-120, 1988.

DAVENPORT, Thomas, PRUSAK, Laurence, *Conhecimento empresarial*, Rio de Janeiro, Campus, São Paulo, Publifolha,1999.

DEPOVER, C. e MARCHAND, L., **E-learning et formation des adultes en contexte professionnel**. Bruxelles, De Boeck Université, 2002.

DEY, A.K. e ABOWD, G.D. **Towards a Better Understanding of Context and**

**Context-Awareness**. Proceedings of the CH12000 Workshop on The What, Who, Where, When, Why and How of Context-Awareness, April, 2000.

DEY, A.K. **Providing Architectural Support for Building Context-Aware Applications**. Ph.D. thesis, Georgia Institute of Techno*log*y, December, 2000.

DIAS, A. P.; WELFER, D. e D' ORNELLAS, M. C. **JENA: uma ferramenta para desenvolver comunidades virtuais de pesquisa científica**. Revista do CCEI, v.8, n. 14, Bagé – RS, 2004.

DOURISH, P. e BLY, S. **Portholes: supporting awareness in a distributed work group**. In: Proceedings of CHI'92. ACM Press, New York, pp. 541–547, 1992.

DOURISH, P. e BELLOTI, V. ***Awareness* and coordination in shared workspaces**, *Proceedings of CSCW'92*, Chapel Hill NC, pp. 107-114, 1992.

EFIOS. **Communities of Practice**. Amsterdan, Holanda, 2004. Disponível em <http://www.efios.com/pdf/efios_communities.pdf>. Acessado em 24/09/2005.

ELLIS, C. A.; GIBBS, S. J. e REIN, G. L. **Groupware: somes issues and experiences**. *Communications of the ACM*, New York, v. 34, n. 1, p. 38-58, jan. 1991.

FREIRE, P. **Pedagogia da Autonomia. Saberes necessários à prática educativa**. Rio de Janeiro: Paz e Terra, 1999.

FRANCO, E. M.. **Gestão do conhecimento na construção civil**: uma aplicação dos mapas cognitivos na concepção ergonômica da tarefa de gerenciamento dos canteiros de obras. Florianópolis: UFSC, 2001.

FUKS, H e ASSIS, R. L. **Facilitating Perception on Virtual Learningware-Based Environments**, *Journal on Systems and Information Technology*, 2001.

FUKS, H.; GEROS, M.A.; PIMENTEL, M.G. **Projeto de Comunicação em Groupware: Desenvolvimento, Interface e Utilização.** XXII Jornada de Atualização em Informática. In: Congresso da Sociedade Brasileira de Computação, 23, 2003. Anais. Vol.2, cap.7.

FUKWYAMA, F. **A grande ruptura: A natureza humana e a reconstituição da ordem social**. Rio de Janeiro: Rocco, 2000.

GIBSON, D. KLEINBERG, J. e RAGHAVAN, P. **Inferring WEB communities from link topo*log*y**, in Proc. ACM Conference on hypertext and hyper-media, 1998, pp.225-234.

GRATHER, W. and PRINZ, W., **The Social WEB Cockpit: Support for Virtual Communities**, *GROUP'01,* Sept. 30-Oct. 3, 2001, Boulder, Colorado, USA.

GREENBERG, S. **Peepholes: Low Cost Awareness of One's Community**. Comp. Proc. CHI '96, 1996, 206-207.

GROSS, T., **CYCLADES: A Distributed System for Virtual Community Support Based on Open Archives**, Eleventh Euromicro Conference on Parallel, Distributed, and Network-Based Processing - PDP 2003 (Feb. 5-7, Genova, Italy). Clematis, A., ed. IEEE Computer Society Press, Los Alamitos, CA, 2003. pp. 484-491.

GROSS, T. e PRINZ W. **Modelling Shared Contexts in Cooperative Environments: Concept, Implementation, and Evaluation**. Computer Supported Cooperative Work 13: 283–303, 2004.

GROSS, T., STARY, C. e TOTTER, A. **User-Centered Awareness in Computer-Supported Cooperative Work-Systems: Structured Embedding of Findings from Social Sciences**. International Journal of Human-Computer Interaction, pp. 323-360, 2005.

GUTWIN, C. e GREENBERG, S. **A Descriptive Framework of Workspace Awareness for Real-Time Groupware**. *Computer Supported Cooperative Work: The Journal of Collaborative Computing 11*, 3-4 (2002). pp. 411-446.

GUTWIN, C. e GREENBERG, S. **Support for Group Awareness in Real-Time Desktop Conferences**. In *Proceedings of the Second New Zealand Computer Science Research Students' Conference* (Apr. 18-21, Waikato, Hamilton, NZ). 1995.

GUTWIN, C.; GREENBERG, S. e ROSEMAN, M. **Workspace Awareness in Real-Time Distributed Groupware: Framework, Widgets, and Evaluation**. In *Proceedings of the Conference on Human-Computer Interaction: People and Computers - HCI'96* (Aug. 20-23, London, UK). Springer-Verlag, Heidelberg, 1996. pp. 281-298.

HAETINGER, D. **Fatores Relevantes à Formação e Manutenção de Comunidades Virtuais Facilitadores da Aprendizagem**. CINTED-UFRGS, Novas Tecnologias na Educação, V.3, Nº 1, Maio, 2005.

HEATH, C.; SVENSSON, M.S.; HINDMARSH, J.; LU, P. e LEHM, D.Vom. **Configuring *Awareness*. Computer Supported Cooperative Work: The Journal of Collaborative Computing**, vol. 11, n. 3–4, p. 317–347, 2002.

HOFFMAN, D. L., NOVAK, T.P. **Marketing in hyper-media computer- mediated environments: conceptual foundations**. *Journal of Marketing*, v. 60, n. 2, p. 50-68, 1996.

ISO - **International Organization for Standardization** (ISO, 9241-11). Disponível em pt.wikipedia.org/wiki/Usabilidade. Acessado em 23 de agosto de 2006.

JENA. **Jena: a semantic WEB framework for Java**. Disponível em: <http://jena.sourceforge.net>. Acesso em: 13 de agosto de 2005.

KAMAHARA, J.; ASAKAWA, T.; SHIMOJO, S. and MIYAHARA, H**., A Community-based Recommendation System to Reveal Unexpected Interests**. Proceedings of the 11th International Multimedia Modelling Conference (MMM'05), IEEE, 2005.

KOBIELUS, J. G. **Workflow Strategies**, IDG Books Worldwide, 1997.

KOCH, M.; GEORG G. e HILLEBRAND C. **Mobile Communities – Extending Online Communities into the Real World**. Proc. Americas Conference on Information Systems (AMCIS 2002), Dallas, Tx, pp. 7-18, Aug 2002.

KOCH, Michael., LACHER, Martin S. **Integrating Community Services – A Common Infraestructure Proposal**. Scientific Literature Digital Library, 2000. [on-line] Disponível em <http://citeseer.nj.nec.com/cs> pp. 1-4. Acessado em 22 de agosto de 2005.

KOLLOCK, P. **Design Principles for Online Communities**, in: *The Internet and Society: Harvard Conference Proceedings*, O'Reilly and Associates, Cambridge, pp. 1997.

LEFEVER, L. ***E-mail* Lists and Message Boards – Where is the Middle Ground?**

2003. Disponível em <http://www.commoncraft.com/archives/000401.html>. Acessado em 24/09/2005.

LIECHTI, O.; MARK, S. and ICHIKAWA, T. **Supporting Social Awareness on the World Wide WEB with the Handheld CyberWindow**, in Proceedings of the International Workshop on Handheld CSCW at ACM CSCW'98, Seattle, WA, November 1998.

LIECHTI, O. **Awareness and the WWW: an overview**, ACM SIGGROUP Bulletin, V. 21 , Issue 3 (Dec 2000) P: 3 – 12, ACM Press.

LIPNACK, J. e STAMPS, J., **Networks: redes de conexões**. São Paulo, Aquariana, 1992.

LUGO, G. G. **Um Modelo de Sistemas Multiagentes para Partilha de Conhecimento Utilizando Redes Sociais Comunitárias**. Tese de Doutorado em Sistemas Digitais. Laboratório de Sistemas Inteligentes (LTI) da Universidade de São Paulo (USP). São Paulo, Brasil, 2004.

MARIANI, J.A. **SISCO: Providing a cooperation filter for a shared information space**. In: ACM SIGGROUP Conference on Supporting Groupwork - GROUP'97. *Proceedings*. Phoenix, Arizona, USA. Novembro, pp. 376-384, 1997.

MARTELETO, R. M. **Análise de Redes Sociais – aplicação nos estudos de transferência da informação**. Programa de Pós-Graduação em Ciência da Informação, ECI - UFMG. Ci. Inf., Brasília, v. 30, n. 1, p. 71-81, jan./abr. 2001.

MARTELETO, R. M. e SILVA, A. B. O. **Redes e Capital Social: o enfoque da informação para o desenvolvimento local**. Ci. Inf., Brasília, v. 33, n. 3, p.41-49, set./dez. 2004.

MITCHELL, T. **Machine Learning**. McGraw-Hill, 1997.

MOECKEL, A. **Modelagem de processos de desenvolvimento em ambiente de engenharia simultânea: implementações com as tecnologias Workflow e BSCW**. Curitiba, 2000. 175 f. Dissertação (Mestrado em Tecnologia) – PPGTE, CEFET-PR.

MONARD, M. C.; BATISTA, G. E. A. P. A.; KAWAMOTO, S.; PUGLIESI, J. B. Uma Introdução ao Aprendizado Simbólico de Máquina por Exemplos. Notas Didáticas Número 29, ICMC-USP, 1997.

MONGOOSE Technology, **The 12 Principles of Civilization - Guidelines for Designing Interactive Internet Services**, 2001, disponível em http://www.mongoosetech.com/realcommunities/12prin.html, acessado em 20/10/2005.

MONTMOLLIN, M.. **Ergonomia**. Lisboa: Instituto Piaget, 1984.

MURATA, T., **Visualizing the Structure of WEB Communities Based on Data Acquired From a Search Engine**. IEEE Transaction on Industrial Electronics, Vol. 50, no. 5, October, 2003.

MYNATT, E.; O'DAY, V.; ADLER, A. and ITO, M. **Design for Network Communities**. In Proc. ACM SIGCHI Conf. on Human Factors in Compt. Syst., 1997.

NIINIVAARA, O. **Agent-Based Recommender Systems**. Software Agent Techno*log*y Course Paper. University of Helsinki, Department of Computer Science, pp. 1-51, 2004.

PALAZZO, L. A. M, ULYSSÉA, M. C. e PORTO, P. R. **Comunidades Virtuais de Aprendizado Adaptativo**, *Conferência Nacional em Ciência, Tecnologia e Inovação*,

Ministério da Ciência e Tecnologia, Florianópolis, pp. 2001.

PALIOURAS, G.; PAPATHEODOROU, C.; KARKALETSIS V. and SPYROPOULOS, C. D, **Discovering user communities on the Internet using unsupervised machine learning techniques**. Interacting with Computers, Vol. 14(6), pp. 761-791, 2002.

PALLOFF, R., PRATT, K., **Building Learning Communities in Cyberspace: effective strategies for the online classroom**. San Francisco: Jossey-Bass, 1999. 320p.

PEDERSEN, E.R., SOKOLER, T., **AROMA : abstract representation of presence supporting mutual awareness**, in *Proceedings of CHI '97*, ACM Press, 1997.

PÓR, G. **What is a "Community of Practice"?**. Disponível em http://www.co-i-l.com/coil/knowledge-garden/cop/definitions.shtml. Acessado em 25/09/2005.

PREECE, J. **Online communities: designing usability, supporting sociability**. Chichester: Wiley, pp. 464, 2000.

PREECE, J. **Sociability and Usability: Twenty years of chatting online**. In Behavior and Information Techno*log*y Journal, 20, 5, 347-356. 2001. Disponível em <http://www.ifsm.umbc.edu/~preece/paper/4 BIT Twenty years.pdf>. Acessado em 24/09/2005.

PRIMO, A.F.T.; BRAMBILLA, A.M.. **Social software e construção do conhecimento**. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 27., 2004. Porto Alegre. Anais... São Paulo: Intercom, 2004.

PRINZ, W.; KOLVENBACH, S. and KLOCKNER, K. **Situative Cooperation Support for Communities**, SIGGROUP Bulletin December 2002/Vol. 23 No. 3.

RECUERO, R. C. Redes **Sociais na Internet:Considerações Iniciais**. In: XXVIII INTERCOM, 2005. Rio de Janeiro/RJ. Ecompos, Internet, v. 2, n. Abril 2005.

REDDY, K. P. and KITSUREGAWA, M., **An approach to relate the WEB communities through bipartite graphs**. IEEE Computer Society Press, 2002.

RESNICK, P. and VARIAN, H., **Recommender Systems**. Communications of the ACM, 40, 3, 56-58, Vol. 40, No. 3, pp. 56-58., 1997.

RESTREPO, Mariluz J. e ANGULO, Jaime Rubio. Intervir en la organización. Bogotá : Significantes de Papel Ediciones, 1992.

RHEINGOLD, H. **The Virtual Community: Homesteading at the Electronic Frontier**, 1993, available at http://www.rheingold.com/vc/book, acessado em setembro de 2005.

RHEINGOLD, H. **La Comunidad Virtual: Una Sociedad sin Fronteras**. Gedisa Editorial. Colección Limites de La Ciência. Barcelona, 1994.

SALTON, G.; McGILL, M. J. **Introduction to Modern Information Retrieval**. McGraw-Hill. New York, 1983.

SALTON, G., **Automatic Text Processing: The Transformations, Analysis, and Retrieval of Information by Computer**, Addison-Wesley, 1989.

SCHILIT, B.N.; ADAMS, N.I. e WANT, R., **Context-Aware Computing Applications**. In Proceedings of the Workshop on Mobile Computing Systems and Applications. IEEE Santa Cruz, CA: Computer Society, 1994. pp. 85–90.

SCHLICHTER, J.; KOCH, M. e XU, C. **Awareness – The Common Link Between Groupware and Community Support Systems**. In T. Ishida, editor, Community Computing and Support Systems, pages pp. 77-93. Springer Verlag, 1998.

SHANKAR, S., and KARYPIS, G. **A Feature Weight Adjustment Algorithm for Document Categorization**, *KDD-2000 Workshop on Text Mining*, Boston, USA, August 2000.

SMITH, Ana Du Val. **Problems in Conflict managemente in Virtual Communities**. *In* KOLLOCK Peter. e Marc Smith. (organizadores) *Communities in Cyberspace.* Routledge. New York, 1999.

SOUZA, C. S. e PREECE, J. **A framework for analyzing and understanding online communities**. In: Interaction with Computers, The Interdisciplinary Journal of Human-Computer Interaction. Inglaterra: Butterworth Scientific, 2004. Disponível em http://www.ifsm.umbc.edu/~preece/Papers/Framework_desouza_preece2003.pdf Accessed in 24/09/2005.

STEINFIELD, C. ; JANG C. Y. e PFAFF, B. **Supporting Virtual Team Collaboration: the TeamSCOPE system**. In GROUP' 99: Proceedings of the international ACM SIGGROUP conference on Supporting Group Work, ACM Press, 1999, pp. 81-90.

STENMARK, D. **The Relationship between Information and Knowledge**. In Proceedings of the 24th Information Systems Research Seminar in Scandinavia - IRIS 2001(Aug. 11-14, Ulvik in Hardanger, Norway). 2001.

TACLA, C. A and ENEMBRECK, F., **An Awareness Mechanism for Enhancing Cooperation in Design Teams**. The 9th International Conference on Computer Supported Cooperative Work in Design Proceedings. 2005. pp. 920-925.

TERRA, J. C. C., **Comunidades de Prática: conceitos, resultados e métodos de gestão**. Disponível em www.terraforum.com.br, acessado em 04 de dezembro de 2005.

TJARA, Sanmya F. **Comunidades Virtuais: um fenômeno na sociedade do conhecimento**. 1ª ed. São Paulo: Érica, 2002.

VALTERSSON, M., **Virtual Communities**. VIRCOM – Virtual Communities, 2002. [on-line] Disponível em http://www.informatik.umu.se/nlrg/valter.html . Acessado em 18 de outubro de 2005.

VAN RIJSBERGEN, C. J. **Information Retrieval**. 2nd edition, London, Butterworths, 1979.

VIVACQUA, A.; MORENO, M. and SOUZA, J., **CUMBIA: An Agent Framework to Detect Opportunities for Collaboration**. Proceedings of the 9th International Conference on Computer Supported Cooperative Work in Design. 2005. pp. 417-422.

VYGOTSKY, Lev S. **Pensamento e linguagem**. 2ª ed. São Paulo: Martins Fontes, 1998.

WATTS, D. J. **Six Degrees. The Science of a Connected Age.** New York: W. W. Norton &Company, 2003.

WELLMAN, B.. **Physical Place and CyberPlace: The Rise of Personalized Networking.** Fevereiro de 2001. Disponível em: <http://www.chass.utoronto.ca/~wellman/publications/individualism/ijurr3a1.htm>.

Acesso em outubro de 2006.

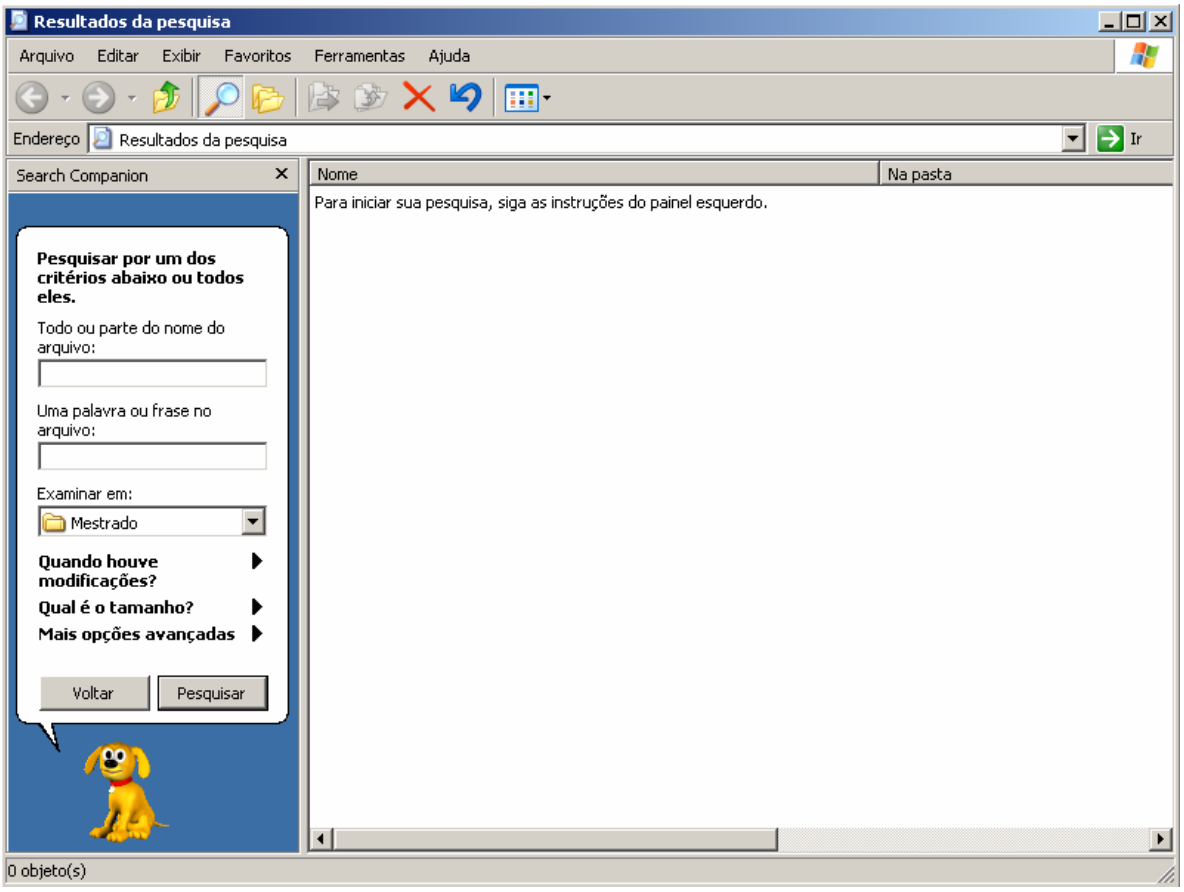
WENGER, E**. Communities of Practice and Social Learning Systems**, Organization, 7 (2), 225-256, 2000.

# APÊNDICE

## APÊNDICE 1 - Orientações para a Coleta de Dados

A tarefa consiste em coletar todos os documentos (com extensões: .doc; .txt; .pdf; .htm; .html) que tenham sido criados, modificados ou acessados no período de Janeiro a Junho de 2006. Esses documentos devem ser separados mês a mês e de acordo com a atividade realizada com cada um deles.
Abaixo seguem os passos a serem seguidos para proceder a coleta.

1. Com o botão direito do mouse, clicar sobre o disco (C:\ ; D:\ ou outro no qual são armazenados os seus documentos de trabalho). Aparece um menu no qual deve ser selecionada a opção Pesquisar.
2. Aparecerá uma janela como essa (ou semelhante):

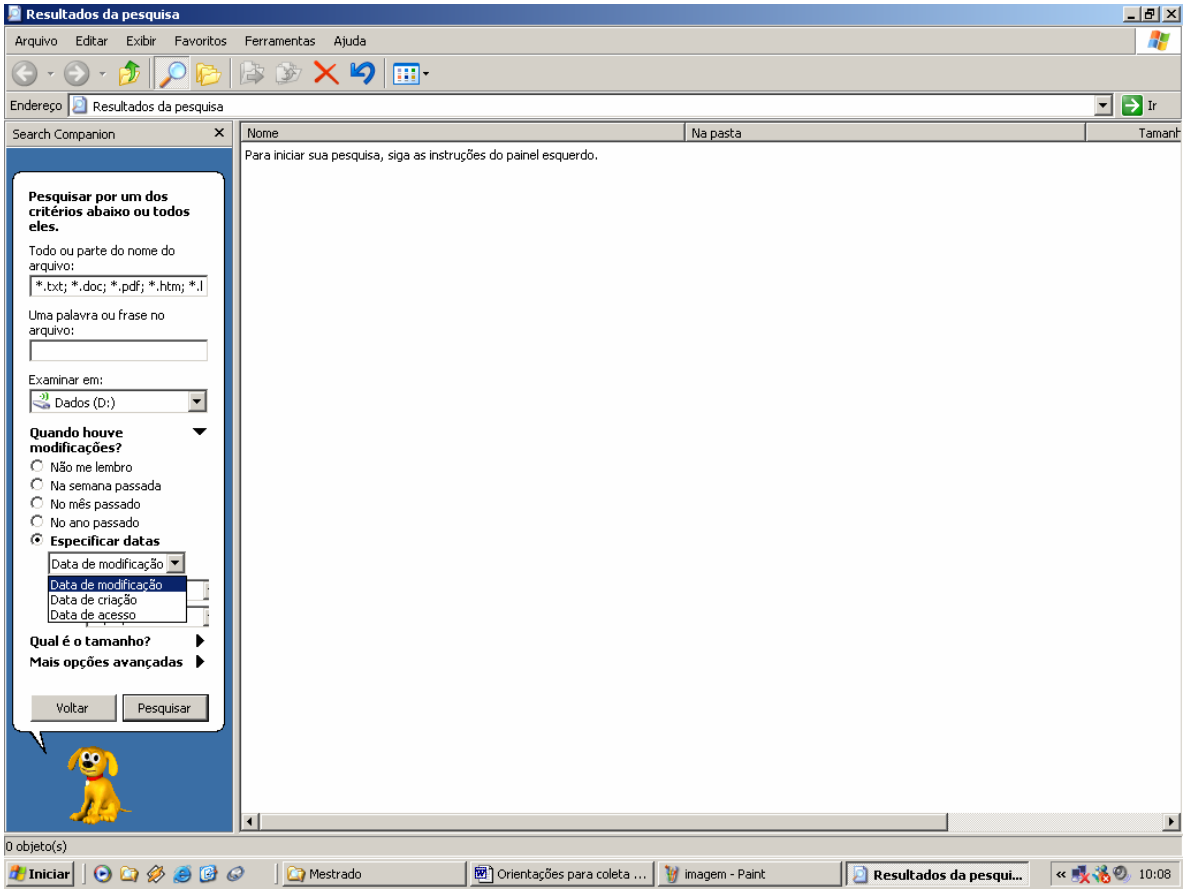



3. No campo onde diz: “Todo ou parte do nome do arquivo:” digitar o seguinte: *.txt; *.doc; *.pdf; *.htm; *.html

Isso fará com que sejam listados todos os arquivos com tais extensões de uma só vez.

