Investigação sobre a Identificação de Assuntos em Mensagens de Chat

Stanley Loh[1,2], Daniel Lichtnow[1], Ramiro Saldaña[1], Thyago Borges[1], Tiago Primo[1], Rodrigo Branco Kickhöfel[1], Gabriel Simões[1]

[1]Universidade Católica de Pelotas (UCPEL) – Grupo de Pesquisa em Sistemas de Informação
R. Félix da Cunha, 412, Pelotas, RS – CEP 96010-000
[2]Universidade Luterana do Brasil (ULBRA) – Faculdade de Informática
R. Miguel Tostes, 101, Canoas, RS – CEP 92420-280

{lichtnow, rsaldana, thyago, rodrigok}@ucpel.tche.br, sloh@terra.com.br,
tiagoprimo@brturbo.com, gsimoes@vetorial.net

Resumo

O objetivo do presente trabalho é investigar o processo de classificaçã o de textos sobre mensagens de um chat na Web. As mensagens de chat possuem algumas particularidades como concisão, pouco cuidado com correções ortográficas e revisões, devido a serem escritas às pressas e geralmente por pessoas leigas. Neste trabalho, foram utilizados métodos simples de classificação para investigar as particularidades do processo de classificação sobre este tipo de texto.

Abstract

This paper investigates the process of classification of textual messages in a web chat. This kind of text has special characteristics, like concision, orthographic mistakes, mainly by being written in a hurry and by naïve people. In this work, simple methods were evaluated to raise some specialties of the process.

1 Introdução

Este trabalho se insere na área de classificação de textos (ou categorização). Muitos trabalhos já foram publicados nesta área, geralmente apresentando métodos novos de classificação ou técnicas para seleção de características. SEBASTIANI (2002) apresenta um *survey* sobre métodos de classificação de textos.

Entretanto, a maioria dos trabalhos realiza as avaliações dos métodos sobre textos bem escritos, como é o caso das coleções Reuters ou OHSUMED (SEBASTIANI, 2002). A primeira coleção contém textos jornalísticos, e a segunda contém títulos e resumos de textos médicos da base de dados MEDLINE. A característica principal de tais textos é que eles contêm informações bem objetivas e são escritos e revisados por profissionais (portanto, com pouquíssimos erros ortográficos). Uma dúvida que surge é se os mesmos métodos teriam o mesmo desempenho sobre textos com características diferentes.

Uns poucos trabalhos utilizaram outro tipo de coleção textual em suas avaliações, entre eles: BAKER & McCALLUM (1998), JOACHIMS (1997), LANG (1995), McCALLUM & NIGAM (1998), McCALLUM ET AL. (1998), NIGAM ET AL. (2000) e SCHAPIRE & SINGER (2000). Estes trabalhos utilizaram mensagens do *newsgroup* Usenet, uma lista de discussão na Web (disponível em http://www.cs.cmu.edu/~textlearning). Tal coleção é formada por mensagens de *e-mail*, postadas para 20 grupos de discussão diferentes. Cada grupo forma uma categoria diferente (um assunto para cada grupo). A diferença deste tipo de texto para os anteriores de Reuters e OHSUMED é que são escritos por pessoas leigas, muitas vezes com pressa e sem muito cuidado com ortografia ou erros

de digitação. A partir destes trabalhos, pode-se pensar em avaliar métodos de classificação de textos sobre mensagens de chat, que têm particularidades diferentes das mensagens de correio eletrônico, como as usadas na coleção do *newsgroup* Usenet.

Neste sentido, pode-se citar o trabalho de Khan e outros (KHAN ET AL., 2002), que analisaram mensagens de um chat mas com o objetivo de identificar relações sociais. Não foram explicados os métodos de extração de assuntos utilizados, nem foi feita nenhuma avaliação neste sentido.

O objetivo do presente trabalho é investigar o processo de classificação de textos sobre mensagens de um chat na Web. As mensagens de chat possuem algumas semelhanças com as de *newsgroups* por poderem ser escritas por qualquer pessoa, por serem escritas às pressas (a pressa é bem maior no chat que no *newsgroup*) e pelo pouquíssimo cuidado com a ortografia e erros de digitação. Mas as mensagens de chat se diferenciam principalmente por serem menores, muito mais concisas. Só para se ter uma idéia, as mensagens de *newsgroups* contêm, na maioria das vezes, mais de um período (texto entre pontos), enquanto que as de chat são geralmente formadas por apenas um período. Além disto, as mensagens de chat se caracterizam pela informalidade da linguagem utilizada.