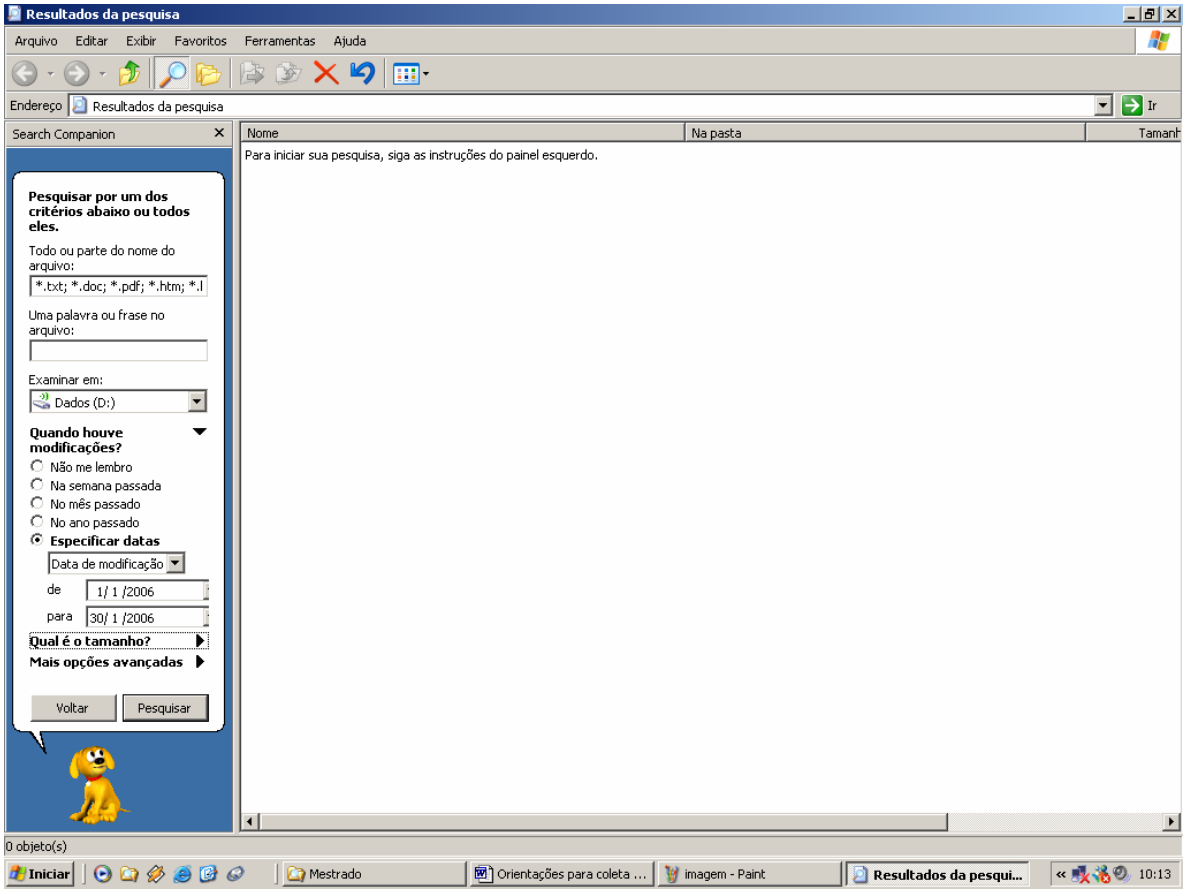
4. No campo “Examinar em:” deve ser selecionado o disco ou diretório onde se encontram os arquivos a serem pesquisados, caso ainda não esteja selecionado.

5. Clicar sobre o link que diz: “Quando houve modificações?”, então ele será expandido da seguinte forma:




6. Você deve clicar sobre a opção Especificar datas, na qual serão abertos alguns campos a serem configurados. No primeiro deles, conforme aparece na janela acima, é para selecionar “Data de modificação”, “Data de criação” e “Data de acesso”. Primeiramente deve ser selecionado “Data de modificação”, ao concluir todo o processo de pesquisa para essa opção (fazendo a coleta mês a mês conforme a seguir) você deve repetir a pesquisa para as outras duas opções.

7. Feito isso, você deve informar o período da pesquisa. Como você precisa fazer a pesquisa mês a mês, de janeiro a junho de 2006, informe a data inicial como 01/01/2006 até 30/01/2006, e clique em Pesquisar, conforme mostrado abaixo.




8. Listados os arquivos modificados para o mês de janeiro, você deve criar uma pasta chamada Janeiro, para a qual irá copiar os arquivos pesquisados.

9. Você deve repetir os itens 7 e 8 para os outros meses, até junho de 2006.

Obs.:
- Ao concluir a pesquisa para a opção “Data de modificação”, você deve repetir os itens 6, 7, 8 e 9 para as outras duas opções: “Data de criação” e “Data de acesso”, copiando os arquivos pesquisados para as mesmas pastas já criadas. Sendo assim, os arquivos modificados, criados e acessados ficarão misturados, separados mês a mês.

- Ao final dessa pesquisa, você deve ter seis pastas denominadas respectivamente: Janeiro, Fevereiro, Março, Abril, Maio e Junho, contendo seus respectivos arquivos modificados, criados e acessados.

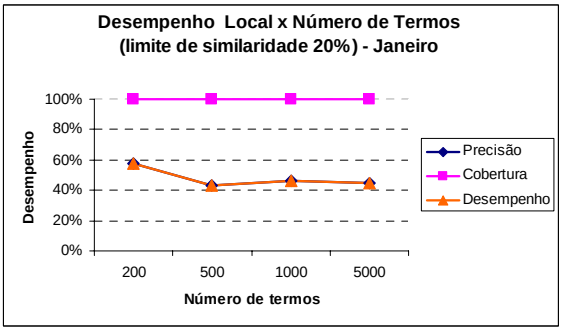
10. Feito isso, para cada pasta criada (Ex. Janeiro), você deve criar sub-pastas cujo nome deve identificar a atividade que foi realizada usando um determinado documento. Ou seja, você deve classificar os documentos de acordo com a atividade realizada com ele, e para isso deverá criar sub-pastas com o nome da atividade e copiá-los para dentro dela. Você poderá criar quantas sub-pastas forem necessárias. As atividades podem ser diversas (Escrever artigo sobre determinado assunto, Preparar aula sobre determinado assunto, Pesquisar na Internet sobre determinado assunto, entre outras).

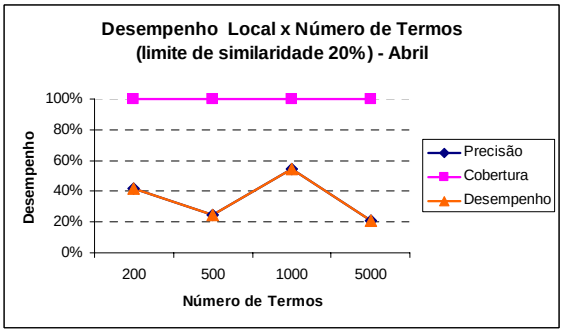
Por exemplo, você utilizou alguns documentos para “Escrever artigo sobre Agentes Inteligentes”, esse pode ser o nome da sub-pasta criada, e todos os arquivos utilizados para essa atividade deverão ser copiados para dentro dela.

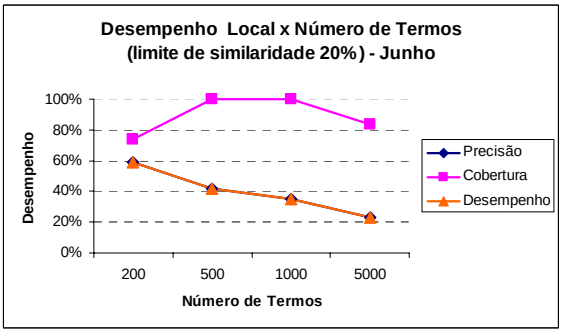
Obs.: O procedimento 10 deve ser realizado para todos os outros meses.

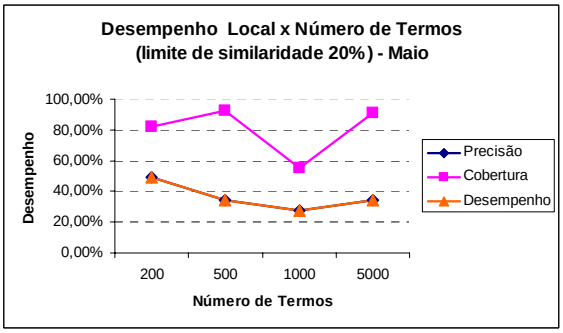
# APÊNDICE 2 – Desempenho Local : Limite de Similaridade x Número de Termos