Neste trabalho, foram utilizados métodos simples de classificação, para investigar as particularidades do processo com este tipo de texto. Não é objetivo apontar o melhor método, mas investigar como algumas características específicas de mensagens de chat podem influenciar o processo.

A identificação de assuntos em mensagens de chat tem várias aplicações, discutidas nas conclusões deste artigo. Uma delas é auxiliar sistemas de recomendação como o SisRecCol – Sistema de Recomendação para Apoio à Colaboração, o qual indica itens de uma Biblioteca Digital para participantes de um chat privado, conforme os assuntos que estão sendo discutidos (protótipo disponível em htt://gpsi.ucpel.tche.br/sisrec).

A seção 2 deste artigo apresenta os métodos utilizados para classificação das mensagens, a seção 3 discute as avaliações feitas e seus resultados, a seção 4 analisa os erros nos métodos e a seção 5 apresenta e discute as conclusões e contribuições.

2 Métodos Utilizados

O método de identificação de assuntos (*Text Mining*) funciona como um *sniffer* examinando cada uma das mensagens enviadas pelos usuários participantes do chat. Os assuntos ou temas são identificados pela comparação dos termos que aparecem nas mensagens com termos que estão relacionados a conceitos presentes numa ontologia (armazenada internamente na mesma ferramenta e descrita na próxima seção).

O método faz na verdade classificação (ou categorização) de textos, isto é, identifica assuntos nos textos presentes nas mensagens. Este método foi apresentado em LOH et al. (2000) e atingiu índices de acerto acima de 60% para textos de prontuários médicos de uma clínica psiquiátrica (o que é considerado ótimo, num domínio complexo como este). Este método está baseado em técnicas probabilísticas e, portanto, não utiliza técnicas de Processamento de Linguagem Natural para analisar a sintaxe de cada mensagem.

O algoritmo usado é baseado nos algoritmos Rocchio e Bayes (ROCCHIO, 1966; RAGAS & KOSTER, 1998; LEWIS, 1998) utilizando vetores de texto para representar textos e assuntos. O método avalia a similaridade entre o texto e um assunto usando uma

função de similaridade que calcula a distância entre os dois vetores, um representando a mensagem do chat e outro representando o assunto (da ontologia). Os vetores que representam os textos e os assuntos são compostos por uma coleção de termos onde existe um peso associado a cada termo. No caso das mensagens, o peso de cada termo é dado pela freqüência relativa de cada termo no texto, isto é, o número de ocorrências de um termo dentro de uma mensagem dividido pelo número total de termos na mesma mensagem. Já o peso de um termo em relação a um assunto representa a probabilidade que o termo tem de indicar um determinado assunto. Este peso é definido na ontologia e será melhor explicado na seção seguinte.

Na montagem dos vetores são ignoradas as chamadas *stopwords* - termos que aparecem com muita freqüência e que não tem relevância na identificação dos assuntos de uma discussão, tais como preposições e artigos, por exemplo. Feita a montagem dos vetores, os dois são comparados por meio de uma função *fuzzy* de similaridade. O método utilizado multiplica os pesos dos termos que estão presentes nos dois vetores, sendo que a soma destes produtos, limitada a 1, é o grau de similaridade existente entre a mensagem e o assunto. Este grau determina qual a probabilidade do assunto estar presente na mensagem. Um limiar mínimo (*threshold*) é usado para cortar graus indesejados, abaixo do qual é improvável que o assunto esteja presente na mensagem. Este limiar é estabelecido por especialistas humanos na configuração do sistema e funciona para todas as sessões. O limiar ainda está sendo testado.

O método é baseado no índice de relevância, proposto por RILOFF & LEHNERT (1994). O índice de relevância é "um conjunto de características que juntas podem predizer com confiança a descrição ou existência de um evento". Partindo desta premissa, o método considera que alguns termos presentes nas mensagens do chat podem portanto indicar a presença de um assunto com um grau de certeza. Conseqüentemente o processo de raciocínio *fuzzy* deverá avaliar a probabilidade de um assunto estar presente em um texto, analisando a intensidade desta s indicações (presença de termos). Isto significa que, se as palavras que descrevem um assunto *c* aparecem em um texto, existe uma probabilidade alta do assunto *c* estar presente no texto. A soma das indicações deve resolver problemas de ambigüidade. Por exemplo, a presença do termo '*inteligência*' gera uma ambigüidade por se referir tanto ao tema "inteligência artificial" quanto ao tema "inteligência competitiva", mas a presença de outros termos ajuda a resolver tal conflito.