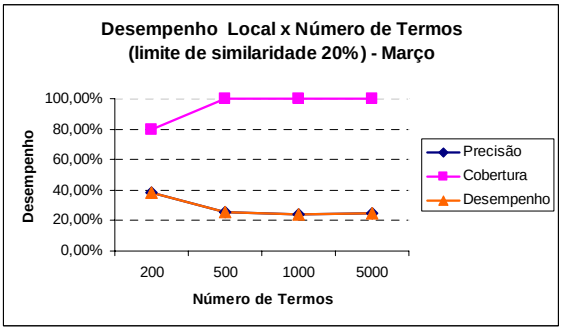
## Limite de similaridade de 20%

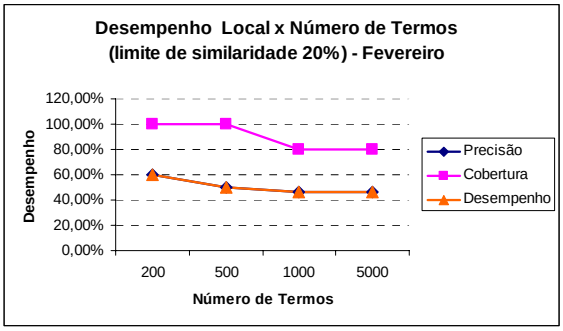


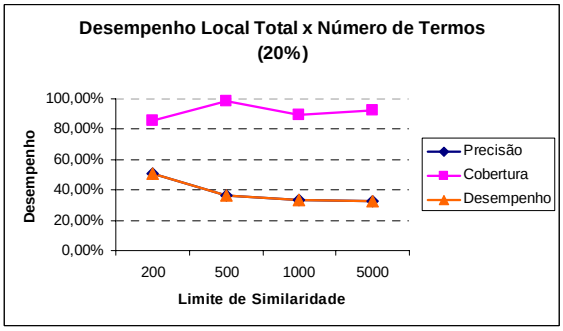



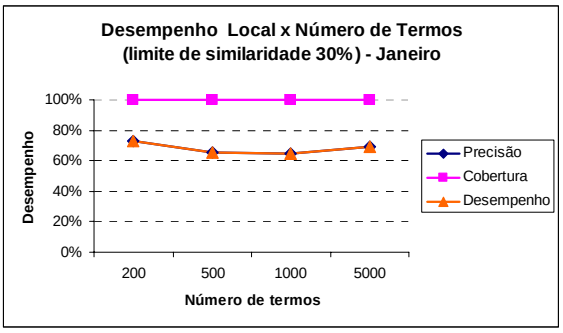



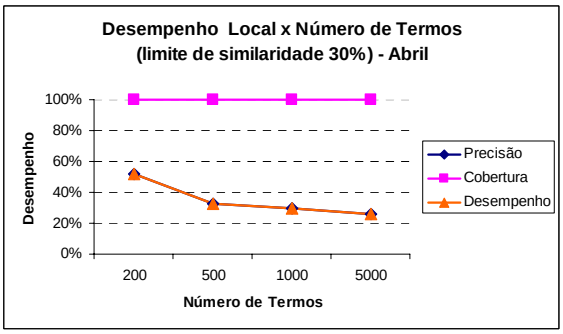



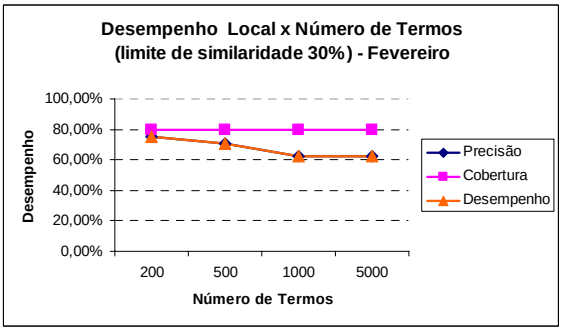



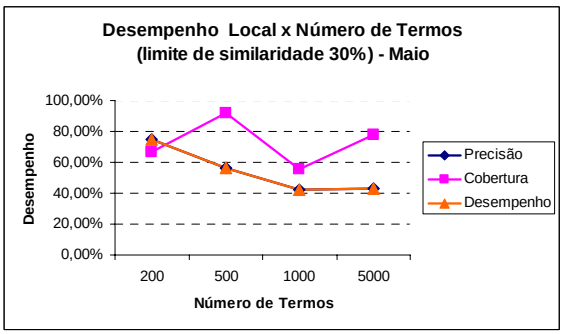



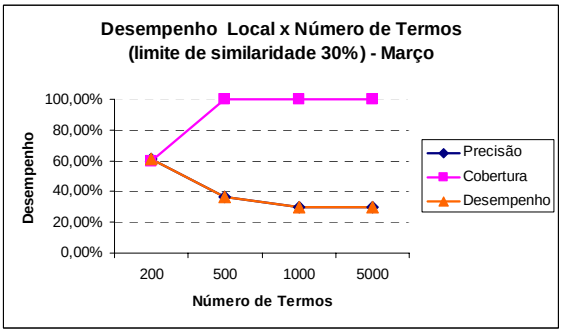
## Limite de similaridade de 30%

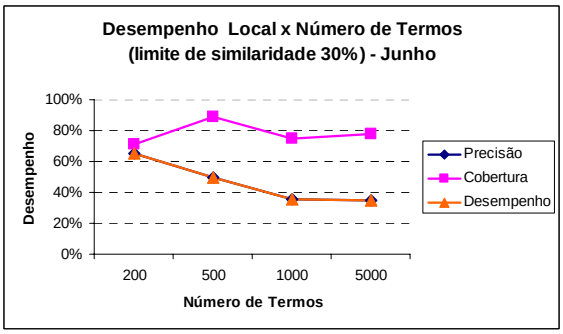


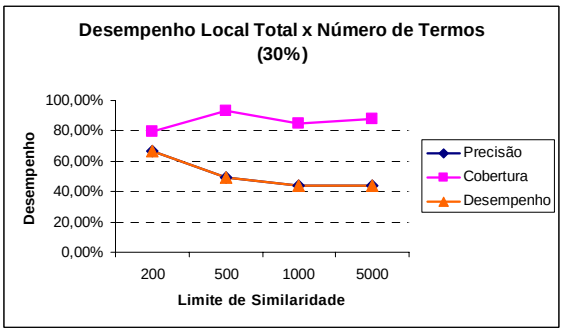



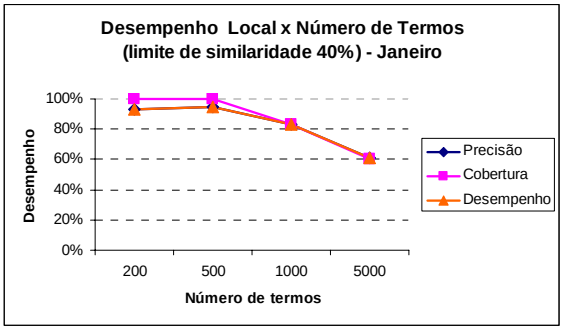



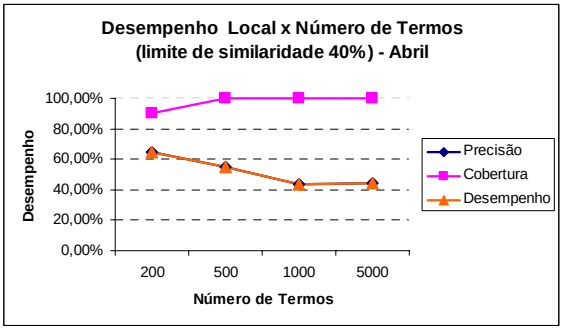



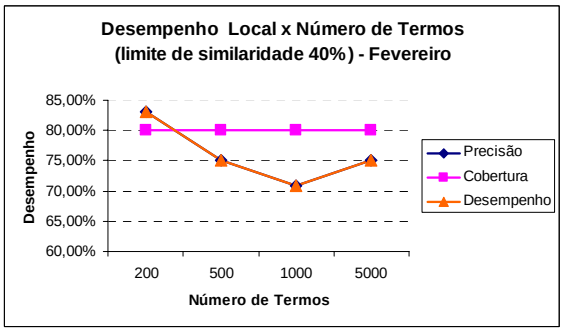



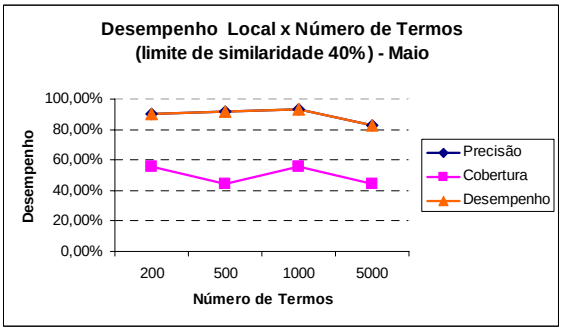



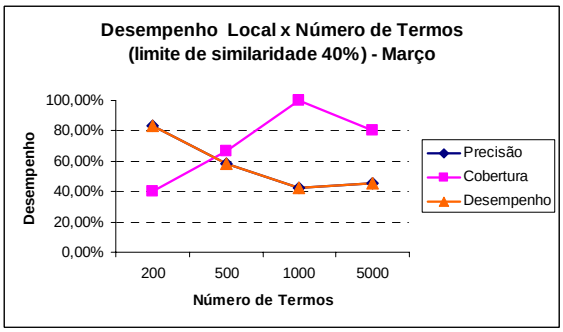
## Limite de similaridade de 40%

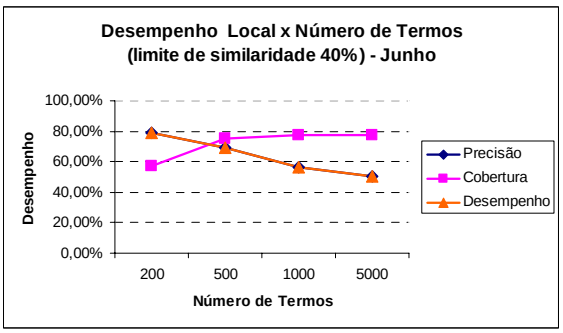


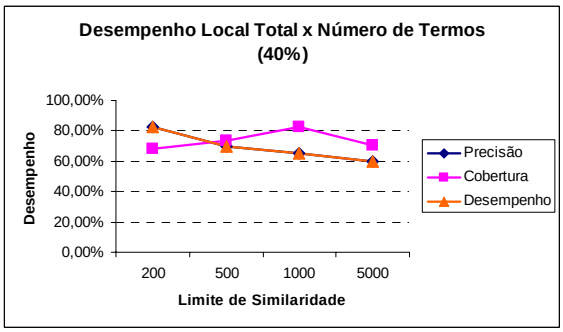



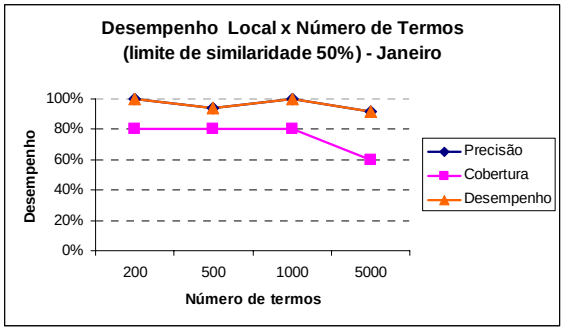



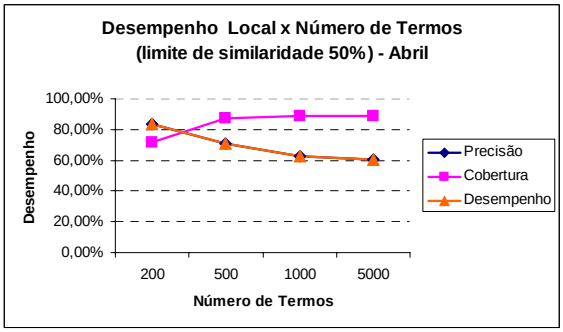



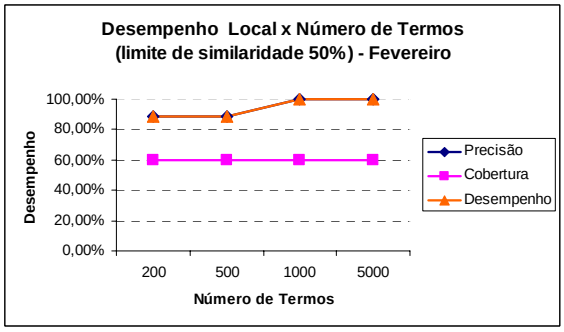



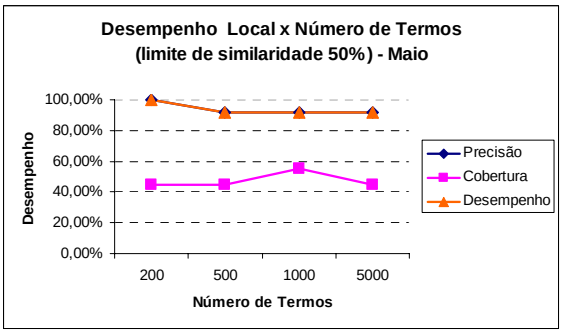



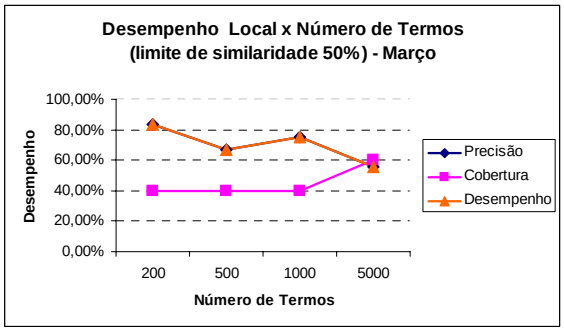
## Limite de similaridade de 50%

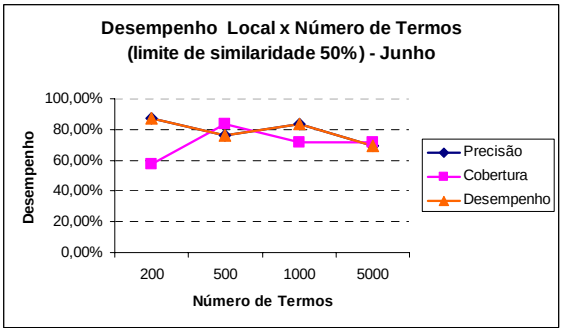


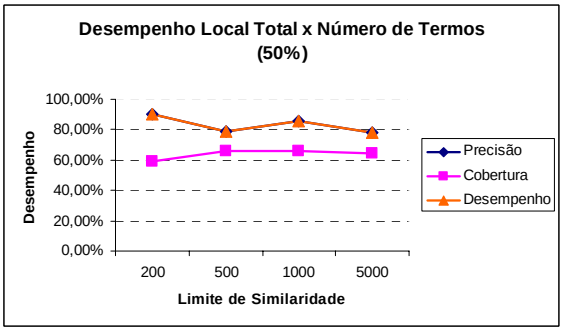



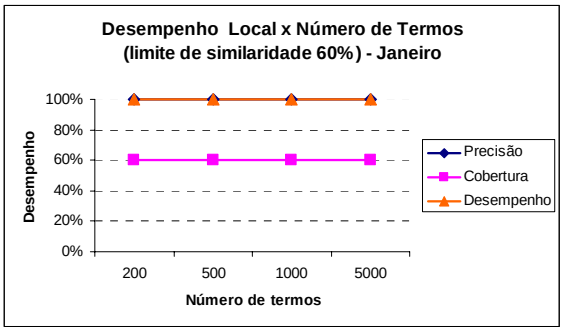



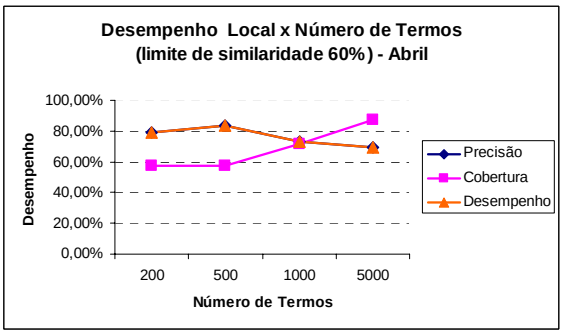



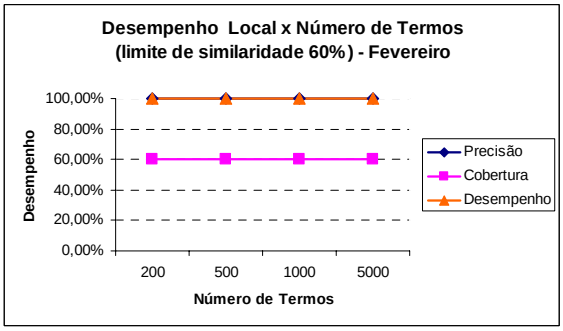



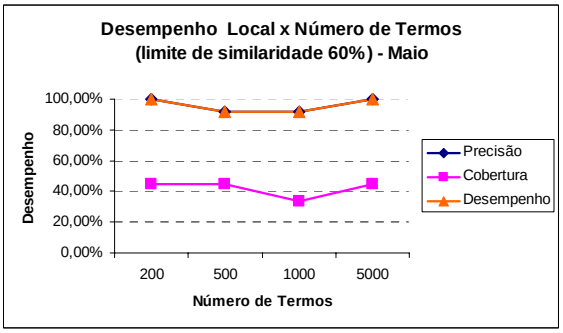



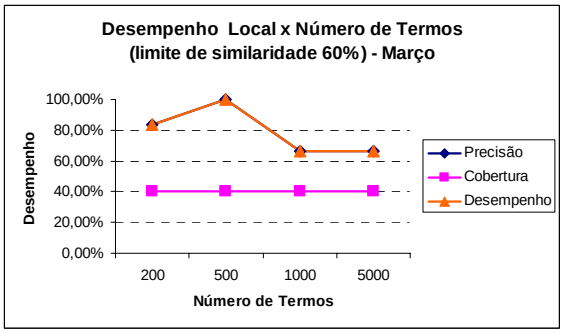
## Limite de Similaridade 60%

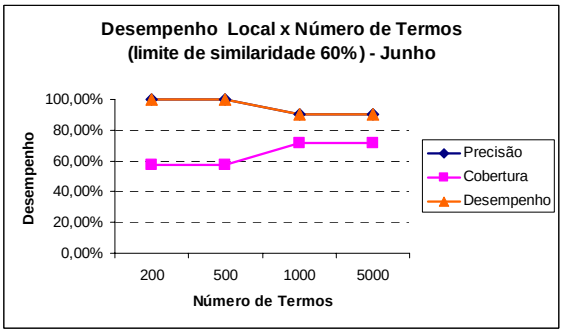


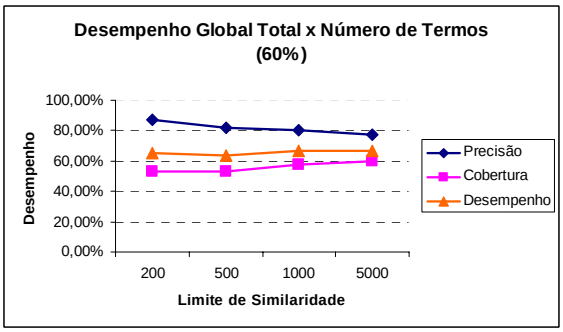



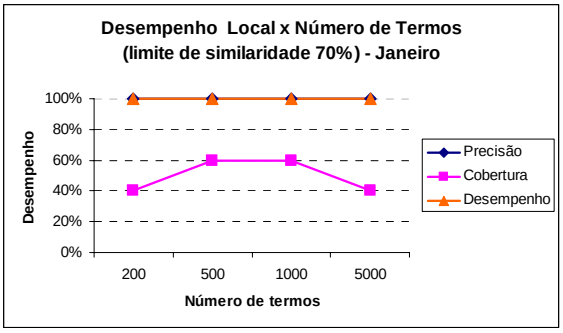



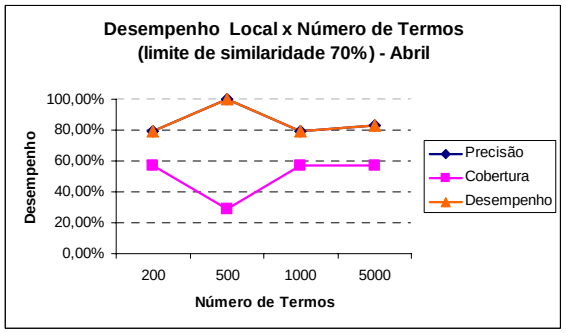



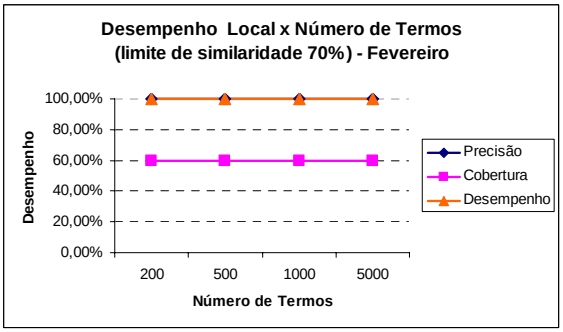



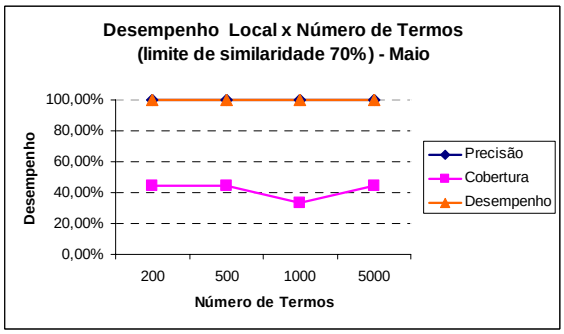



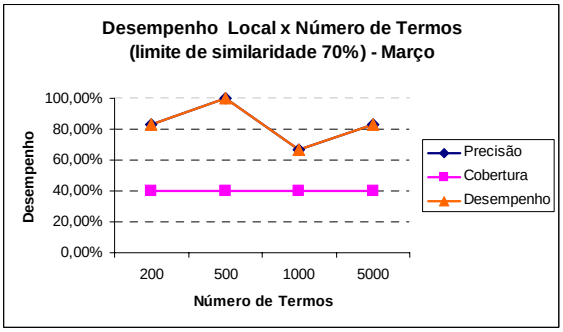
## Limite de Similaridade de 70%

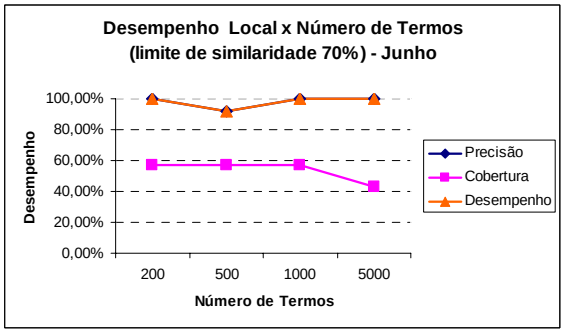


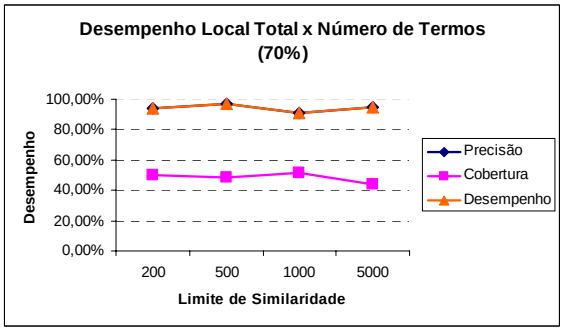



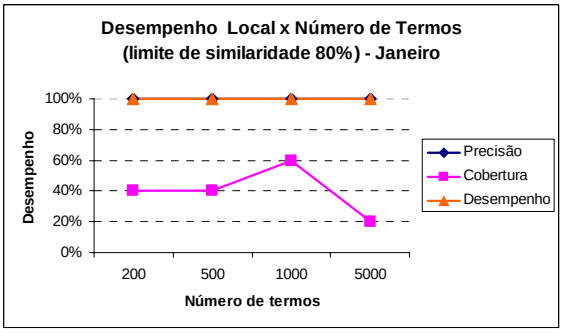



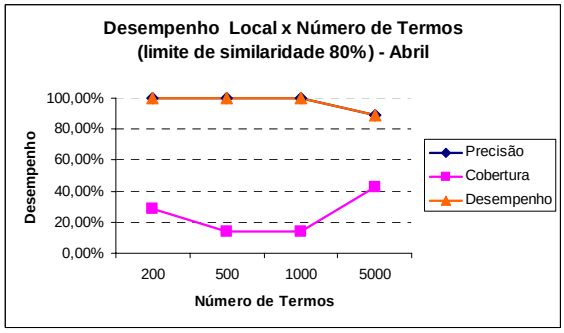



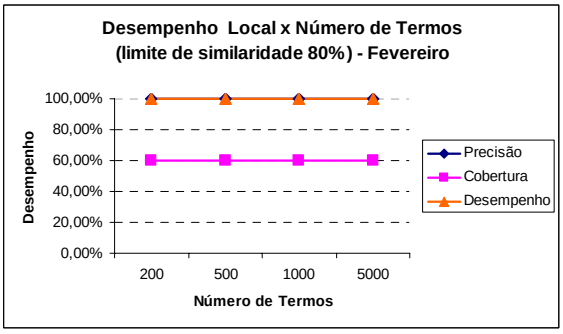



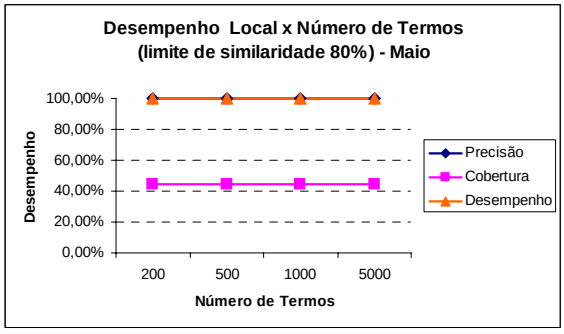



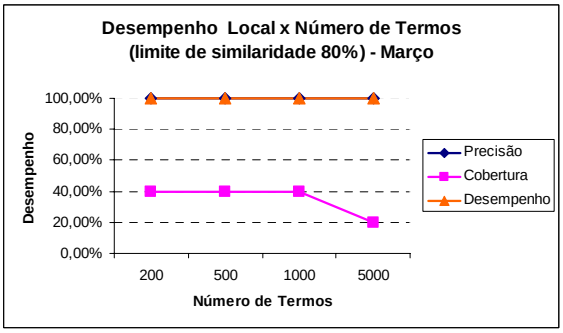
## Limite de Similaridade de 80%

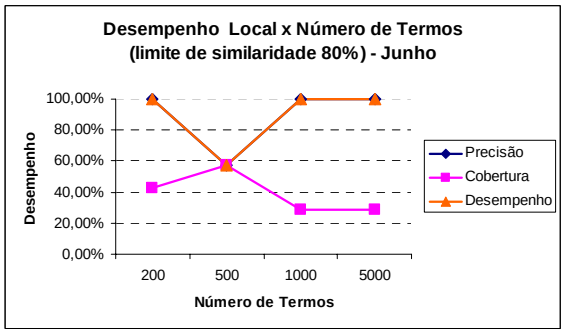


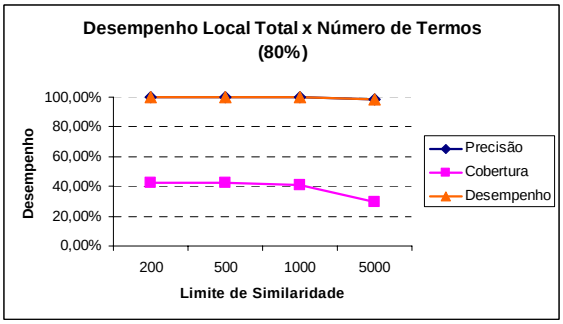



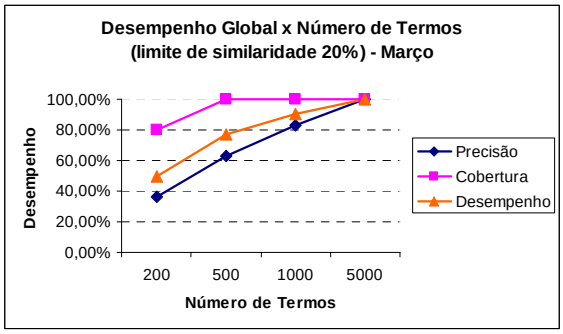



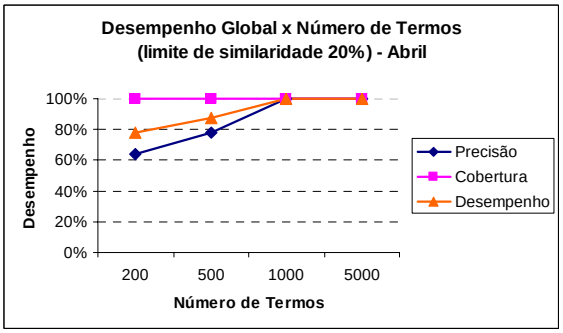



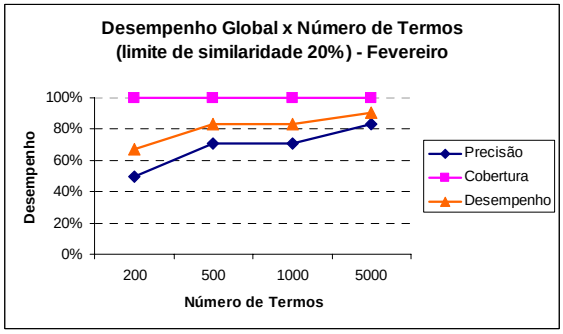



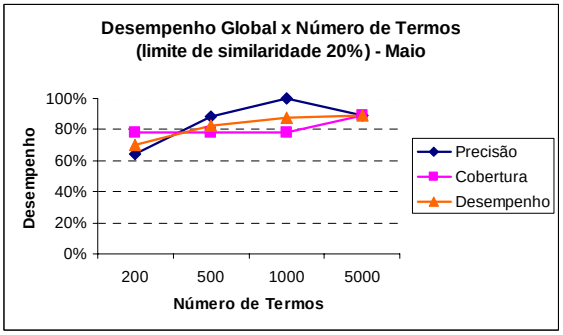
# APÊNDICE 3 – Desempenho Global: Limite de Similaridade x Número de Termos

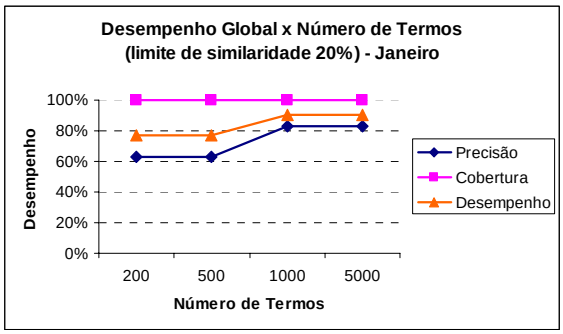
## Limite de similaridade de 20%

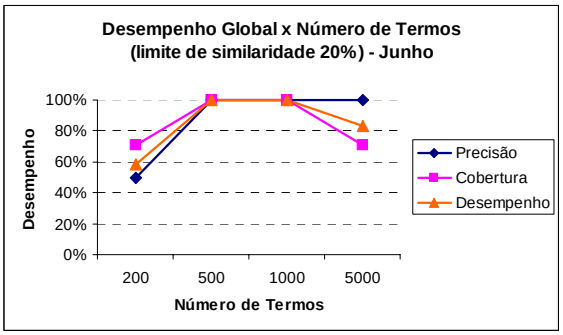


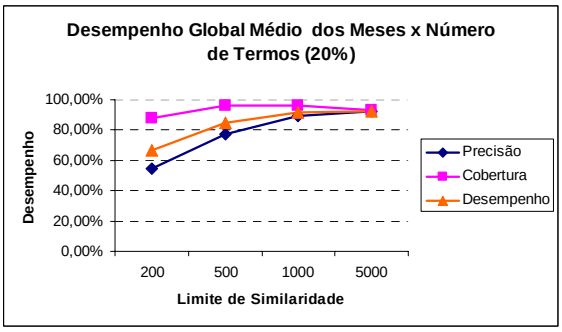



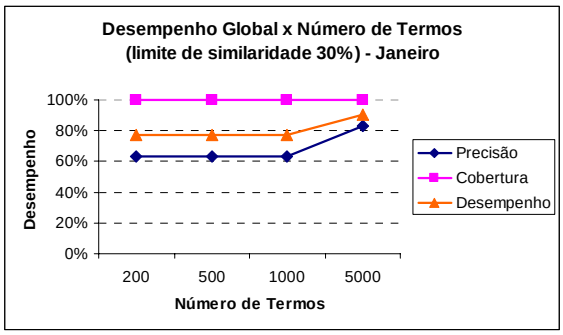



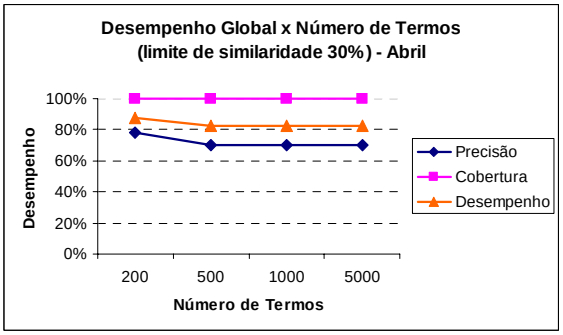



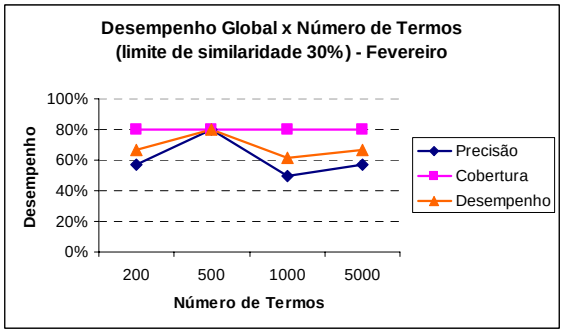



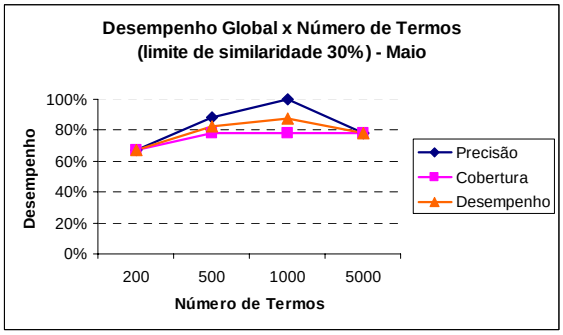



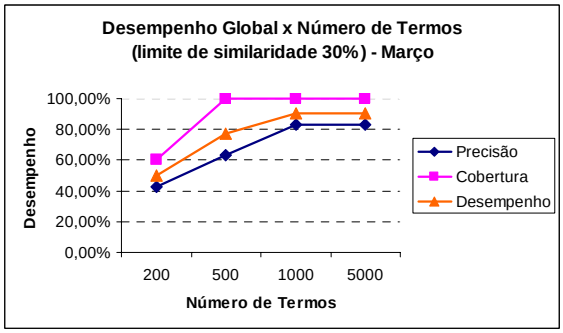
## Limite de similaridade de 30%

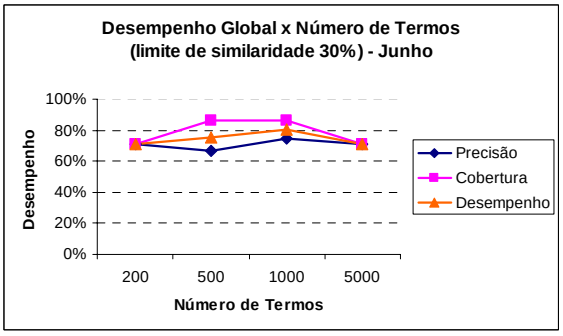


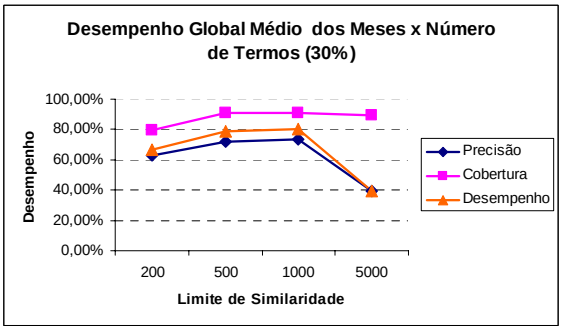



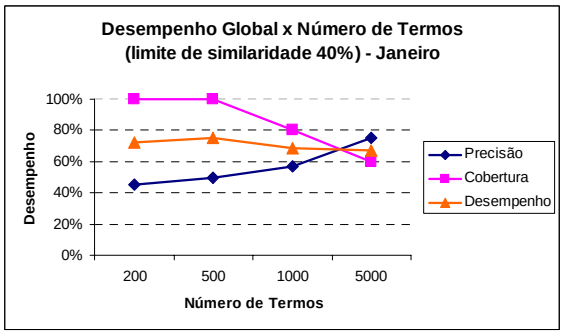



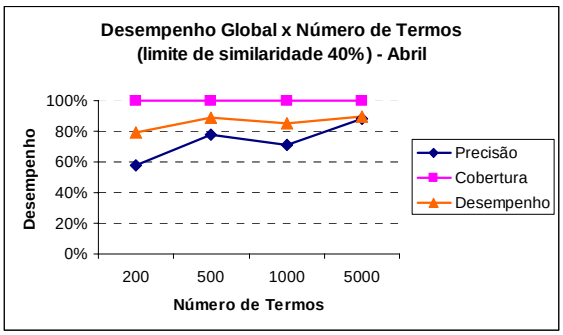



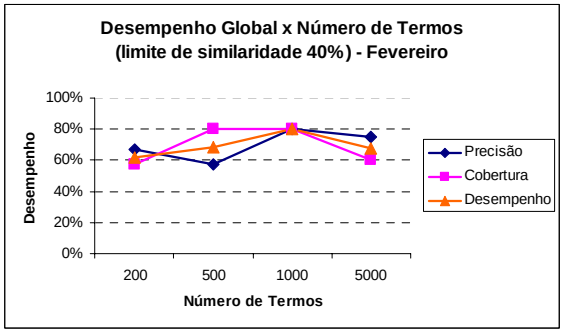



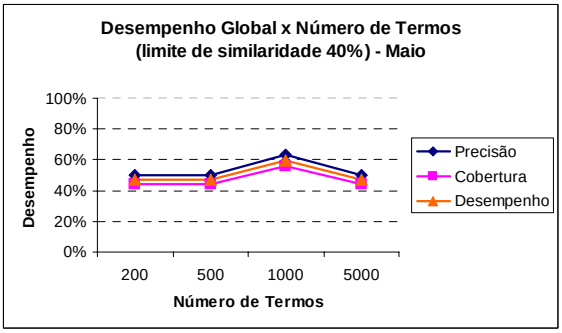



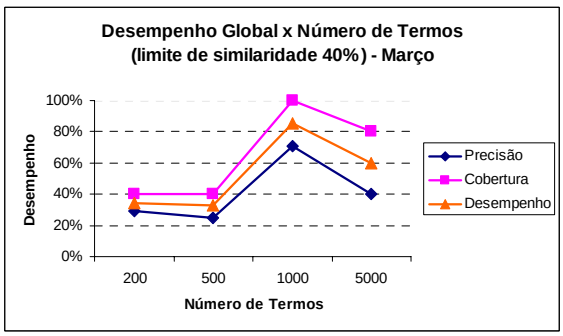
## Limite de similaridade de 40%

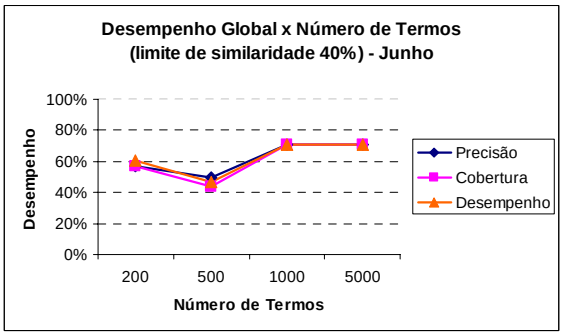


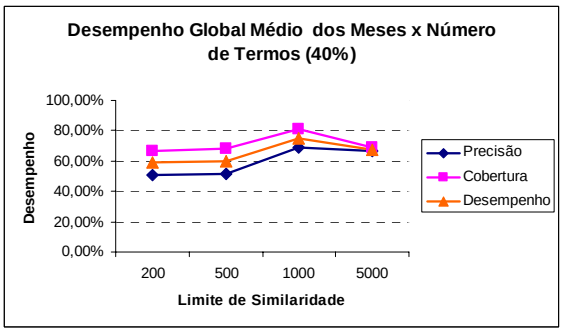



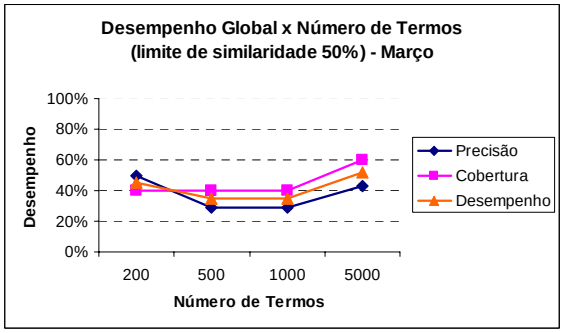



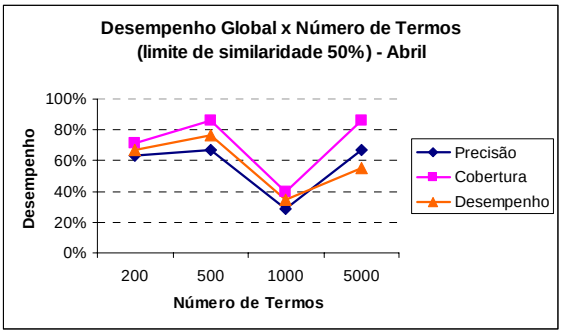



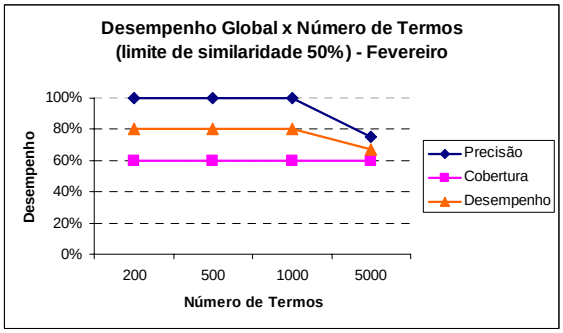



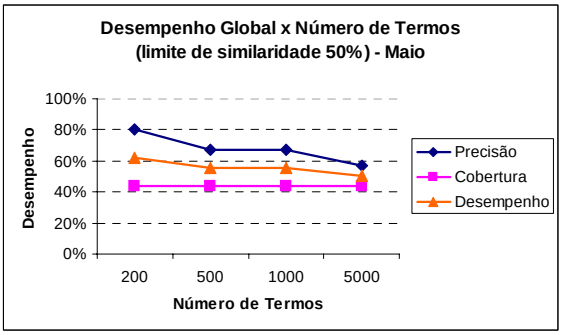



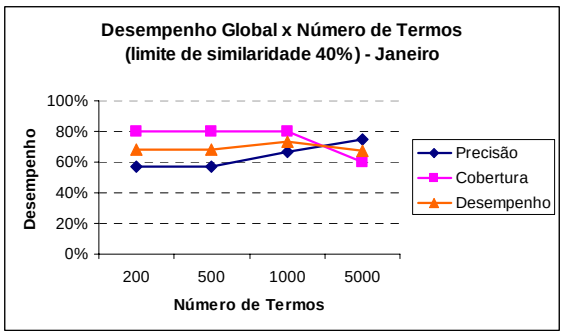
## Limite de similaridade de 50%

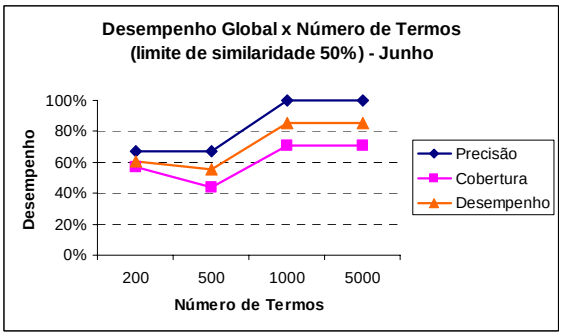


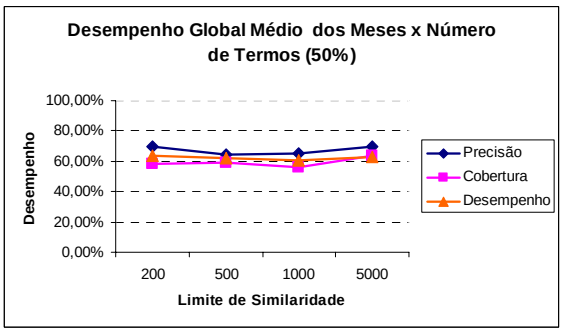



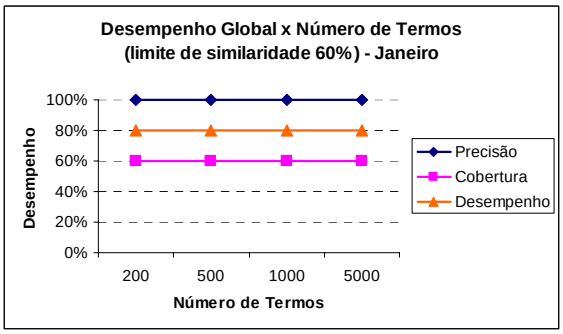



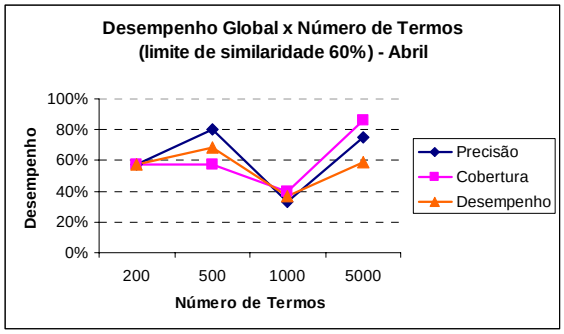



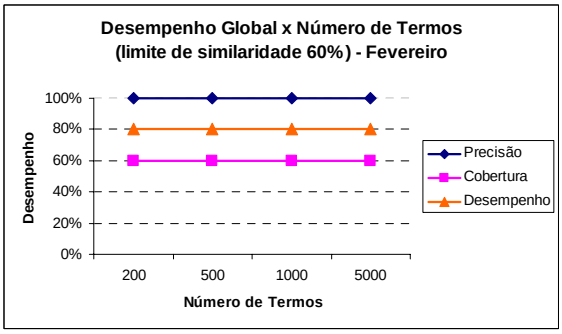



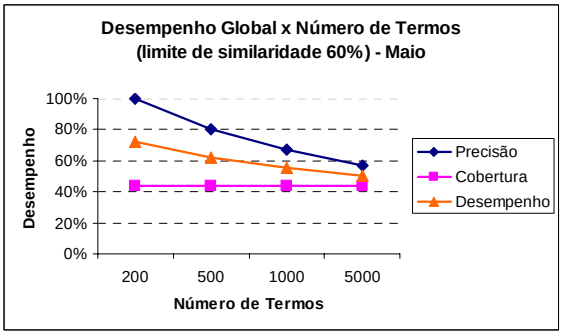



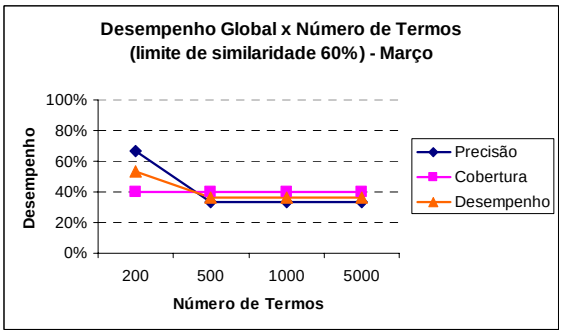
## Limite de Similaridade 60%

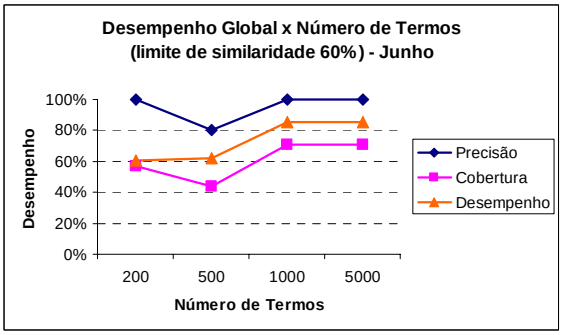


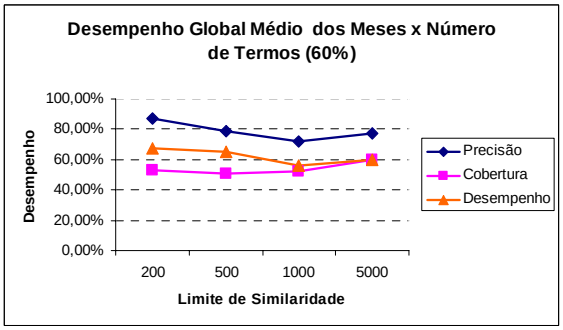



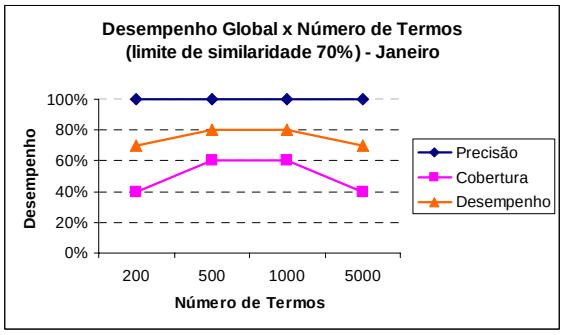



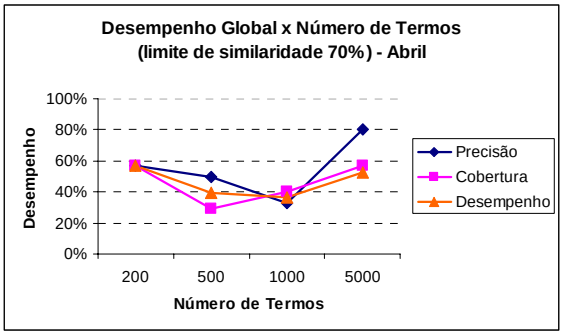



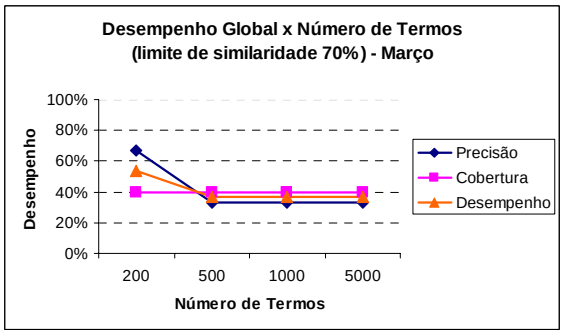



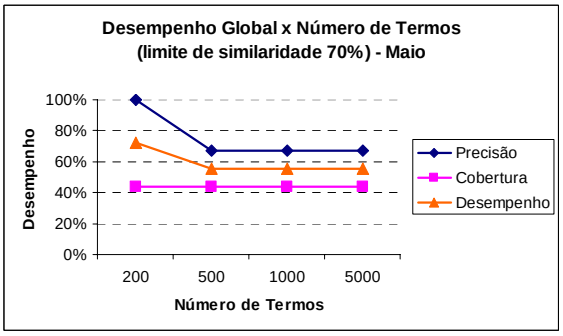



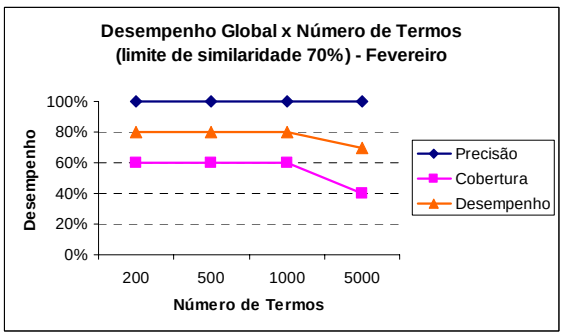
## Limite de Similaridade de 70%

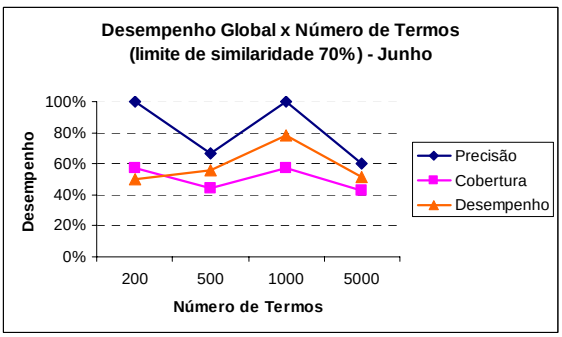


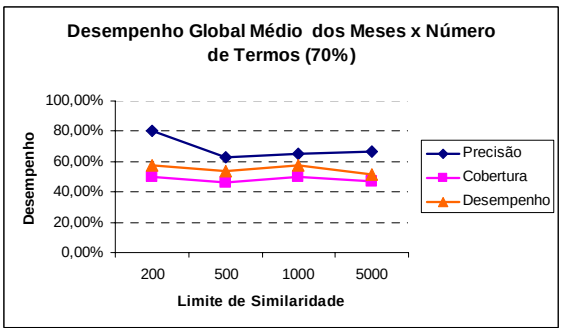



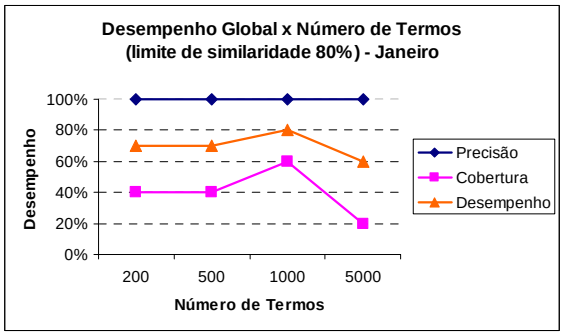



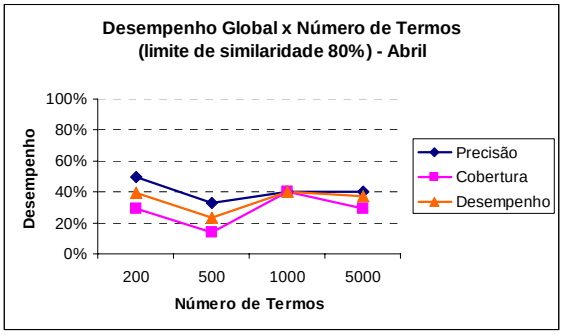



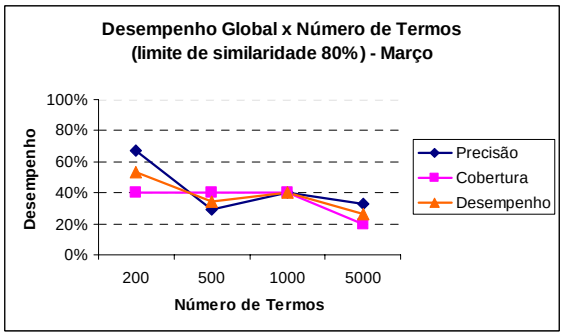



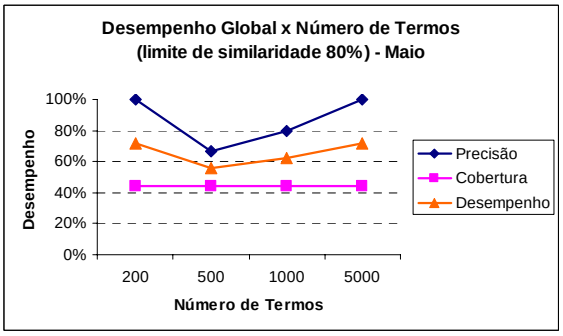



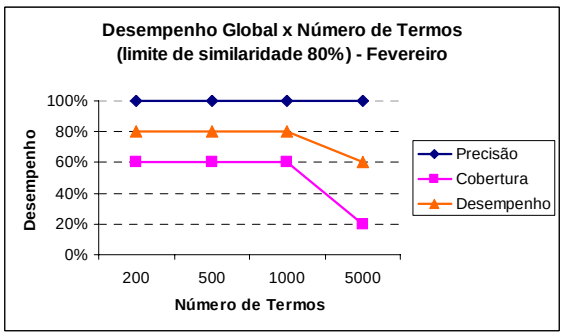
## Limite de Similaridade de 80%

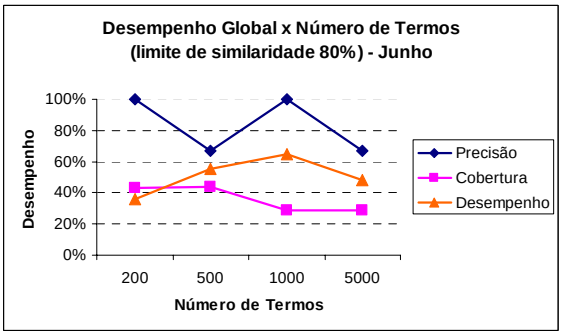


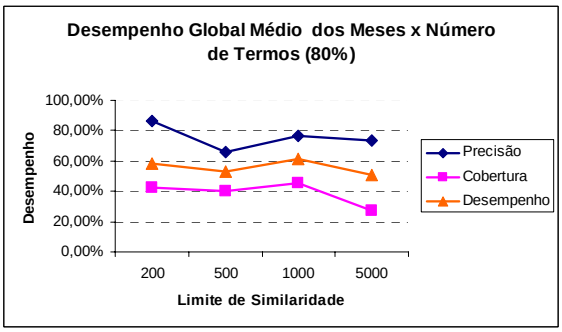



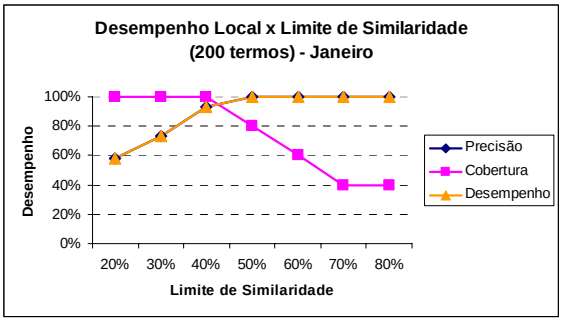



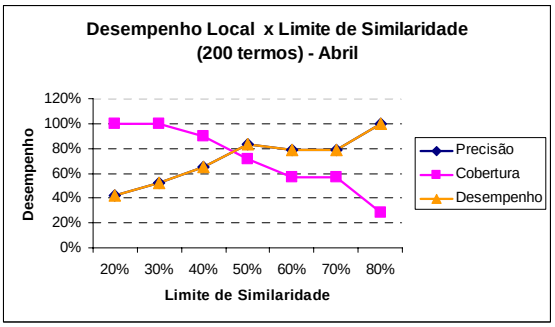



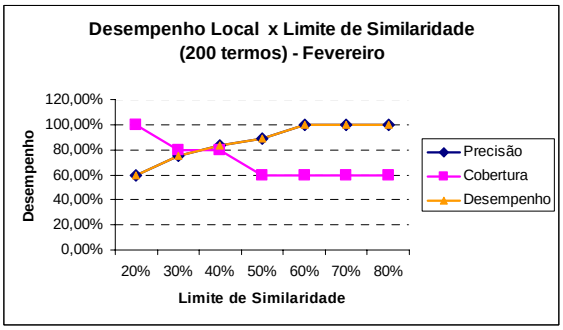



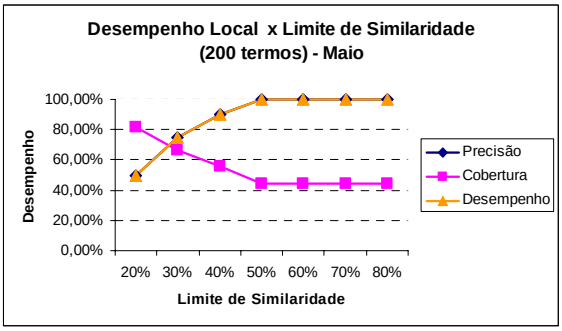
# APÊNDICE 4 – Desempenho Local: Tamanho do Vetor de Contexto x Limite de Similaridade

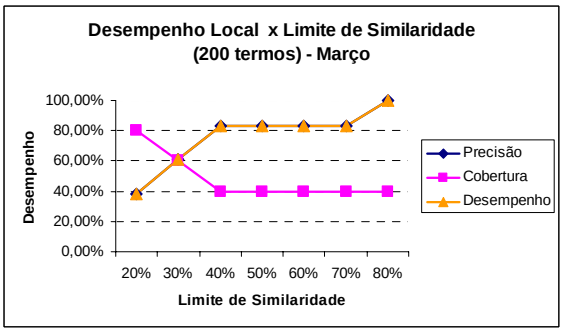
## Vetores de 200 termos

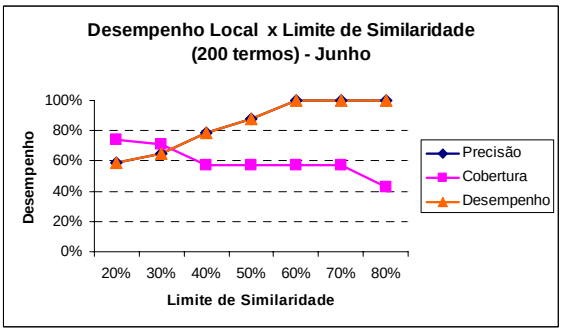


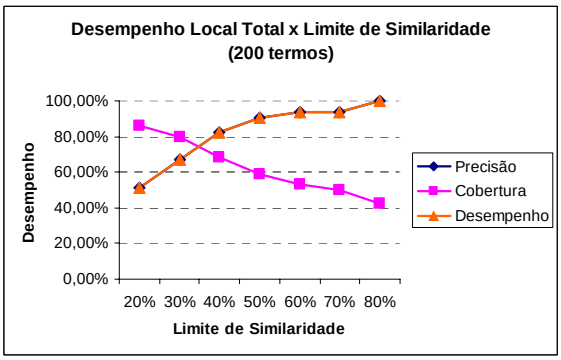



## Desempenho Local Total - Vetores de 200 termos

### Precisão

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 57,86% | 60,00% | 38,00% | 42,00% | 49,26% | 59,00% | ***51,02%*** |
| **30%** | 72,92% | 75,00% | 61,00% | 52,00% | 75,00% | 65,00% | ***66,82%*** |
| **40%** | 93,33% | 83,00% | 83,33% | 64,80% | 90,00% | 79,00% | ***82,24%*** |
| **50%** | 100,00% | 88,88% | 83,33% | 83,33% | 100,00% | 87,50% | ***90,51%*** |
| **60%** | 100,00% | 100,00% | 83,33% | 79,16% | 100,00% | 100,00% | ***93,75%*** |
| **70%** | 100,00% | 100,00% | 83,33% | 79,16% | 100,00% | 100,00% | ***93,75%*** |
| **80%** | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | ***100,00%*** |