A abordagem pode ser considerada sob o paradigma de processamento estatístico de linguagem natural, conforme classificação de KNIGHT (1999), uma vez que analisa freqüência de palavras e probabilidades.

Inicialmente estão sendo testados 2 métodos derivados, a saber:

a) um método que analisa todas as mensagens enviadas e, para cada uma delas, aceita como verdadeiro apenas o assunto que tem maior grau; neste método, se dois assuntos são identificados com o mesmo grau e se existe entre eles uma hierarquia (pai e filho), então o assunto mais específico é utilizado; caso o empate ocorra entre assuntos que estão em um mesmo nível da hierarquia, então um dos assuntos será tomado como verdadeiro de forma aleatória; uma variação seria permitir que vários assuntos fossem identificados para cada mensagem (formando um *ranking* pelo grau ou probabilidade);
b) um método que avalia um conjunto de mensagens para identificar o assunto sendo tratado na discussão; no método anterior, cada mensagem era analisada individualmente

para se identificar o assunto; neste, os pesos ou graus associados a cada assunto para cada mensagem vão sendo somados sendo um assunto identificado somente quando a soma dos pesos passa um determinado limiar (configurado manualmente) .

Conceito Identificado	Peso	Mensagem enviada
PROJETO DE BANCO DE DADOS	0.000900	éé, rodrigo como funcia a esquema, a tabela q grava os pesos fica sempre gravando? ou ela e limpa em determinado ponto?
REDES DE COMPUTADORES	0.002365	tô no notebook, mas pode ser minha conexão cable modem
REDES DE COMPUTADORES	0.001152	Abri as recomendações de redes parece que está bem
PROJETO DE BANCO DE DADOS	0.000900	uma chave estrangeira pode ser ao mesmo tempo uma chave primária?

Figura 1: Trecho extraído de uma sessão no chat

Para exemplificar a diferença entre estes 2 métodos, note-se a figura 1, contendo um trecho extraído de uma sessão no chat. Pelo primeiro método, tem-se na primeira coluna o assunto identificado para cada mensagem enviada (e o peso associando o assunto à mensagem). Neste caso, cada mensagem individualmente gera um assunto identificado. Pelo segundo método, os pesos gerados para cada mensagem devem ser somados, agrupados por assunto. Assumindo que o limiar mínimo fosse de 0,001 (que é o limiar que está sendo testado no momento), a primeira mensagem não geraria nenhum assunto. Já a segunda e a terceira mensagens gerariam assuntos de forma individual. A quarta mensagem geraria o assunto "Projeto de Banco de Dados" pela soma (o peso do assunto identificado nela com o peso da primeira mensagem, que também identificou este mesmo assunto).

Os métodos foram implementados utilizando as tecnologias livres como linguagens de programação PHP e Javascript, banco de dados MySql, servidor *Web* Apache e sistema operacional Linux. Estes métodos compõem um protótipo de sistema de recomendação, que encontra-se disponível no endereço http://gpsi.ucpel.tche.br/sisrec.

2.1 A Ontologia

O sistema utiliza uma ontologia de domínio para classificar documentos, para identificar temas nas mensagens e para traçar o perfil dos usuários. Uma ontologia de domínio (*domain ontology*) é uma descrição de "coisas" que existem ou podem existir em um domínio (SOWA, 2002) e descreve o vocabulário relacionado ao domínio em questão (GUARINO, 1998).

Neste trabalho, a ontologia foi implementada como um conjunto de assuntos em uma estrutura hierárquica (um nó raiz, e nós pais e filhos), onde cada assunto tem associado a si uma lista de termos e seus respectivos pesos, que ajudam e identificar o assunto presente nos textos das mensagens. Os pesos associados aos termos determinam a importância relativa ou a probabilidade de um determinado termo identificar o assunto em um texto.

Para gerenciar e manter a ontologia foram implementadas algumas ferramentas que permitem sua visualização (hierarquia de assuntos, termos associados aos assuntos), a inclusão e a remoção de assuntos e de termos e a modificação dos pesos dos termos associados aos assuntos. Na implementação atual, a ontologia está voltada para a área de Ciência da Computação, sendo os assuntos baseados na classificação da ACM (www.acm.org), mas novas sub-áreas foram acrescentadas. Entretanto, outras ontologias podem ser adicionadas.