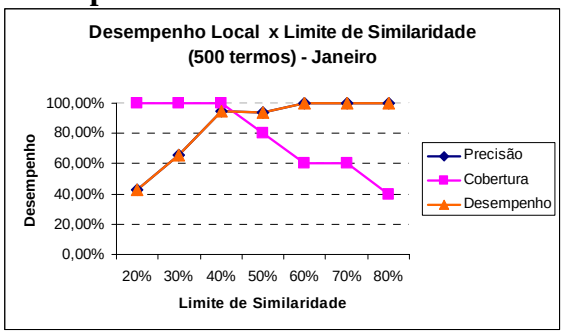
### Cobertura

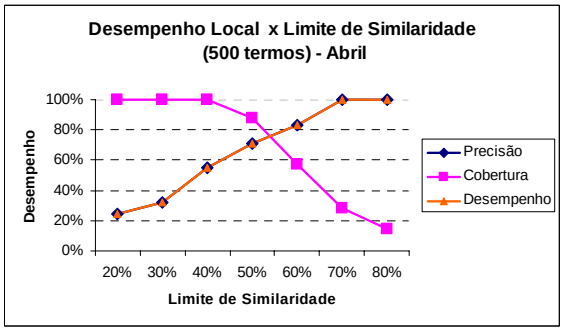
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 100,00% | 80,00% | 80,00% | 100,00% | 81,82% | 74,07% | ***85,98%*** |
| **30%** | 100,00% | 80,00% | 60,00% | 100,00% | 66,67% | 70,83% | ***79,58%*** |
| **40%** | 100,00% | 80,00% | 40,00% | 90,00% | 44,44% | 57,14% | ***68,60%*** |
| **50%** | 80,00% | 60,00% | 40,00% | 71,43% | 44,44% | 57,14% | ***58,84%*** |
| **60%** | 60,00% | 60,00% | 40,00% | 57,14% | 44,44% | 57,14% | ***53,12%*** |
| **70%** | 40,00% | 60,00% | 40,00% | 57,14% | 44,44% | 57,14% | ***49,79%*** |
| **80%** | 40,00% | 60,00% | 40,00% | 28,57% | 44,44% | 42,86% | ***42,65%*** |

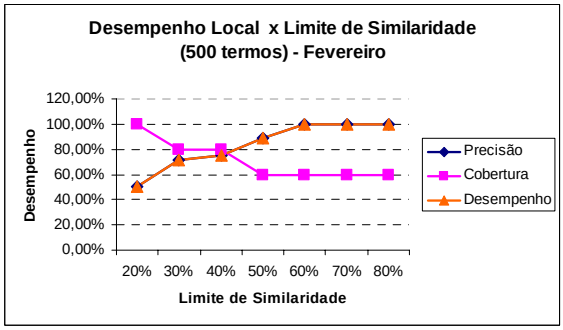
### Desempenho

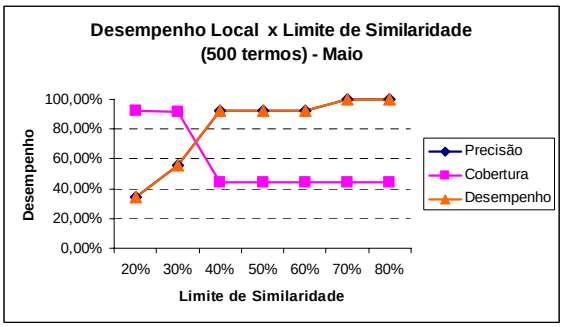
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 57,86% | 60,00% | 38,00% | 42,00% | 49,26% | 59,00% | ***51,02%*** |
| **30%** | 72,92% | 75,00% | 61,00% | 52,00% | 75,00% | 65,00% | ***66,82%*** |
| **40%** | 93,33% | 83,00% | 83,33% | 64,80% | 90,00% | 79,00% | ***82,24%*** |
| **50%** | 100,00% | 88,88% | 83,33% | 83,33% | 100,00% | 87,50% | ***90,51%*** |
| **60%** | 100,00% | 100,00% | 83,33% | 79,16% | 100,00% | 100,00% | ***93,75%*** |
| **70%** | 100,00% | 100,00% | 83,33% | 79,16% | 100,00% | 100,00% | ***93,75%*** |
| **80%** | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | ***100,00%*** |

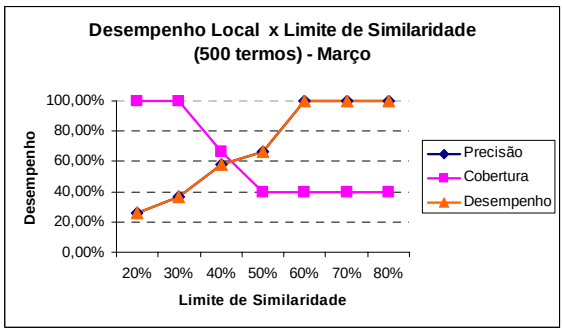
## Desempenho Local Total - Vetores de 500 termos

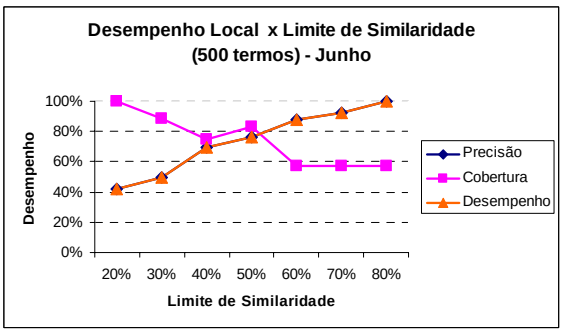


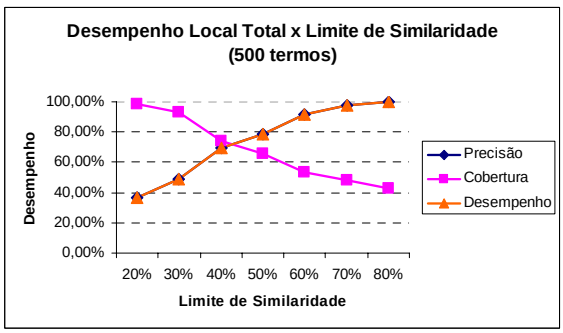



## Desempenho Local Total - Vetores de 500 termos

### Precisão

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 41,98% | 50,00% | 25,93% | 24,63% | 34,00% | 41,98% | ***36,42%*** |
| **30%** | 49,63% | 71,00% | 36,67% | 32,29% | 56,00% | 49,63% | ***49,20%*** |
| **40%** | 69,44% | 75,00% | 58,00% | 54,60% | 92,00% | 69,44% | ***69,75%*** |
| **50%** | 76,40% | 88,88% | 66,66% | 71,00% | 92,00% | 76,40% | ***78,56%*** |
| **60%** | 87,50% | 100,00% | 100,00% | 83,33% | 92,00% | 87,50% | ***91,72%*** |
| **70%** | 92,00% | 100,00% | 100,00% | 100,00% | 100,00% | 92,00% | ***97,33%*** |
| **80%** | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | ***100,00%*** |

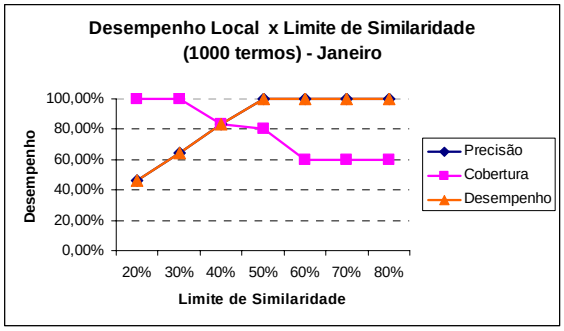
### Cobertura

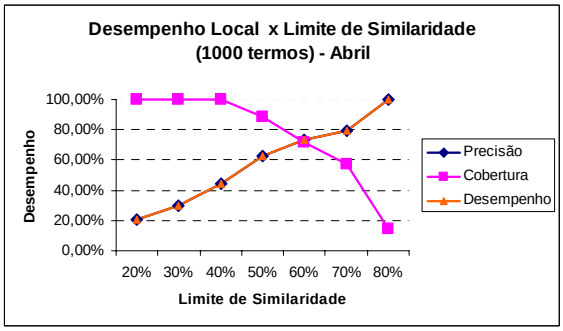
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 100,00% | 100,00% | 100,00% | 100,00% | 92,31% | 100,00% | ***98,72%*** |
| **30%** | 100,00% | 80,00% | 100,00% | 100,00% | 91,67% | 88,89% | ***93,43%*** |
| **40%** | 100,00% | 83,33% | 40,00% | 100,00% | 44,44% | 75,00% | ***73,80%*** |
| **50%** | 80,00% | 60,00% | 40,00% | 87,50% | 44,44% | 83,33% | ***65,88%*** |
| **60%** | 60,00% | 60,00% | 40,00% | 57,14% | 44,44% | 57,14% | ***53,12%*** |
| **70%** | 60,00% | 60,00% | 40,00% | 28,57% | 44,44% | 57,14% | ***48,36%*** |
| **80%** | 40,00% | 60,00% | 40,00% | 14,29% | 44,44% | 57,14% | ***42,65%*** |

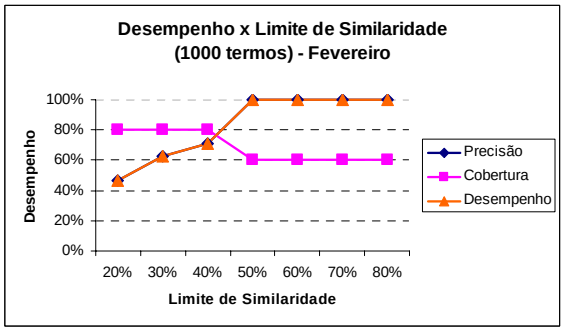
### Desempenho

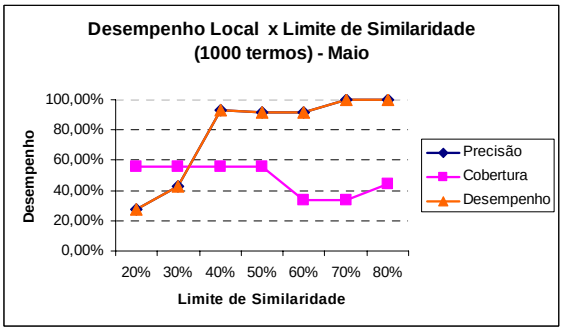
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 41,98% | 50,00% | 25,93% | 24,63% | 34,00% | 41,98% | ***36,42%*** |
| **30%** | 49,63% | 71,00% | 36,67% | 32,29% | 56,00% | 49,63% | ***49,20%*** |
| **40%** | 69,44% | 75,00% | 58,00% | 54,60% | 92,00% | 69,44% | ***69,75%*** |
| **50%** | 76,40% | 88,88% | 66,66% | 71,00% | 92,00% | 76,40% | ***78,56%*** |
| **60%** | 87,50% | 100,00% | 100,00% | 83,33% | 92,00% | 87,50% | ***91,72%*** |
| **70%** | 92,00% | 100,00% | 100,00% | 100,00% | 100,00% | 92,00% | ***97,33%*** |
| **80%** | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | ***100,00%*** |

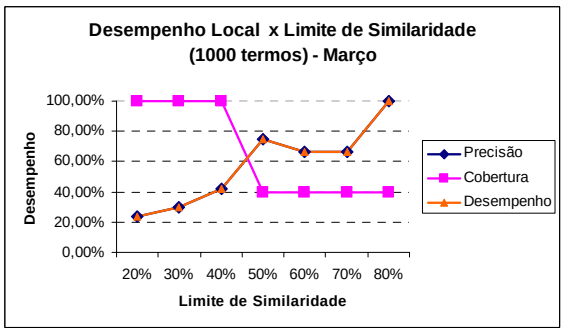
## Vetores de 1.000 termos

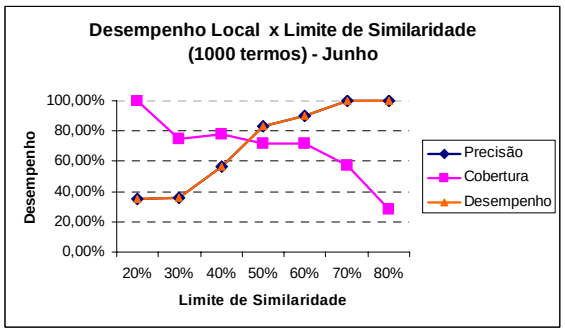


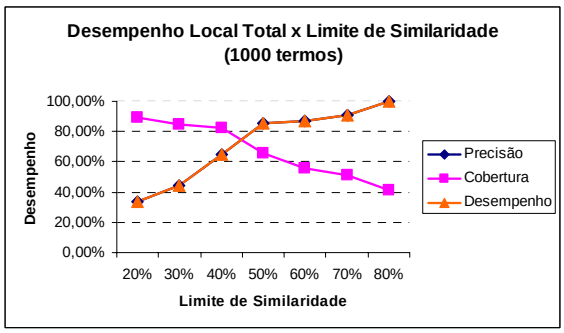



## Desempenho Local Total - Vetores de 1.000 termos

### Precisão

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 46,19% | 46,67% | 24,00% | 20,83% | 27,62% | 35,43% | ***33,46%*** |
| **30%** | 64,68% | 62,50% | 30,00% | 30,00% | 42,43% | 35,55% | ***44,19%*** |
| **40%** | 83,33% | 70,83% | 42,22% | 43,97% | 93,33% | 56,19% | ***64,98%*** |
| **50%** | 100,00% | 100,00% | 75,00% | 62,50% | 91,66% | 83,33% | ***85,42%*** |
| **60%** | 100,00% | 100,00% | 66,67% | 73,33% | 91,66% | 90,00% | ***86,94%*** |
| **70%** | 100,00% | 100,00% | 66,67% | 79,17% | 100,00% | 100,00% | ***90,97%*** |
| **80%** | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | ***100,00%*** |

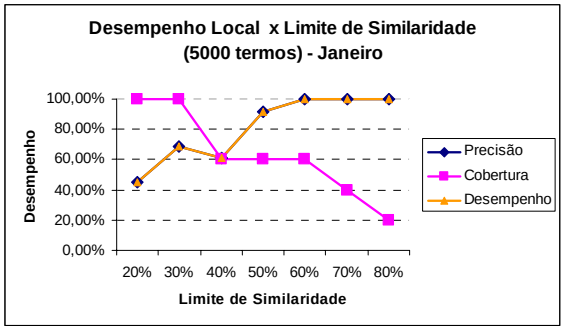
### Cobertura

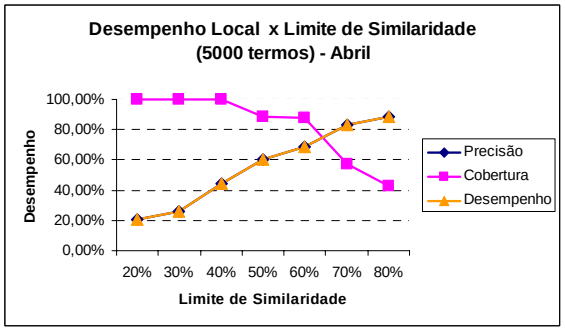
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 100,00% | 80,00% | 100,00% | 100,00% | 55,56% | 100,00% | ***89,26%*** |
| **30%** | 100,00% | 80,00% | 100,00% | 100,00% | 55,56% | 75,00% | ***85,09%*** |
| **40%** | 83,33% | 80,00% | 100,00% | 100,00% | 55,56% | 77,78% | ***82,78%*** |
| **50%** | 80,00% | 60,00% | 40,00% | 88,89% | 55,56% | 71,43% | ***65,98%*** |
| **60%** | 60,00% | 60,00% | 40,00% | 71,43% | 33,33% | 71,43% | ***56,03%*** |
| **70%** | 60,00% | 60,00% | 40,00% | 57,14% | 33,33% | 57,14% | ***51,27%*** |
| **80%** | 60,00% | 60,00% | 40,00% | 14,28% | 44,44% | 28,57% | ***41,22%*** |

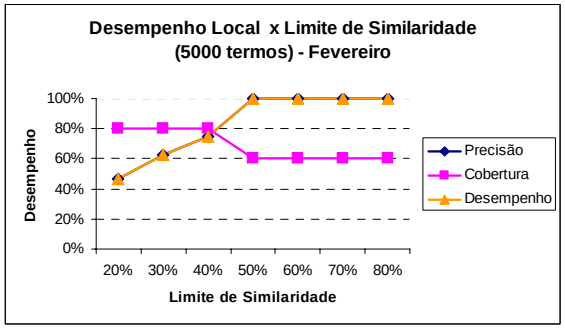
### Desempenho

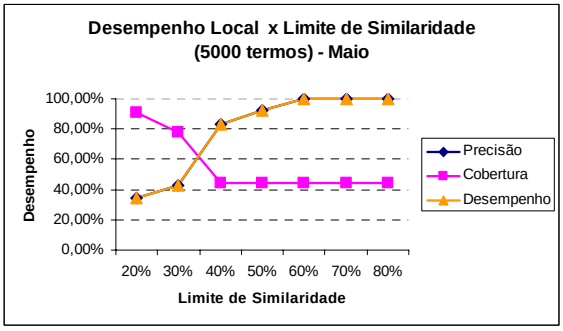
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 46,19% | 46,67% | 24,00% | 20,83% | 27,62% | 35,43% | ***33,46%*** |
| **30%** | 64,68% | 62,50% | 30,00% | 30,00% | 42,43% | 35,55% | ***44,19%*** |
| **40%** | 83,33% | 70,83% | 42,22% | 43,97% | 93,33% | 56,19% | ***64,98%*** |
| **50%** | 100,00% | 100,00% | 75,00% | 62,50% | 91,66% | 83,33% | ***85,42%*** |
| **60%** | 100,00% | 100,00% | 66,67% | 73,33% | 91,66% | 90,00% | ***86,94%*** |
| **70%** | 100,00% | 100,00% | 66,67% | 79,17% | 100,00% | 100,00% | ***90,97%*** |
| **80%** | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | ***100,00%*** |

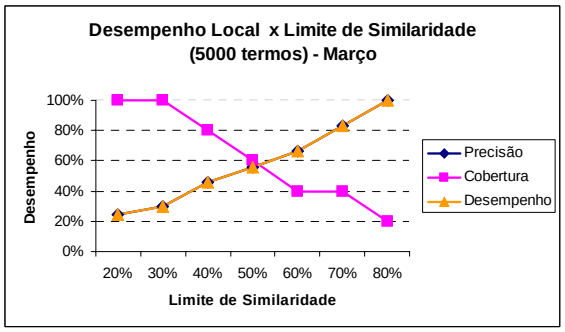
## Vetores de 5.000 termos

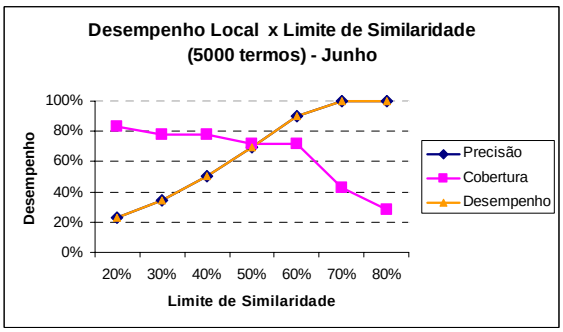


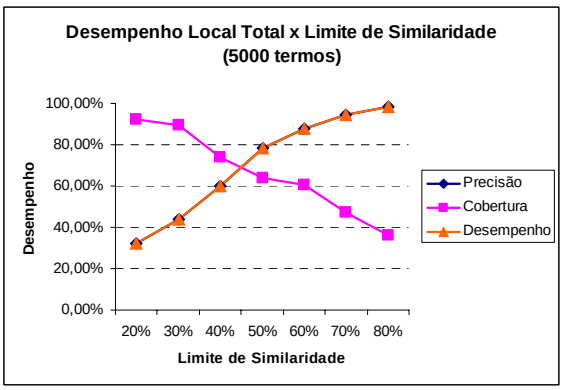



## Desempenho Local Total - Vetores de 5.000 termos

### Precisão

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 44,98% | 46,67% | 24,76% | 20,83% | 34,00% | 22,88% | ***32,35%*** |
| **30%** | 69,05% | 62,50% | 29,94% | 25,79% | 43,00% | 34,55% | ***44,14%*** |
| **40%** | 61,11% | 75,00% | 45,83% | 44,33% | 83,00% | 50,47% | ***59,96%*** |
| **50%** | 91,67% | 100,00% | 55,55% | 60,42% | 92,00% | 69,33% | ***78,16%*** |
| **60%** | 100,00% | 100,00% | 66,66% | 69,05% | 100,00% | 90,00% | ***87,62%*** |
| **70%** | 100,00% | 100,00% | 83,33% | 83,33% | 100,00% | 100,00% | ***94,44%*** |
| **80%** | 100,00% | 100,00% | 100,00% | 88,89% | 100,00% | 100,00% | ***98,15%*** |

### Cobertura

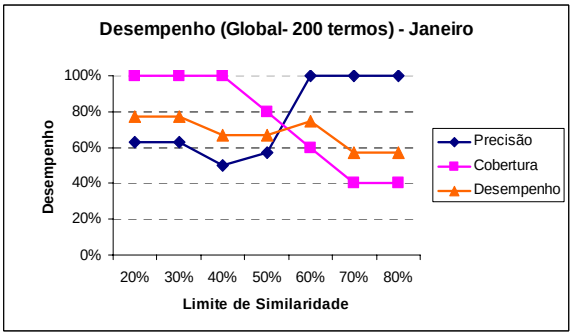
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 100,00% | 80,00% | 100,00% | 100,00% | 90,91% | 83,33% | ***92,37%*** |
| **30%** | 100,00% | 80,00% | 100,00% | 100,00% | 77,78% | 77,78% | ***89,26%*** |
| **40%** | 60,00% | 80,00% | 80,00% | 100,00% | 44,44% | 77,78% | ***73,70%*** |
| **50%** | 60,00% | 60,00% | 60,00% | 88,89% | 44,44% | 71,43% | ***64,13%*** |
| **60%** | 60,00% | 60,00% | 40,00% | 87,50% | 44,44% | 71,43% | ***60,56%*** |
| **70%** | 40,00% | 60,00% | 40,00% | 57,14% | 44,44% | 42,86% | ***47,41%*** |
| **80%** | 20,00% | 60,00% | 20,00% | 42,86% | 44,44% | 28,57% | ***35,98%*** |