A ontologia foi criada de forma semi-automática. A seleção de assuntos (áreas e sub-áreas da hierarquia) foi feita manualmente por especialistas na área de Computação. Após, ferramentas automatizadas foram utilizadas para identificar termos que pudessem indicar cada assunto, seguindo um processo tipicamente de aprendizado de máquina (*machine learning*). Neste processo, especialistas nos assuntos selecionaram documentos eletrônicos de cada assunto (aproximadamente 100 documentos para cada assunto) e uma ferramenta de software identificou os termos mais relevantes e determinou o peso de cada termo. Este peso foi calculado com base na freqüência do termo dentro dos documentos e também avaliando o número de documentos daquele assunto onde o termo aparecia. Este é um procedimento típico do método Rocchio, sendo que o vetor gerado chama-se centróide.

Uma revisão dos termos e pesos foi feita por especialistas nas áreas relacionadas, observando os termos que apareciam em mais de um assunto (considerados termos genéricos, os quais tiveram seu peso diminuído) e procurando normalizar os pesos (os maiores pesos em cada assunto deveriam estar em patamares semelhantes).

A ontologia possui termos em português e inglês. Para tanto, foi necessário gerar termos nas duas línguas, quando o processo automático não conseguiu identificá-los. Como não se usa nenhum tratamento de radicais (*stemming*), foi necessário gerar manualmente as variações lingüísticas (número, gênero e as principais conjugações verbais).

3 Avaliação Formal do Método

O método de identificação de assuntos em mensagens de chat foi avaliado de duas formas. A primeira foi uma avaliação feita de forma *offline*, tomando como entrada resumos (abstracts) e parágrafos de textos (artigos científicos) selecionados manualmente por assunto. A avaliação comparou os textos (resumos e parágrafos) de forma completa em relação a frases (períodos) extraídos destes textos. O objetivo desta avaliação era saber o grau de acerto do método quando textos bem escritos e bem objetivos eram utilizados como entrada para representar mensagens de um chat. Neste caso, procurou-se observar se o texto todo ou parte dele (frases) poderiam gerar resultados diferentes.

A segunda avaliação foi feita *online* sobre as mensagens enviadas pelo chat, durante sessões reais de discussão de grupos de pesquisa ou comunidades virtuais.

3.1 Avaliação Offline

Foram selecionados 15 artigos científicos de diversas área de Computação, coletados a partir da Biblioteca Digital Citeseer ou ResearchIndex (www.researchindex.com). Destes artigos, foram extraídos os resumos (abstracts) e 2 parágrafos do meio de cada artigo (escolha aleatória). Procurou-se observar o nível de acerto do método nas seguintes situações:

a) quando o resumo todo era submetido como entrada (admitindo uma mensagem extensa no chat);
b) quando cada frase do resumo era submetida individualmente como entrada (uma por vez, gerando cada frase uma avaliação diferente, ou seja, um assunto identificado para cada uma);
c) quando cada parágrafo extraído do artigo era submetido como entrada;
d) quando cada frase dos 2 parágrafos de cada texto era submetida como entrada.

Cada artigo foi previamente associado por especialistas a um assunto da ontologia, e para fins de avaliação dos métodos, este assunto foi assumido como o correto tanto para os resumos e parágrafos extraídos deste texto, quanto para as frases extraídas.

A saber, foram avaliadas no total 78 frases extraídas de resumos e 418 frases extraídas de parágrafos. Os resultados são apresentados na tabela 1. Os textos maiores geraram melhores resultados. Isto já era esperado, uma vez que o método é probabilístico e, portanto, identifica melhor o assunto quando existe um maior número de características presentes. Acredita-se que os resumos geraram melhores resultados que os parágrafos por terem informações mais abrangentes sobre a área, enquanto que os parágrafos poderiam tratar mais de detalhes do artigo.

Tabela 1: Resultados da Avaliação Offline

Tipo de entrada	% de acertos
Resumos	91,66%
Frases dos resumos	60,97%
Parágrafos	83,33%
Frases dos parágrafos	58,73%

3.2 Avaliação Online e Comparação entre Métodos

Para a avaliação *online*, foram selecionadas 3 sessões de discussão, onde grupos de pesquisa utilizaram o chat. As mensagens enviadas para o chat foram analisadas pelo sistema de identificação de assuntos conforme os métodos descritos anteriormente. As mensagens da primeira e da segunda sessão foram analisadas com o primeiro método, que identifica assunto para cada mensagem individualmente. Já as mensagens da terceira sessão foram analisadas com o segundo método, que procura identificar o assunto pela soma dos pesos das mensagens (agrupadas por assunto).

Os próprios participantes das sessões analisaram os assuntos identificados para as mensagens e decidiram o que estava correto ou errado, bem como as mensagens que deveriam ter gerado assunto e não o fizeram.

A primeira sessão teve um total de 168 mensagens enviadas para o chat, sendo que em 48 mensagens foi identificado um assunto, mas em 120 mensagens não foi possível identificar nenhum assunto. Destas 120 mensagens, 9 mensagens deveriam ter permitido identificar algum assunto. Das 48 mensagens, 18 permitiram identificar o assunto correto.

A segunda sessão teve um total de 184 mensagens enviadas para o chat, sendo que em 52 mensagens foi possível identificar um assunto, mas em 132 mensagens não foi identificado nenhum assunto. Em 26 mensagens, foi identificado o assunto correto. Novamente em 9 mensagens deveria ter sido identificado um assunto e isto não ocorreu.

Na primeira sessão, houve uma precisão de 37,5% (proporção de assuntos corretamente identificados) e abrangência de 66,6% (proporção de mensagens com assunto corretamente identificado em relação ao total de mensagens que deveriam gerar assunto). Já na segunda sessão, a precisão ficou em 50%, sendo a abrangência igual a 74,3%. Uma das explicações para a melhora de precisão é que a segunda sessão foi mais técnica, isto é, poucas mensagens estavam relacionadas a aspectos administrativos do grupo.

A avaliação do segundo método (terceira sessão) levou em conta um conjunto de mensagens para identificar um assunto, e não mensagens individualmente. Foi utilizado um limiar para somar os pesos, como explicado anteriormente. Procurou-se determinar, além do grau de acerto em geral, qual limiar gerava melhores resultados. A terceira sessão teve

um total de 374 mensagens, sendo que em 258 não foi possível identificar nenhum conceito. Em 116 mensagens, foi possível identificar um assunto. Entretanto, algumas destas mensagens tinham associado a elas um peso abaixo do limiar. Portanto, para fins de avaliação deste segundo método, somente foram consideradas as mensagens que geraram a identificação de um assunto, isto é, quando a soma dos pesos ficava acima do limiar estabelecido.

Com um limiar igual a 0,001, foram identificados assuntos para 83 mensagens, isto é, 83 mensagens geraram um assunto quando a soma ultrapassou o limiar. Portanto, das 116 mensagens, 33 não ultrapassaram o limiar pela soma. Das 83, 56 tiveram o assunto correto identificado. Em 9 mensagens deveria ter sido identificado um assunto mas nada foi identificado. Assim, a precisão ficou em 67,5% e a abrangência em 86,1%. Com o limiar em 0,005, foram identificados assuntos em 63 mensagens, com 54 corretas, e tendo 9 mensagens sido deixadas de fora. Neste caso, a precisão melhorou para 85,7% e a abrangência baixou um pouco para 85,7%. Aumentando o limiar para 0,01, 54 mensagens geraram assuntos (45 corretas e 9 deixadas de fora), resultando numa precisão de 83,3%, mas a abrangência caiu levemente para 83,3%.

Deste resultado, conclui-se que o limiar 0,005 poderia ser utilizado neste método que considera a soma dos pesos. As mensagens para as quais deveria ter sido identificado um assunto e não foi, não foram deixadas de fora por causa do limiar, mas sim por erros de outro tipo.

Comparando os métodos, o que se nota é que o método que utiliza a soma dos pesos para identificar um assunto (usado na terceira sessão) aumenta bastante a precisão, pois somente identifica um assunto quando várias mensagens com pesos médios forem enviadas ou quando uma mensagem com peso alto é enviada (simulando uma análise de contexto).

4 Análise dos Erros

A partir dos experimentos de avaliação dos métodos utilizados, procurou-se analisar causas dos possíveis erros na identificação de assunto. Uma conclusão a que se chegou é que erros ortográficos, gírias e abreviaturas tendem a confundir os métodos de identificação dos assuntos. Um corretor ortográfico e a inclusão na ontologia dos demais termos que puderem ser previstos (os mais comuns) devem minimizar tais problemas.

Pôde-se notar também a presença de muitas expressões fora do contexto técnico da discussão (como *"oi"*, *"o fulano entrou no chat"*) ou que retrucavam ("*que acham*?", *"por quê*?"). Estas, em sua