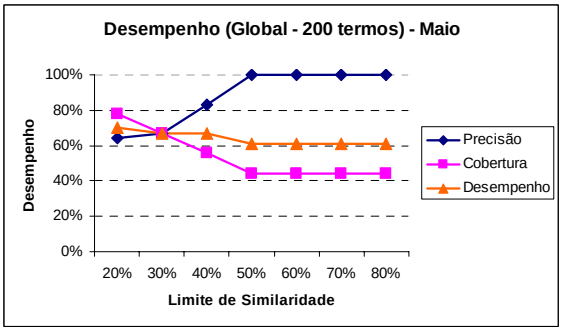
### Desempenho

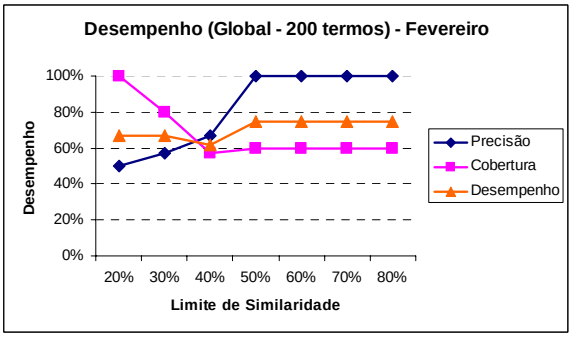
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 44,98% | 46,67% | 24,76% | 20,83% | 34,00% | 22,88% | ***32,35%*** |
| **30%** | 69,05% | 62,50% | 29,94% | 25,79% | 43,00% | 34,55% | ***44,14%*** |
| **40%** | 61,11% | 75,00% | 45,83% | 44,33% | 83,00% | 50,47% | ***59,96%*** |
| **50%** | 91,67% | 100,00% | 55,55% | 60,42% | 92,00% | 69,33% | ***78,16%*** |
| **60%** | 100,00% | 100,00% | 66,66% | 69,05% | 100,00% | 90,00% | ***87,62%*** |
| **70%** | 100,00% | 100,00% | 83,33% | 83,33% | 100,00% | 100,00% | ***94,44%*** |
| **80%** | 100,00% | 100,00% | 100,00% | 88,89% | 100,00% | 100,00% | ***98,15%*** |

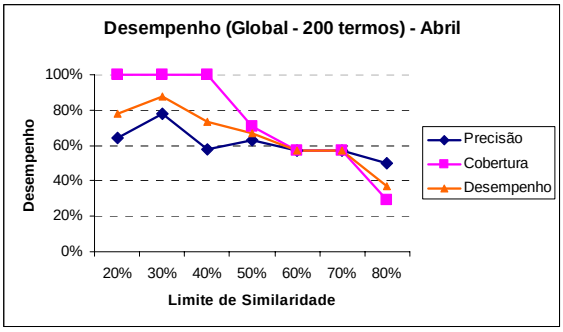
## APÊNDICE 5 – Desempenho Global Mensal: Tamanho do Vetor de Contexto x Limite de Similaridade

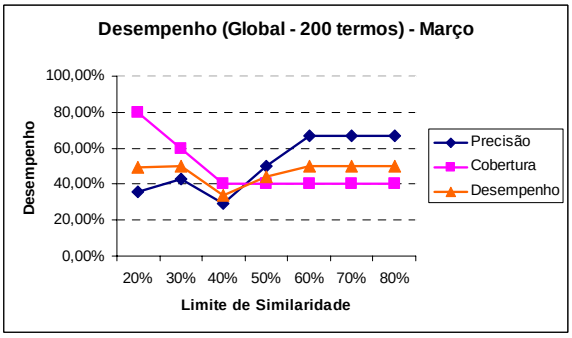
### Vetores de 200 termos

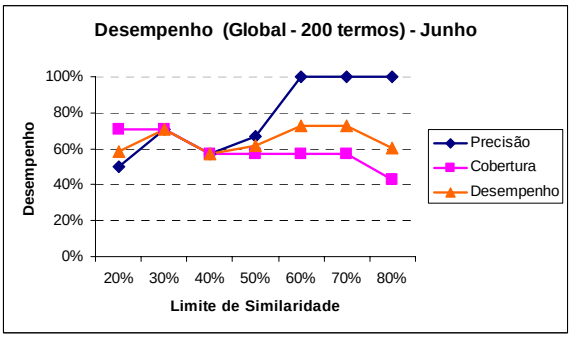


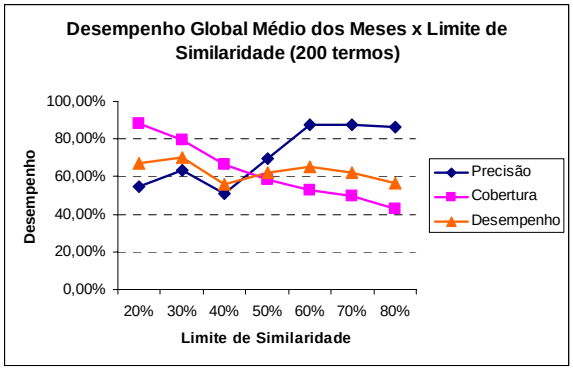



## Vetores de 200 termos

### Precisão

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 63,00% | 50,00% | 36,00% | 64,00% | 64,00% | 50,00% | ***54,50%*** |
| **30%** | 63,00% | 57,00% | 43,00% | 78,00% | 67,00% | 71,00% | ***63,17%*** |
| **40%** | 45,00% | 67,00% | 29,00% | 58,00% | 50,00% | 57,00% | ***51,00%*** |
| **50%** | 57,00% | 100,00% | 50,00% | 63,00% | 80,00% | 67,00% | ***69,50%*** |
| **60%** | 100,00% | 100,00% | 67,00% | 57,00% | 100,00% | 100,00% | ***87,33%*** |
| **70%** | 100,00% | 100,00% | 67,00% | 57,00% | 100,00% | 100,00% | ***87,33%*** |
| **80%** | 100,00% | 100,00% | 67,00% | 50,00% | 100,00% | 100,00% | ***86,17%*** |

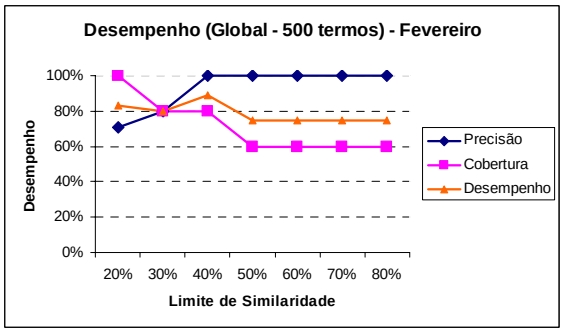
### Cobertura

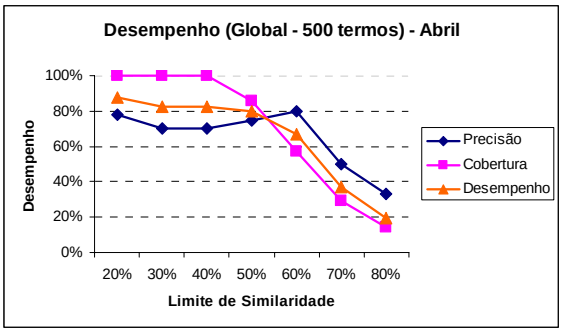
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 100,00% | 100,00% | 80,00% | 100,00% | 78,00% | 71,00% | ***88,17%*** |
| **30%** | 100,00% | 80,00% | 60,00% | 100,00% | 67,00% | 71,00% | ***79,67%*** |
| **40%** | 100,00% | 57,00% | 40,00% | 100,00% | 44,00% | 57,00% | ***66,33%*** |
| **50%** | 80,00% | 60,00% | 40,00% | 71,00% | 44,00% | 57,00% | ***58,67%*** |
| **60%** | 60,00% | 60,00% | 40,00% | 57,00% | 44,00% | 57,00% | ***53,00%*** |
| **70%** | 40,00% | 60,00% | 40,00% | 57,00% | 44,00% | 57,00% | ***49,67%*** |
| **80%** | 40,00% | 60,00% | 40,00% | 29,00% | 44,00% | 43,00% | ***42,67%*** |

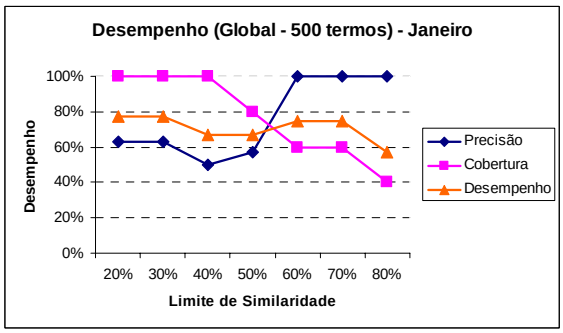
### Desempenho

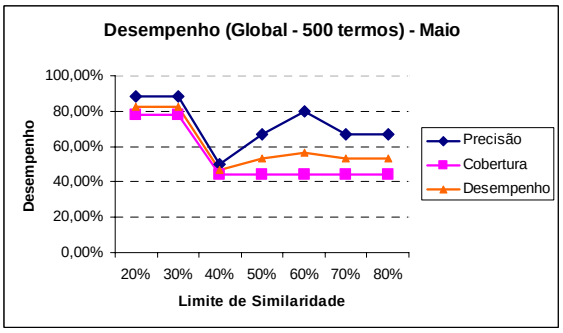
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 77,00% | 67,00% | 50,00% | 78,05% | 70,00% | 58,68% | ***66,79%*** |
| **30%** | 77,00% | 67,00% | 50,00% | 87,64% | 67,00% | 71,00% | ***69,94%*** |
| **40%** | 62,07% | 61,60% | 33,62% | 73,42% | 46,81% | 57,00% | ***55,75%*** |
| **50%** | 66,57% | 75,00% | 44,44% | 66,76% | 56,77% | 61,60% | ***61,86%*** |
| **60%** | 75,00% | 75,00% | 50,09% | 57,00% | 61,11% | 72,61% | ***65,14%*** |
| **70%** | 57,14% | 75,00% | 50,09% | 57,00% | 61,11% | 72,61% | ***62,16%*** |
| **80%** | 57,14% | 75,00% | 50,09% | 36,71% | 61,11% | 60,14% | ***56,70%*** |

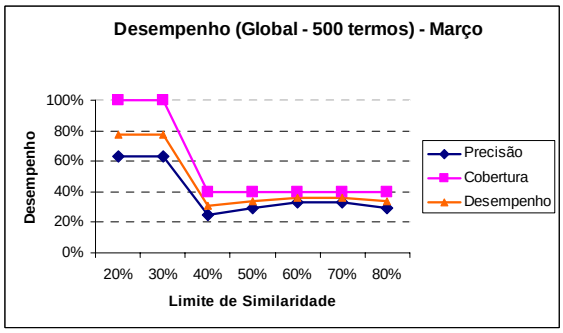
## Vetores de 500 termos

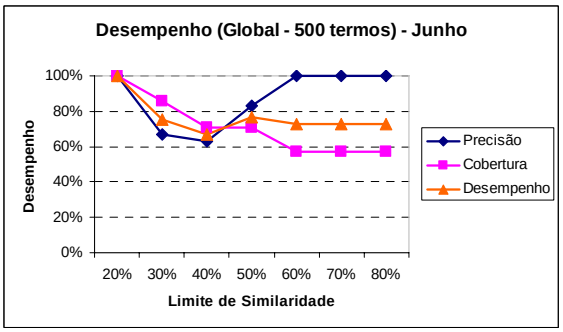


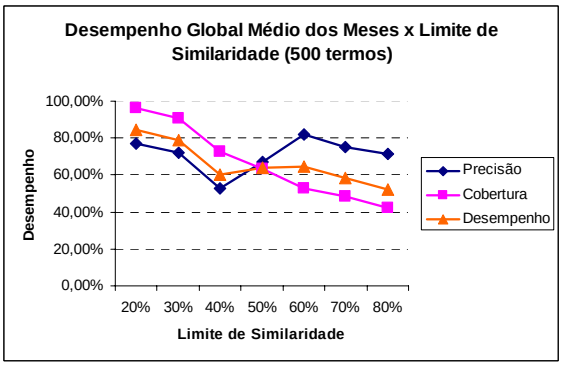



## Vetores de 500 termos

### Precisão

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 63,00% | 71,00% | 63,00% | 78,00% | 88,00% | 100,00% | ***77,17%*** |
| **30%** | 63,00% | 80,00% | 63,00% | 70,00% | 88,00% | 67,00% | ***71,83%*** |
| **40%** | 50,00% | 57,00% | 25,00% | 78,00% | 50,00% | 56,00% | ***52,67%*** |
| **50%** | 57,00% | 100,00% | 29,00% | 67,00% | 67,00% | 83,00% | ***67,17%*** |
| **60%** | 100,00% | 100,00% | 33,00% | 80,00% | 80,00% | 100,00% | ***82,17%*** |
| **70%** | 100,00% | 100,00% | 33,00% | 50,00% | 67,00% | 100,00% | ***75,00%*** |
| **80%** | 100,00% | 100,00% | 29,00% | 33,00% | 67,00% | 100,00% | ***71,50%*** |

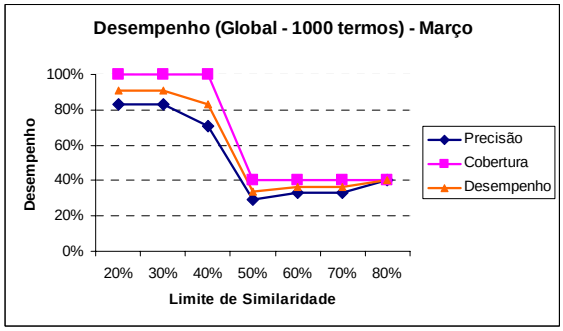
### Cobertura

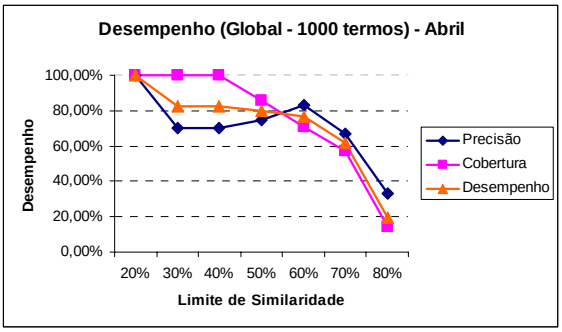
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 100,00% | 100,00% | 100,00% | 100,00% | 78,00% | 100,00% | ***96,33%*** |
| **30%** | 100,00% | 80,00% | 100,00% | 100,00% | 78,00% | 86,00% | ***90,67%*** |
| **40%** | 100,00% | 80,00% | 40,00% | 100,00% | 44,00% | 71,00% | ***72,50%*** |
| **50%** | 80,00% | 60,00% | 40,00% | 86,00% | 44,00% | 71,00% | ***63,50%*** |
| **60%** | 60,00% | 60,00% | 40,00% | 57,00% | 44,00% | 57,00% | ***53,00%*** |
| **70%** | 60,00% | 60,00% | 40,00% | 29,00% | 44,00% | 57,00% | ***48,33%*** |
| **80%** | 40,00% | 60,00% | 40,00% | 14,00% | 44,00% | 57,00% | ***42,50%*** |

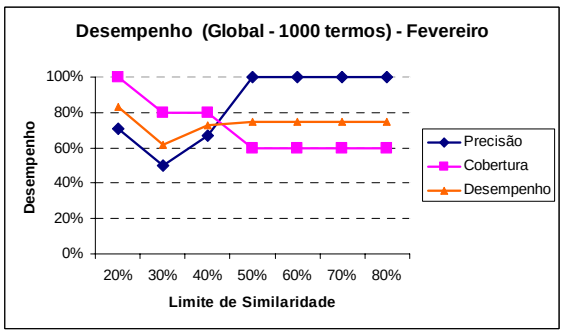
### Desempenho

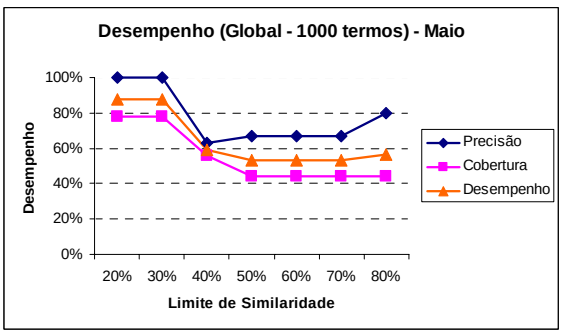
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 77,00% | 83,04% | 77,30% | 87,64% | 82,70% | 100,00% | ***84,61%*** |
| **30%** | 77,00% | 80,00% | 77,30% | 82,35% | 82,70% | 75,32% | ***79,11%*** |
| **40%** | 66,67% | 66,57% | 30,77% | 87,64% | 47,00% | 62,61% | ***60,21%*** |
| **50%** | 66,57% | 75,00% | 33,62% | 75,32% | 55,50% | 76,53% | ***63,76%*** |
| **60%** | 75,00% | 75,00% | 36,16% | 66,57% | 62,00% | 72,61% | ***64,56%*** |
| **70%** | 75,00% | 75,00% | 36,16% | 36,71% | 55,50% | 72,61% | ***58,50%*** |
| **80%** | 57,14% | 75,00% | 33,62% | 19,66% | 55,50% | 72,61% | ***52,26%*** |

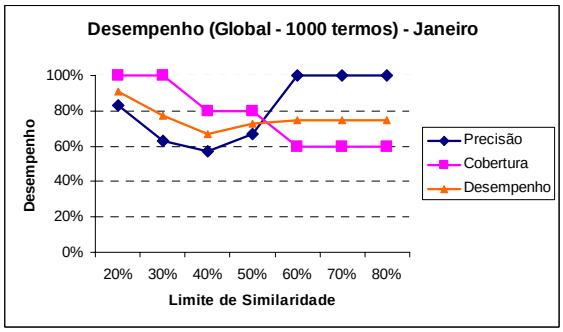
## Vetores de 1.000 termos

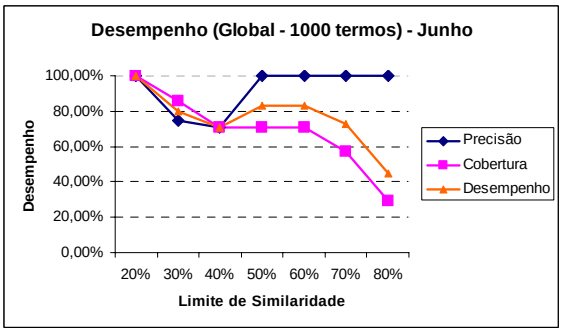


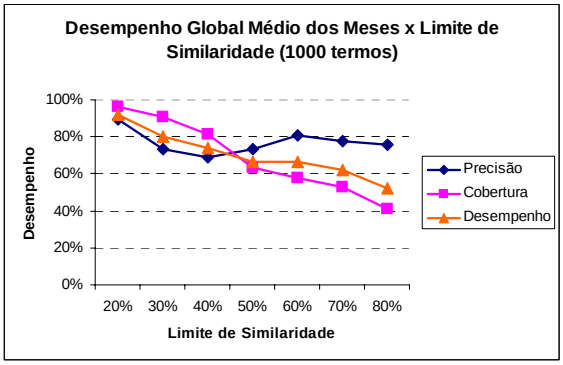



## Vetores de 1.000 termos

### Precisão

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 83,00% | 71,00% | 83,00% | 100,00% | 100,00% | 100,00% | ***89,50%*** |
| **30%** | 63,00% | 50,00% | 83,00% | 70,00% | 100,00% | 75,00% | ***73,50%*** |
| **40%** | 57,00% | 80,00% | 71,00% | 70,00% | 63,00% | 71,00% | ***68,67%*** |
| **50%** | 67,00% | 100,00% | 29,00% | 75,00% | 67,00% | 100,00% | ***73,00%*** |
| **60%** | 100,00% | 100,00% | 33,00% | 83,00% | 67,00% | 100,00% | ***80,50%*** |
| **70%** | 100,00% | 100,00% | 33,00% | 67,00% | 67,00% | 100,00% | ***77,83%*** |
| **80%** | 100,00% | 100,00% | 40,00% | 33,00% | 80,00% | 100,00% | ***75,50%*** |

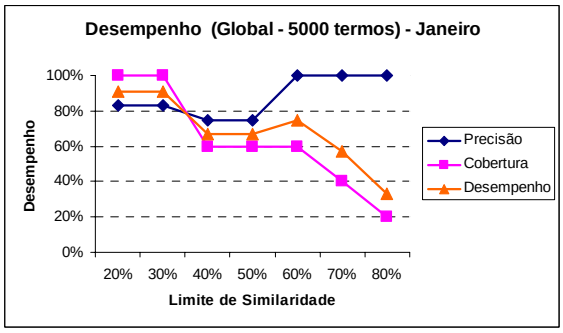
### Cobertura

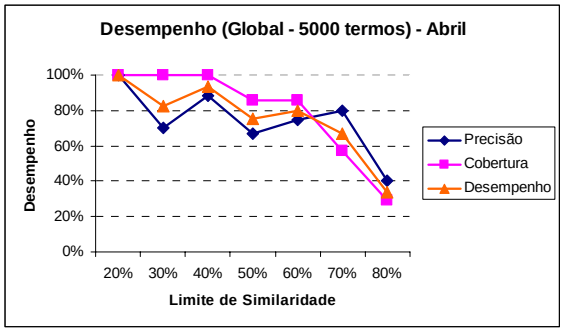
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 100,00% | 100,00% | 100,00% | 100,00% | 78,00% | 100,00% | ***96,33%*** |
| **30%** | 100,00% | 80,00% | 100,00% | 100,00% | 78,00% | 86,00% | ***90,67%*** |
| **40%** | 80,00% | 80,00% | 100,00% | 100,00% | 56,00% | 71,00% | ***81,17%*** |
| **50%** | 80,00% | 60,00% | 40,00% | 86,00% | 44,00% | 71,00% | ***63,50%*** |
| **60%** | 60,00% | 60,00% | 40,00% | 71,00% | 44,00% | 71,00% | ***57,67%*** |
| **70%** | 60,00% | 60,00% | 40,00% | 57,00% | 44,00% | 57,00% | ***53,00%*** |
| **80%** | 60,00% | 60,00% | 40,00% | 14,00% | 44,00% | 29,00% | ***41,17%*** |

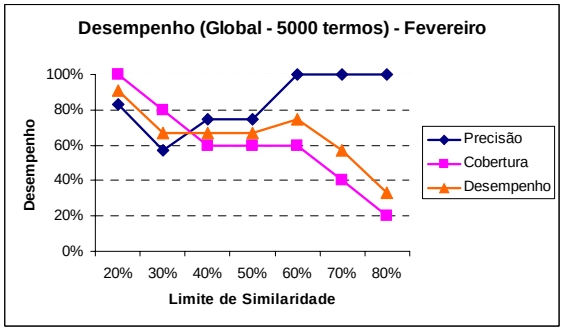
### Desempenho

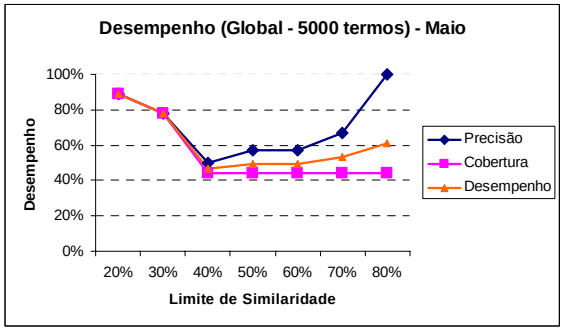
| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 90,71% | 83,04% | 90,71% | 100,00% | 87,64% | 100,00% | ***92,02%*** |
| **30%** | 77,30% | 61,54% | 90,71% | 82,35% | 87,64% | 80,12% | ***79,94%*** |
| **40%** | 66,57% | 80,00% | 83,04% | 82,00% | 59,29% | 71,00% | ***73,65%*** |
| **50%** | 72,93% | 75,00% | 33,62% | 80,00% | 53,12% | 83,04% | ***66,28%*** |
| **60%** | 75,00% | 75,00% | 36,16% | 77,00% | 53,12% | 83,04% | ***66,55%*** |
| **70%** | 75,00% | 75,00% | 36,16% | 62,00% | 53,12% | 72,61% | ***62,32%*** |
| **80%** | 75,00% | 75,00% | 40,00% | 20,00% | 56,77% | 44,96% | ***51,96%*** |

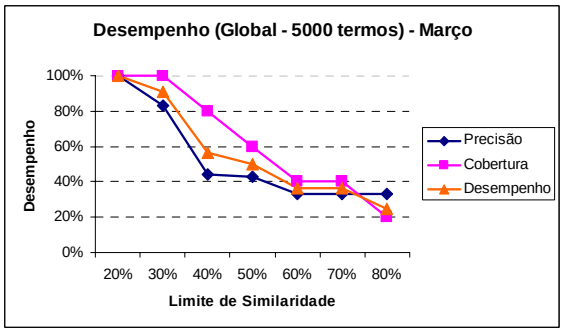
## Vetores de 5.000 termos

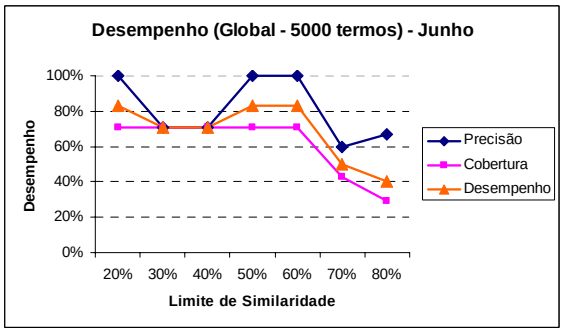


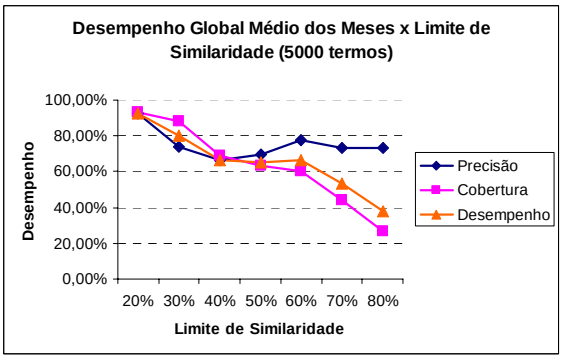



## Vetores de 5.000 termos

### Precisão

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 83,00% | 83,00% | 100,00% | 100,00% | 89,00% | 100,00% | ***92,50%*** |
| **30%** | 83,00% | 57,00% | 83,00% | 70,00% | 78,00% | 71,00% | ***73,67%*** |
| **40%** | 75,00% | 75,00% | 40,00% | 88,00% | 50,00% | 71,00% | ***66,50%*** |
| **50%** | 75,00% | 75,00% | 43,00% | 67,00% | 57,00% | 100,00% | ***69,50%*** |
| **60%** | 100,00% | 100,00% | 33,00% | 75,00% | 57,00% | 100,00% | ***77,50%*** |
| **70%** | 100,00% | 100,00% | 33,00% | 80,00% | 67,00% | 60,00% | ***73,33%*** |
| **80%** | 100,00% | 100,00% | 33,00% | 40,00% | 100,00% | 67,00% | ***73,33%*** |

### Cobertura

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 100,00% | 100,00% | 100,00% | 100,00% | 89,00% | 71,00% | ***93,33%*** |
| **30%** | 100,00% | 80,00% | 100,00% | 100,00% | 78,00% | 71,00% | ***88,17%*** |
| **40%** | 60,00% | 60,00% | 80,00% | 100,00% | 44,00% | 71,00% | ***69,17%*** |
| **50%** | 60,00% | 60,00% | 60,00% | 86,00% | 44,00% | 71,00% | ***63,50%*** |
| **60%** | 60,00% | 60,00% | 40,00% | 86,00% | 44,00% | 71,00% | ***60,17%*** |
| **70%** | 40,00% | 40,00% | 40,00% | 57,00% | 44,00% | 43,00% | ***44,00%*** |
| **80%** | 20,00% | 20,00% | 20,00% | 29,00% | 44,00% | 29,00% | ***27,00%*** |

### Desempenho

| | janeiro | fevereiro | março | abril | maio | junho | Média |
|---|---|---|---|---|---|---|---|
| **20%** | 90,71% | 90,71% | 100,00% | 100,00% | 89,00% | 83,04% | ***92,24%*** |
| **30%** | 90,71% | 66,57% | 90,71% | 82,35% | 78,00% | 71,00% | ***79,89%*** |
| **40%** | 66,67% | 66,67% | 53,33% | 93,62% | 46,81% | 71,00% | ***66,35%*** |
| **50%** | 66,67% | 66,67% | 50,10% | 75,32% | 49,66% | 83,04% | ***65,24%*** |
| **60%** | 75,00% | 75,00% | 36,16% | 80,12% | 49,66% | 83,04% | ***66,50%*** |
| **70%** | 57,14% | 57,14% | 36,16% | 66,57% | 53,12% | 50,10% | ***53,37%*** |
| **80%** | 33,33% | 33,33% | 24,91% | 33,62% | 61,11% | 40,48% | ***37,80%*** |

## APÊNDICE 6 – Codificação das principais funções implementadas no algoritmo

**//Calculo do TF/IDF**

```
import org.apache.lucene.document.Document;
import org.apache.lucene.search.IndexSearcher;
import org.apache.lucene.search.Query;
import org.apache.lucene.search.Hits;
import org.apache.lucene.store.FSDirectory;
import org.apache.lucene.store.Directory;
import org.apache.lucene.index.FilterIndexReader.FilterTermPositions;
import org.apache.lucene.index.IndexReader;
import org.apache.lucene.queryParser.QueryParser;
import org.apache.lucene.analysis.standard.StandardAnalyzer;
import org.apache.lucene.index.IndexWriter;
import org.apache.lucene.document.Field;
import java.util.*;
import java.io.*;
import java.text.*;
import java.util.*;
import org.apache.lucene.index.*;

/**
  Eliminar as stopwords e calcular  o TF/IDF
**/

public class Lexicon {
    public IndexReader idxReader;

    public Collection calculateTfidf(String path) throws Exception {

        ArrayList< TermoTfidf > termList = new ArrayList< TermoTfidf >();
        TermoComparator comp= new TermoComparator();
        idxReader = null;
        TermEnum termEnum = null;
        try {
            idxReader = IndexReader.open(path); //caminho do índice
            int nd = idxReader.numDocs(), df = 0, tf = 0, tfa = 0;
            termEnum = idxReader.terms();

            double tfidf = 0.0;


            Term term = null;
            while (termEnum.next()) {

                TermoTfidf TTfidf = new TermoTfidf();

                tfidf = 0.0; tfa = 0;
                term = termEnum.term();
                double tfidf_maior = 0.0;
                double tfidf_medio = 0.0;

                if ((!isStopWord(term.text().toUpperCase())))  {
                     df = idxReader.docFreq(term);
                     FilterTermPositions ftp = new
FilterTermPositions(idxReader.termPositions(term));
                    while (ftp.next()) {
                        tf = ftp.freq();
```

```
                        tfa += tf;

                        tfidf += tf*(Math.log10(nd/(double)df));
                    }

                    tfidf_medio = (tfidf/nd);
                    TTfidf.setTermo(term.text());
                    TTfidf.setTfidf_termo(tfidf_medio);
                    if (TTfidf.getTfidf_termo() > 0.1)
                    {
                       termList.add(TTfidf);
                    }

                   Collections.sort(termList, comp);
                }
            }

        } finally {
            if (termEnum != null) termEnum.close();
            if (idxReader != null) idxReader.close();
        }

        return termList; //termList é a lista de termos
   }
```

**//Remocao de stop words**

```
public static HashMap loadStopWords() throws Exception {
      HashMap stopWord = new HashMap();
      stopWord.put("a", "");
      stopWord.put("à", "");
      stopWord.put("ab", "");
      stopWord.put("abaixo", "");
      stopWord.put("about", "");
      stopWord.put("above", "");
      stopWord.put("abroad", "");
      stopWord.put("acaso", "");
      stopWord.put("acerca", "");
      stopWord.put("acima", "");
      stopWord.put("acha", "");
      stopWord.put("acho", "");
      stopWord.put("acola", "");
      stopWord.put("across", "");
      stopWord.put("ademais", "");
      stopWord.put("adentro", "");
      stopWord.put("adiante", "");
      stopWord.put("afim", "");
      stopWord.put("afin", "");
      stopWord.put("afins", "");
      stopWord.put("afora", "");
      stopWord.put("after", "");
      stopWord.put("against", "");
      stopWord.put("agora", "");
      stopWord.put("ahead", "");
      stopWord.put("ai", "");
      stopWord.put("ailleurs", "");
      stopWord.put("aind","");
      stopWord.put("ainda", "");
      stopWord.put("aiansi", "");
      stopWord.put("alem", "");
      stopWord.put("algo", "");
```

```
stopWord.put("alguem", "");
stopWord.put("algum", "");
stopWord.put("alguma", "");
stopWord.put("algumas", "");
stopWord.put("algunas", "");
stopWord.put("algunos", "");
stopWord.put("alguns", "");
stopWord.put("ali", "");
stopWord.put("alias", "");
stopWord.put("all", "");
stopWord.put("alli", "");
stopWord.put("alo", "");
stopWord.put("alone", "");
stopWord.put("along", "");
stopWord.put("alors", "");
stopWord.put("also", "");
stopWord.put("although", "");
stopWord.put("always", "");
stopWord.put("ambas", "");
stopWord.put("ambito", "");
stopWord.put("ambos", "");
stopWord.put("among", "");
stopWord.put("amongst", "");
stopWord.put("amplo", "");
stopWord.put("an", "");
stopWord.put("ano", "");
stopWord.put("and", "");
stopWord.put("anex", "");
stopWord.put("another", "");
stopWord.put("ante", "");
stopWord.put("antes", "");
stopWord.put("anti", "");
stopWord.put("any", "");
stopWord.put("ao", "");
stopWord.put("aonde", "");
stopWord.put("aos", "");
stopWord.put("apenas", "");
stopWord.put("apesar", "");
stopWord.put("apos", "");
stopWord.put("apres", "");
stopWord.put("apresent", "");
stopWord.put("aquela", "");
stopWord.put("aquelas", "");
stopWord.put("aquele", "");
stopWord.put("aqueles", "");
stopWord.put("aqui", "");
stopWord.put("aquilo", "");
stopWord.put("are", "");
stopWord.put("as", "");
stopWord.put("assim", "");
stopWord.put("at", "");
stopWord.put("ate", "");
stopWord.put("ati", "");
stopWord.put("atras", "");
stopWord.put("atraves", "");
stopWord.put("attendant", "");
stopWord.put("atual", "");
stopWord.put("aucun", "");
stopWord.put("aucune", "");
stopWord.put("auquel", "");
stopWord.put("aussi", "");
```

```
stopWord.put("aussitot", "");
stopWord.put("autant", "");
stopWord.put("autour", "");
stopWord.put("auxquelles", "");
stopWord.put("auxquels", "");
stopWord.put("avante", "");
stopWord.put("avec", "");
stopWord.put("b", "");
stopWord.put("back", "");
stopWord.put("baixissima", "");
stopWord.put("baixissimas", "");
stopWord.put("baixissimo", "");
stopWord.put("baixissimos", "");
stopWord.put("balela", "");
stopWord.put("barato", "");
stopWord.put("baseada", "");
stopWord.put("basicamente", "");
stopWord.put("basta", "");
stopWord.put("bastante", "");
stopWord.put("be", "");
stopWord.put("beaucoup", "");
stopWord.put("because", "");
stopWord.put("been", "");
stopWord.put("before", "");
stopWord.put("bem", "");
stopWord.put("best", "");
stopWord.put("better", "");
stopWord.put("between", "");
stopWord.put("beyond", "");
stopWord.put("big", "");
stopWord.put("both", "");
stopWord.put("but", "");
stopWord.put("by", "");
stopWord.put("c", "");
stopWord.put("ca", "");
stopWord.put("cada", "");
stopWord.put("capaz", "");
stopWord.put("capazes", "");
stopWord.put("caro", "");
stopWord.put("caros", "");
stopWord.put("caso", "");
stopWord.put("casos", "");
stopWord.put("cd", "");
stopWord.put("cela", "");
stopWord.put("celle", "");
stopWord.put("celles", "");
stopWord.put("celui", "");
stopWord.put("cependant", "");
stopWord.put("cerca", "");
stopWord.put("certa", "");
stopWord.put("certain", "");
stopWord.put("certaine", "");
stopWord.put("certaines", "");
stopWord.put("certains", "");
stopWord.put("certas", "");
stopWord.put("certo", "");
stopWord.put("certos", "");
stopWord.put("ces", "");
stopWord.put("cet", "");
stopWord.put("cette", "");
stopWord.put("ceux", "");
```

```
stopWord.put("chacun", "");
stopWord.put("chacune", "");
stopWord.put("chang", "");
stopWord.put("chaque", "");
stopWord.put("chez", "");
stopWord.put("chose", "");
stopWord.put("cima", "");
stopWord.put("coisa", "");
stopWord.put("com", "");
stopWord.put("combien", "");
stopWord.put("comigo", "");
stopWord.put("comme", "");
stopWord.put("comment", "");
stopWord.put("como", "");
stopWord.put("comuns", "");
stopWord.put("concernant", "");
stopWord.put("conforme", "");
stopWord.put("conformidade", "");
stopWord.put("conosco", "");
stopWord.put("conseguinte", "");
stopWord.put("consequentemente", "");
stopWord.put("consigo", "");
stopWord.put("contexto", "");
stopWord.put("contida", "");
stopWord.put("contidas", "");
stopWord.put("contido", "");
stopWord.put("contidos", "");
stopWord.put("contigo", "");
stopWord.put("contra", "");
stopWord.put("contre", "");
stopWord.put("contudo", "");
stopWord.put("convosco", "");
stopWord.put("corp", "");
stopWord.put("could", "");
stopWord.put("cuando", "");
stopWord.put("cuja", "");
stopWord.put("cujas", "");
stopWord.put("cujo", "");
stopWord.put("cujos", "");
stopWord.put("d", "");
stopWord.put("da", "");
stopWord.put("dans", "");
stopWord.put("daquela", "");
stopWord.put("daquelas", "");
stopWord.put("daquele", "");
stopWord.put("daqueles", "");
stopWord.put("daqui", "");
stopWord.put("daquilodas", "");
stopWord.put("das", "");
stopWord.put("de", "");
stopWord.put("debaixo", "");
stopWord.put("dedans", "");
stopWord.put("defin","");
stopWord.put("defronte", "");
stopWord.put("dehors", "");
stopWord.put("deix", "");
stopWord.put("deja", "");
stopWord.put("del", "");
stopWord.put("dela", "");
stopWord.put("delas", "");
stopWord.put("dele", "");
```

```
stopWord.put("deles", "");
stopWord.put("della", "");
stopWord.put("delle", "");
stopWord.put("demais", "");
stopWord.put("dentre", "");
stopWord.put("depois", "");
stopWord.put("depressa", "");
stopWord.put("depuis", "");
stopWord.put("des", "");
stopWord.put("desde", "");
stopWord.put("desquelles", "");
stopWord.put("desquels", "");
stopWord.put("dess", "");
stopWord.put("dessa", "");
stopWord.put("dessas", "");
stopWord.put("desse", "");
stopWord.put("desses", "");
stopWord.put("dessus", "");
stopWord.put("desta", "");
stopWord.put("destas", "");
stopWord.put("deste", "");
stopWord.put("destes", "");
stopWord.put("detras", "");
stopWord.put("dev", "");
stopWord.put("dever", "");
stopWord.put("diante", "");
stopWord.put("disso", "");
stopWord.put("disto", "");
stopWord.put("do", "");
stopWord.put("donc", "");
stopWord.put("donde", "");
stopWord.put("donne", "");
stopWord.put("dont", "");
stopWord.put("doutra", "");
stopWord.put("doutras", "");
stopWord.put("doutro", "");
stopWord.put("doutros", "");
stopWord.put("dos", "");
stopWord.put("du", "");
stopWord.put("dum", "");
stopWord.put("duma", "");
stopWord.put("duns", "");
stopWord.put("duquel", "");
stopWord.put("duramente", "");
stopWord.put("durant", "");
stopWord.put("durante", "");
stopWord.put("during", "");
stopWord.put("e", "");
stopWord.put("each", "");
stopWord.put("easy", "");
stopWord.put("either", "");
stopWord.put("el", "");
stopWord.put("ela", "");
stopWord.put("elas", "");
stopWord.put("ele", "");
stopWord.put("eles", "");
stopWord.put("ell", "");
stopWord.put("ella", "");
stopWord.put("elle", "");
stopWord.put("elles", "");
stopWord.put("ello", "");
```

```
stopWord.put("em", "");
stopWord.put("embaixo", "");
stopWord.put("embora", "");
stopWord.put("en", "");
stopWord.put("encore", "");
stopWord.put("enough", "");
stopWord.put("enquanto", "");
stopWord.put("entanto", "");
stopWord.put("entao", "");
stopWord.put("então", "");
stopWord.put("entr", "");
stopWord.put("entre", "");
stopWord.put("entretanto", "");
stopWord.put("era", "");
stopWord.put("ess", "");
stopWord.put("essa", "");
stopWord.put("essas", "");
stopWord.put("esse", "");
stopWord.put("esses", "");
stopWord.put("est", "");
stopWord.put("esta", "");
stopWord.put("está", "");
stopWord.put("estar", "");
stopWord.put("estas", "");
stopWord.put("este", "");
stopWord.put("estes", "");
stopWord.put("et", "");
stopWord.put("etaient", "");
stopWord.put("etait", "");
stopWord.put("etant", "");
stopWord.put("etc", "");
stopWord.put("eu", "");
stopWord.put("eux", "");
stopWord.put("ever", "");
stopWord.put("every", "");
stopWord.put("everybody", "");
stopWord.put("everyone", "");
stopWord.put("everything", "");
stopWord.put("everywhere", "");
stopWord.put("exemp", "");
stopWord.put("exempl", "");
stopWord.put("exist", "");
stopWord.put("f", "");
stopWord.put("facil", "");
stopWord.put("favor", "");
stopWord.put("faz", "");
stopWord.put("few", "");
stopWord.put("fim", "");
stopWord.put("fin", "");
stopWord.put("finali", "");
stopWord.put("fins", "");
stopWord.put("foi", "");
stopWord.put("for", "");
stopWord.put("fora", "");
stopWord.put("forward", "");
stopWord.put("from", "");
stopWord.put("fulano", "");
stopWord.put("furent", "");
stopWord.put("g", "");
stopWord.put("grace", "");
stopWord.put("grand", "");
```

```
stopWord.put("grandemente", "");
stopWord.put("h", "");
stopWord.put("had", "");
stopWord.put("has", "");
stopWord.put("hav", "");
stopWord.put("have", "");
stopWord.put("he", "");
stopWord.put("hem", "");
stopWord.put("her", "");
stopWord.put("here", "");
stopWord.put("hers", "");
stopWord.put("herself", "");
stopWord.put("him", "");
stopWord.put("himself", "");
stopWord.put("his", "");
stopWord.put("hoje", "");
stopWord.put("hormis", "");
stopWord.put("hors", "");
stopWord.put("how", "");
stopWord.put("however", "");
stopWord.put("hoy", "");
stopWord.put("hum", "");
stopWord.put("i", "");
stopWord.put("ici", "");
stopWord.put("if", "");
stopWord.put("igual", "");
stopWord.put("igualmente", "");
stopWord.put("il", "");
stopWord.put("ils", "");
stopWord.put("importânc", "");
stopWord.put("in", "");
stopWord.put("inc", "");
stopWord.put("inter", "");
stopWord.put("into", "");
stopWord.put("is", "");
stopWord.put("isso", "");
stopWord.put("isto", "");
stopWord.put("it", "");
stopWord.put("item", "");
stopWord.put("its", "");
stopWord.put("itself", "");
stopWord.put("iusque", "");
stopWord.put("j", "");
stopWord.put("ja", "");
stopWord.put("já", "");
stopWord.put("jadis", "");
stopWord.put("jamais", "");
stopWord.put("je", "");
stopWord.put("jeito", "");
stopWord.put("juntamente", "");
stopWord.put("junto", "");
stopWord.put("juntos", "");
stopWord.put("jusqu", "");
stopWord.put("jusque", "");
stopWord.put("just", "");
stopWord.put("justa", "");
stopWord.put("justo", "");
stopWord.put("k", "");
stopWord.put("kg", "");
stopWord.put("l", "");
stopWord.put("la", "");
```

```
stopWord.put("lado", "");
stopWord.put("lados", "");
stopWord.put("laquelle", "");
stopWord.put("las", "");
stopWord.put("last", "");
stopWord.put("le", "");
stopWord.put("lequel", "");
stopWord.put("les", "");
stopWord.put("lesquelles", "");
stopWord.put("lesquels", "");
stopWord.put("less", "");
stopWord.put("leur", "");
stopWord.put("leurs", "");
stopWord.put("lha", "");
stopWord.put("lhas", "");
stopWord.put("lhe", "");
stopWord.put("lhes", "");
stopWord.put("lho", "");
stopWord.put("lhos", "");
stopWord.put("lo", "");
stopWord.put("lors", "");
stopWord.put("lorsque", "");
stopWord.put("los", "");
stopWord.put("ltda", "");
stopWord.put("lui", "");
stopWord.put("m", "");
stopWord.put("maior", "");
stopWord.put("mais", "");
stopWord.put("malgre", "");
stopWord.put("maneira", "");
stopWord.put("maneiras", "");
stopWord.put("many", "");
stopWord.put("mas", "");
stopWord.put("me", "");
stopWord.put("mean","");
stopWord.put("médi", "");
stopWord.put("mediante", "");
stopWord.put("melhor", "");
stopWord.put("meme", "");
stopWord.put("memes", "");
stopWord.put("menos", "");
stopWord.put("mera", "");
stopWord.put("meramente", "");
stopWord.put("meras", "");
stopWord.put("mero", "");
stopWord.put("meros", "");
stopWord.put("mes", "");
stopWord.put("mesm", "");
stopWord.put("mesma", "");
stopWord.put("mesmas", "");
stopWord.put("mesmo", "");
stopWord.put("mesmos", "");
stopWord.put("meu", "");
stopWord.put("meus", "");
stopWord.put("mien", "");
stopWord.put("mienne", "");
stopWord.put("miennes", "");
stopWord.put("miens", "");
stopWord.put("mim", "");
stopWord.put("minha", "");
stopWord.put("minhas", "");
```

```
stopWord.put("moins", "");
stopWord.put("moment", "");
stopWord.put("mon", "");
stopWord.put("more", "");
stopWord.put("most", "");
stopWord.put("moyennant", "");
stopWord.put("mr", "");
stopWord.put("mrs", "");
stopWord.put("my", "");
stopWord.put("myself", "");
stopWord.put("mz", "");
stopWord.put("muit", "");
stopWord.put("muita", "");
stopWord.put("muitas", "");
stopWord.put("muito", "");
stopWord.put("muitos", "");
stopWord.put("n", "");
stopWord.put("na", "");
stopWord.put("nada", "");
stopWord.put("nao", "");
stopWord.put("naquela", "");
stopWord.put("naquelas", "");
stopWord.put("naquele", "");
stopWord.put("naqueles", "");
stopWord.put("naquilo", "");
stopWord.put("nas", "");
stopWord.put("ne", "");
stopWord.put("near", "");
stopWord.put("necess", "");
stopWord.put("neither", "");
stopWord.put("nela", "");
stopWord.put("nelas", "");
stopWord.put("nele", "");
stopWord.put("neles", "");
stopWord.put("nenhum", "");
stopWord.put("nenhuma", "");
stopWord.put("nenhumas", "");
stopWord.put("nenhuns", "");
stopWord.put("nesse", "");
stopWord.put("nesses", "");
stopWord.put("nest", "");
stopWord.put("nesta", "");
stopWord.put("nestas", "");
stopWord.put("neste", "");
stopWord.put("nestes", "");
stopWord.put("never", "");
stopWord.put("ni", "");
stopWord.put("ninguem", "");
stopWord.put("nisso", "");
stopWord.put("no", "");
stopWord.put("nom","");
stopWord.put("non", "");
stopWord.put("none", "");
stopWord.put("nos", "");
stopWord.put("nossa", "");
stopWord.put("nossas", "");
stopWord.put("nosso", "");
stopWord.put("nossos", "");
stopWord.put("not", "");
stopWord.put("notamment", "");
stopWord.put("notres", "");
```

```
stopWord.put("nous", "");
stopWord.put("now", "");
stopWord.put("nulle", "");
stopWord.put("nulles", "");
stopWord.put("num", "");
stopWord.put("numa", "");
stopWord.put("numas", "");
stopWord.put("nunca", "");
stopWord.put("nuns", "");
stopWord.put("o", "");
stopWord.put("of", "");
stopWord.put("often", "");
stopWord.put("oi", "");
stopWord.put("ola", "");
stopWord.put("on", "");
stopWord.put("onde", "");
stopWord.put("one", "");
stopWord.put("only", "");
stopWord.put("or", "");
stopWord.put("os", "");
stopWord.put("other", "");
stopWord.put("ou", "");
stopWord.put("our", "");
stopWord.put("ourselves", "");
stopWord.put("out", "");
stopWord.put("outr", "");
stopWord.put("outra", "");
stopWord.put("outras", "");
stopWord.put("outrem", "");
stopWord.put("outro", "");
stopWord.put("outros", "");
stopWord.put("over", "");
stopWord.put("p", "");
stopWord.put("págin","");
stopWord.put("par", "");
stopWord.put("para", "");
stopWord.put("parce", "");
stopWord.put("parmi", "");
stopWord.put("part","");
stopWord.put("paul", "");
stopWord.put("pel", "");
stopWord.put("pela", "");
stopWord.put("pelas", "");
stopWord.put("per", "");
stopWord.put("perante", "");
stopWord.put("perto", "");
stopWord.put("plus", "");
stopWord.put("plusieurs", "");
stopWord.put("pod", "");
stopWord.put("pode", "");
stopWord.put("pois", "");
stopWord.put("por", "");
stopWord.put("possui", "");
stopWord.put("poucos", "");
stopWord.put("pour", "");
stopWord.put("pourquoi", "");
stopWord.put("pr", "");
stopWord.put("pres", "");
stopWord.put("proxima", "");
stopWord.put("proximamente", "");
stopWord.put("proximas", "");
```

```
stopWord.put("proximos", "");
stopWord.put("psiu", "");
stopWord.put("puis", "");
stopWord.put("puisque", "");
stopWord.put("q", "");
stopWord.put("quais", "");
stopWord.put("quaisquer", "");
stopWord.put("qual", "");
stopWord.put("qualquer", "");
stopWord.put("quand", "");
stopWord.put("quando", "");
stopWord.put("quant", "");
stopWord.put("quanta", "");
stopWord.put("quantas", "");
stopWord.put("quanto", "");
stopWord.put("quantos", "");
stopWord.put("quase", "");
stopWord.put("que", "");
stopWord.put("quel", "");
stopWord.put("quelle", "");
stopWord.put("quelqu''un", "");
stopWord.put("quelqu''une", "");
stopWord.put("quelque", "");
stopWord.put("quelques-unes", "");
stopWord.put("quelques-uns", "");
stopWord.put("quels", "");
stopWord.put("quem", "");
stopWord.put("quiconque", "");
stopWord.put("quoi", "");
stopWord.put("quoique", "");
stopWord.put("r", "");
stopWord.put("rather", "");
stopWord.put("s", "");
stopWord.put("sa", "");
stopWord.put("sans", "");
stopWord.put("satisfaz", "");
stopWord.put("sauf", "");
stopWord.put("says", "");
stopWord.put("se", "");
stopWord.put("sej", "");
stopWord.put("ser", "");
stopWord.put("seguinte", "");
stopWord.put("selon", "");
stopWord.put("sem", "");
stopWord.put("sempre", "");
stopWord.put("senao", "");
stopWord.put("sendo", "");
stopWord.put("ses", "");
stopWord.put("seu", "");
stopWord.put("seus", "");
stopWord.put("shall", "");
stopWord.put("she", "");
stopWord.put("should", "");
stopWord.put("si", "");
stopWord.put("sien", "");
stopWord.put("sienne", "");
stopWord.put("siennes", "");
stopWord.put("siens", "");
stopWord.put("sim", "");
stopWord.put("simplesmente", "");
stopWord.put("since", "");
```

```
stopWord.put("sob", "");
stopWord.put("sobr", "");
stopWord.put("sobre", "");
stopWord.put("small", "");
stopWord.put("soi", "");
stopWord.put("soi-meme", "");
stopWord.put("soit", "");
stopWord.put("some", "");
stopWord.put("soment", "");
stopWord.put("somente", "");
stopWord.put("sont", "");
stopWord.put("soon", "");
stopWord.put("sous", "");
stopWord.put("still", "");
stopWord.put("sua", "");
stopWord.put("suas", "");
stopWord.put("such", "");
stopWord.put("suis", "");
stopWord.put("t", "");
stopWord.put("ta", "");
stopWord.put("tal", "");
stopWord.put("talvez", "");
stopWord.put("tambem", "");
stopWord.put("tampouco", "");
stopWord.put("tandis", "");
stopWord.put("tant", "");
stopWord.put("tanta", "");
stopWord.put("tantas", "");
stopWord.put("tanto", "");
stopWord.put("tantos", "");
stopWord.put("te", "");
stopWord.put("telle", "");
stopWord.put("telles", "");
stopWord.put("tem", "");
stopWord.put("teremos", "");
stopWord.put("tes", "");
stopWord.put("teus", "");
stopWord.put("than", "");
stopWord.put("that", "");
stopWord.put("the", "");
stopWord.put("their", "");
stopWord.put("themselves", "");
stopWord.put("then", "");
stopWord.put("there", "");
stopWord.put("these", "");
stopWord.put("they", "");
stopWord.put("thing", "");
stopWord.put("this", "");
stopWord.put("those", "");
stopWord.put("though", "");
stopWord.put("through", "");
stopWord.put("tienne", "");
stopWord.put("tiennes", "");
stopWord.put("tiens", "");
stopWord.put("tod", "");
stopWord.put("toda", "");
stopWord.put("todas", "");
stopWord.put("todavia", "");
stopWord.put("today", "");
stopWord.put("todo", "");
stopWord.put("todos", "");
```

```
stopWord.put("toi", "");
stopWord.put("tomara", "");
stopWord.put("ton", "");
stopWord.put("to", "");
stopWord.put("too", "");
stopWord.put("toujours", "");
stopWord.put("tous", "");
stopWord.put("toute", "");
stopWord.put("toutes", "");
stopWord.put("tranquila", "");
stopWord.put("tranquilamente", "");
stopWord.put("tranquilas", "");
stopWord.put("tranquilo", "");
stopWord.put("tranquilos", "");
stopWord.put("tras", "");
stopWord.put("tres", "");
stopWord.put("trop", "");
stopWord.put("tu", "");
stopWord.put("tua", "");
stopWord.put("tuas", "");
stopWord.put("tudo", "");
stopWord.put("u", "");
stopWord.put("ulterior", "");
stopWord.put("um", "");
stopWord.put("uma", "");
stopWord.put("umas", "");
stopWord.put("un", "");
stopWord.put("una", "");
stopWord.put("under", "");
stopWord.put("uno", "");
stopWord.put("uns", "");
stopWord.put("until", "");
stopWord.put("up", "");
stopWord.put("upon", "");
stopWord.put("upres", "");
stopWord.put("use", "");
stopWord.put("v", "");
stopWord.put("vai", "");
stopWord.put("varia", "");
stopWord.put("varias", "");
stopWord.put("vario", "");
stopWord.put("varios", "");
stopWord.put("vem", "");
stopWord.put("very", "");
stopWord.put("vez", "");
stopWord.put("vezes", "");
stopWord.put("voce", "");
stopWord.put("você", "");
stopWord.put("voces", "");
stopWord.put("vos", "");
stopWord.put("vossa", "");
stopWord.put("vossas", "");
stopWord.put("vosso", "");
stopWord.put("vossos", "");
stopWord.put("votre", "");
stopWord.put("votres", "");
stopWord.put("vous", "");
stopWord.put("vu", "");
stopWord.put("vide", "");
stopWord.put("w", "");
stopWord.put("was", "");
```

```
stopWord.put("we", "");
stopWord.put("were", "");
stopWord.put("what", "");
stopWord.put("when", "");
stopWord.put("wher", "");
stopWord.put("where", "");
stopWord.put("which", "");
stopWord.put("while", "");
stopWord.put("who", "");
stopWord.put("whom", "");
stopWord.put("whose", "");
stopWord.put("will", "");
stopWord.put("with", "");
stopWord.put("within", "");
stopWord.put("without", "");
stopWord.put("would", "");
stopWord.put("www", "");
stopWord.put("x", "");
stopWord.put("y", "");
stopWord.put("yet", "");
stopWord.put("you", "");
stopWord.put("your", "");
stopWord.put("yours", "");
stopWord.put("yourself", "");
stopWord.put("yourselves", "");
stopWord.put("z", "");
stopWord.put("i","");
stopWord.put("me","");
stopWord.put("my","");
stopWord.put("myself","");
stopWord.put("we","");
stopWord.put("us","");
stopWord.put("our","");
stopWord.put("ours","");
stopWord.put("ourselves","");
stopWord.put("you","");
stopWord.put("your","");
stopWord.put("yours","");
stopWord.put("yourself","");
stopWord.put("yourselves","");
stopWord.put("he","");
stopWord.put("him","");
stopWord.put("his","");
stopWord.put("himself","");
stopWord.put("she","");
stopWord.put("her","");
stopWord.put("hers","");
stopWord.put("herself","");
stopWord.put("it","");
stopWord.put("its","");
stopWord.put("itself","");
stopWord.put("they","");
stopWord.put("them","");
stopWord.put("their","");
stopWord.put("theirs","");
stopWord.put("themselves","");
stopWord.put("what","");
stopWord.put("which","");
stopWord.put("who","");
stopWord.put("whom","");
stopWord.put("this","");
```

```
stopWord.put("that","");
stopWord.put("these","");
stopWord.put("those","");
stopWord.put("am","");
stopWord.put("is","");
stopWord.put("are","");
stopWord.put("was","");
stopWord.put("were","");
stopWord.put("be","");
stopWord.put("been","");
stopWord.put("being","");
stopWord.put("have","");
stopWord.put("has","");
stopWord.put("had","");
stopWord.put("having","");
stopWord.put("do","");
stopWord.put("does","");
stopWord.put("did","");
stopWord.put("doing","");
stopWord.put("will","");
stopWord.put("would","");
stopWord.put("shall","");
stopWord.put("should","");
stopWord.put("can","");
stopWord.put("could","");
stopWord.put("may","");
stopWord.put("might","");
stopWord.put("must","");
stopWord.put("ought","");
stopWord.put("a","");
stopWord.put("an","");
stopWord.put("the","");
stopWord.put("and","");
stopWord.put("but","");
stopWord.put("if","");
stopWord.put("or","");
stopWord.put("because","");
stopWord.put("as","");
stopWord.put("until","");
stopWord.put("while","");
stopWord.put("of","");
stopWord.put("at","");
stopWord.put("by","");
stopWord.put("for","");
stopWord.put("with","");
stopWord.put("about","");
stopWord.put("against","");
stopWord.put("between","");
stopWord.put("into","");
stopWord.put("through","");
stopWord.put("during","");
stopWord.put("before","");
stopWord.put("after","");
stopWord.put("above","");
stopWord.put("below","");
stopWord.put("to","");
stopWord.put("from","");
stopWord.put("up","");
stopWord.put("down","");
stopWord.put("in","");
stopWord.put("out","");
```

```
stopWord.put("on","");
stopWord.put("off","");
stopWord.put("over","");
stopWord.put("under","");
stopWord.put("again","");
stopWord.put("further","");
stopWord.put("then","");
stopWord.put("once","");
stopWord.put("here","");
stopWord.put("there","");
stopWord.put("when","");
stopWord.put("where","");
stopWord.put("why","");
stopWord.put("how","");
stopWord.put("all","");
stopWord.put("any","");
stopWord.put("both","");
stopWord.put("each","");
stopWord.put("few","");
stopWord.put("more","");
stopWord.put("most","");
stopWord.put("other","");
stopWord.put("some","");
stopWord.put("such","");
stopWord.put("no","");
stopWord.put("nor","");
stopWord.put("not","");
stopWord.put("only","");
stopWord.put("own","");
stopWord.put("same","");
stopWord.put("so","");
stopWord.put("than","");
stopWord.put("too","");
stopWord.put("very","");
stopWord.put("one","");
stopWord.put("every","");
stopWord.put("least","");
stopWord.put("less","");
stopWord.put("many","");
stopWord.put("now","");
stopWord.put("ever","");
stopWord.put("never","");
stopWord.put("say","");
stopWord.put("says","");
stopWord.put("said","");
stopWord.put("also","");
stopWord.put("get","");
stopWord.put("go","");
stopWord.put("goes","");
stopWord.put("just","");
stopWord.put("made","");
stopWord.put("make","");
stopWord.put("put","");
stopWord.put("see","");
stopWord.put("seen","");
stopWord.put("whether","");
stopWord.put("like","");
stopWord.put("well","");
stopWord.put("back","");
stopWord.put("even","");
stopWord.put("still","");
```

```
        stopWord.put("way","");
        stopWord.put("take","");
        stopWord.put("since","");
        stopWord.put("another","");
        stopWord.put("however","");
        stopWord.put("two","");
        stopWord.put("three","");
        stopWord.put("four","");
        stopWord.put("five","");
        stopWord.put("first","");
        stopWord.put("second","");
        stopWord.put("new","");
        stopWord.put("old","");
        stopWord.put("high","");
        stopWord.put("long","");
        stopWord.put("0","");
        stopWord.put("1","");
        stopWord.put("2","");
        stopWord.put("3","");
        stopWord.put("4","");
        stopWord.put("5","");
        stopWord.put("6","");
        stopWord.put("7","");
        stopWord.put("8","");
        stopWord.put("9","");
        stopWord.put("1998","");
        stopWord.put("1999","");
        stopWord.put("2000","");
        stopWord.put("2001","");
        stopWord.put("2002","");
        stopWord.put("2003","");
        stopWord.put("100","");
        stopWord.put("120","");
        stopWord.put("-","");
        stopWord.put("/","");
        stopWord.put("é","");
        stopWord.put("às","");
        stopWord.put("as","");
        stopWord.put("não","");
        stopWord.put("há","");
        stopWord.put("também","");
        stopWord.put("assim","");
        stopWord.put("são","");
        stopWord.put("será","");
        return stopWord;
}


public static boolean isStopWord(String pWord) throws Exception {
     HashMap stopWords = loadStopWords();
     return stopWords.containsKey(pWord.trim().toLowerCase());
}

}
```

**//Comparação entre os vetores de contexto**

```
import org.apache.lucene.document.*;
import org.apache.lucene.search.*;
import org.apache.lucene.store.*;
import org.apache.lucene.index.FilterIndexReader.FilterTermPositions;
import org.apache.lucene.index.IndexReader;
import org.apache.lucene.queryParser.QueryParser;
import org.apache.lucene.analysis.standard.StandardAnalyzer;
import org.apache.lucene.index.*;
import java.io.*;
import java.util.*;
import java.io.*;
import java.text.*;


class BuscaTexto {

    public int i, j, k, l, m, prox, pos, posUsr = 0;
    public Object key = "";
    public Vector vetOrdenado = new Vector();
    public Vector todosVetUsr = new Vector();
    public Vector todosVetUsrNovo = new Vector();
    public Vector novoComparado = new Vector();
    public Vector termoUsuario;
    public Vector vetUsr;
    public String arqUsuario = "";



    public BuscaTexto()
    {
        Boolean fExiste = false;
        Boolean fPassou = false;
        File root = new File("D:\\Mestrado\\Atual\\DirUsuarios\\");
        File[] all = root.listFiles();


        try
        {
            for (i=0; i<all.length; i++) //Percorre todos os diretórios
            {
               if (all[i].isDirectory())
               {
                   System.out.println("\n"+"USUARIO:
                                      "+all[i].getName()+"\n");
                   String pathArquivos = "D:\\Mestrado\\Atual\\\
                                          DirUsuarios\\"+all[i].
                                          getName()+"\\Janeiro";
                   String pathIndice = "D:\\Mestrado\\Atual\\\
                                        IndiceUsuarios\\"+"
                                        Index"+all[i].getName()+
                                        "\\Janeiro";
                   File indexDir = new File(pathIndice);
                   File dataDir = new File (pathArquivos);
                   System.out.println("Janeiro");
                   Indexer.index(indexDir, dataDir); //Faz a indexação

                   //Calcula o TfIdf
                   Lexicon lex = new Lexicon();
```

```
        Collection termList = lex.calculateTfidf(pathIndice);
        Iterator list = termList.iterator();

        vetOrdenado.clear();
        while (list.hasNext()==true) {
           vetOrdenado.addElement(list.next());
        }
        System.out.println("Tamanho de vetOrdenado: " +
                           vetOrdenado.size());

        Vector termoUsuario = new Vector();
        for(prox=0;prox<200;prox++)
        {
           TermoTfidf objTfidf = new TermoTfidf();
           if (vetOrdenado.size() <= prox)
           {
              break;
           }
           else
           {
              objTfidf = (TermoTfidf) vetOrdenado.get(prox);
              termoUsuario.add(objTfidf);
           }
      }

        todosVetUsr.add(termoUsuario);
   }
}

//COMPARAÇÃO ENTRE OS VETORES DE CONTEXTO

//Cada j representa um Vector de um usuário
for(j=0;j<todosVetUsr.size();j++)
{
    System.out.println("\nCOMPARAÇÃO ENTRE O USUARIO " + j +
                       " E OS DEMAIS...");
    Vector vetUsr = new Vector();
    vetUsr = (Vector) todosVetUsr.get(j);
    for(l=0;l<todosVetUsr.size();l++)
    {
        Vector novoComparado = new Vector();
         Vector vetUsr1 = new Vector();
        vetUsr1 = (Vector) todosVetUsr.get(l);
        System.out.println("TAMANHO DE vetUsr1:
                           "+vetUsr1.size()+"\n");

        for(k=0;k<vetUsr.size();k++)
        {
            TermoTfidf TTfidf = new TermoTfidf();
            TTfidf = (TermoTfidf) vetUsr.get(k);
            System.out.println("\nO termo comparador K é: "+
            TTfidf.getTermo());

            fExiste = false;

            for(m=0;m<vetUsr1.size();m++)
            {
                TermoTfidf TTfidf1 = new TermoTfidf();
                TTfidf1 = (TermoTfidf) vetUsr1.get(m);
                System.out.println("O termo M é: "+
                TTfidf1.getTermo());
```

```
                If (TTfidf.getTermo().
                    equals(TTfidf1.getTermo()))
                {
                    System.out.println("Termo igual TTfidf:
                              "+TTfidf.getTermo());
                    System.out.println("Termo igual TTfidf:
                              "+TTfidf.getTfidf_termo());
                    System.out.println("Termo igual TTfidf1:
                              "+TTfidf1.getTermo());
                    System.out.println("Termo igual TTfidf1:
                              "+TTfidf1.getTfidf_termo());

                    fExiste = true;
                    if (j!=l)
                    {
                        novoComparado.add(TTfidf1);
                    }
                }

            }
            if ((j!=l) && (fExiste == false) &&
                 (vetUsr.size() != novoComparado.size()))
            {
                TermoTfidf TTfidf_Diferente =
                                        new TermoTfidf();
                TTfidf_Diferente.setTermo(TTfidf.termo);

                 Tfidf_Diferente.setTfidf_termo(TTfidf.tfidf_
                 termo);
                TTfidf_Diferente.setTfidf_termo(0);
                novoComparado.add(TTfidf_Diferente);
            }

            todosVetUsrNovo.add(j,vetUsr);
            if(todosVetUsrNovo.size()>1)
            {
                todosVetUsrNovo.remove(j+1);
            }
            if (j!=l){
               todosVetUsrNovo.add(l, novoComparado);
               if(todosVetUsrNovo.size()>(l+1)){
                  todosVetUsrNovo.remove(l+1);
               }
            }
        }
    }
    //LISTAGEM PARA CONFERÊNCIA

    for(pos=0;pos<todosVetUsrNovo.size();pos++)
    {
        Vector vetTeste = (Vector) todosVetUsrNovo.get(pos);
        System.out.println("\nTAMANHO DE vetTeste:
                           "+vetTeste.size());
        System.out.println("\nUSUÁRIO: "+ pos);
        for(i=0;i<vetTeste.size();i++)
        {
            TermoTfidf objeto = (TermoTfidf) vetTeste.get(i);
            System.out.println("Termo: "+ objeto.getTermo()+"
                        = "+ objeto.getTfidf_termo());
        }
```

```
                }

                //GERA ARQUIVO PARA O EXCEL CONTENDO TERMOS E TFIDF
                GeraArqExcel.GeraAE(this.todosVetUsrNovo, this.j);
                arqUsuario = "D:\\Mestrado\\Atual\\
                             ArquivosGerados\\tfidf_usr" + j +".txt";

                //É FEITA A GRAVAÇÃO NO ARQUIVO TFIDF.TXT
                OutputStreamWriter objOutput =
                              new OutputStreamWriter(new
                               FileOutputStream(arqUsuario, true));

                for(posUsr=0;posUsr<todosVetUsrNovo.size();posUsr++)
                {
                    objOutput.write("u" +posUsr+ " ");
                    Vector vetCapturaUsr = (Vector)
                    todosVetUsrNovo.get(posUsr);
                    for(i=0;i<vetCapturaUsr.size();i++)
                    {
                        TermoTfidf objeto =
                                    TermoTfidf)vetCapturaUsr.get(i);
                         key = objeto.getTfidf_termo();
                        objOutput.write(key + " ");
                    }
                    objOutput.write("\r\n");
                }
                objOutput.close(); //fecha o arquivo
            }
        } catch (Exception e) {
           System.out.println("Deu craps: " + e);
        }
    }
}
```

**//Calculo da similaridade**

```
import java.util.*;
import java.text.*;

class CalcSimilar {

    public Vector vectorUsuario = new Vector();
    public int i = 0;

    public CalcSimilar(Vector objetoUsuario,double vetPi[], int maior,
                       int j)
    {
        int posUsr = 0;
        int m = 0;
        int usuarioAtual = -1;
        int usuarioComp = 0; //usuário comparado com o usuário atual
        double soma = 0;
        Object userAtual =  "";
        double MatSimilar[][];

        for (posUsr=0;posUsr < objetoUsuario.size();posUsr++)
        {
            Usuario objUsr = (Usuario) objetoUsuario.get(posUsr);
            System.out.println("\nUSUARIO: "+ objUsr.idUsr + " SERA
                               SIMILAR A...");
            userAtual = objUsr.idUsr;
            usuarioAtual++;
            usuarioComp = 0;

            objUsr.vetSim = new double [maior][objetoUsuario.size()];
            objUsr.vetSimilaridade = new double [objetoUsuario.size()];
            MatSimilar = new double
            [objetoUsuario.size()][objetoUsuario.size()];


            while (i < objetoUsuario.size()) //é o usuario comparado
            {
                Usuario usrSimilar = (Usuario) objetoUsuario.get(i);
                System.out.println("SIMILAR A: " + usrSimilar.idUsr);


                    for (m=0;m<maior;m++)
                    {
                        if (m > objUsr.MatUsr.length-1)
                        {
                           objUsr.vetSim[m][i] = 0;
                        }
                        else
                        {
                           objUsr.vetSim[m][i] =
                          ((Double.parseDouble(objUsr.MatUsr[m])) *
                           usrSimilar.VetCompara[m][i] * vetPi[m]);
                        }

                        if (Double.isNaN(objUsr.vetSim[m][i]) == true)
                        {
                            objUsr.vetSim[m][i] = 0;
                        }

                        soma = soma + objUsr.vetSim[m][i];
```

```
                    }
                    objUsr.vetSimilaridade[i] = soma;
                    System.out.println("SIMILARIDADE DE " + objUsr.idUsr
                    + " E " + usrSimilar.idUsr + " = " +
                    (objUsr.vetSimilaridade[i])+"\n");

                MatSimilar[usuarioAtual][usuarioComp] =
                objUsr.vetSimilaridade[i];
                usuarioComp++;
                i++;
                soma = 0;
            }
            vectorUsuario.add(posUsr, objUsr);
            if (posUsr == j)
            {
               GeraArqSim.GeraASim(this.vectorUsuario, posUsr);
            }
            i = 0;
        }
    }
}
```

# Livros Grátis

( http://www.livrosgratis.com.br )

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo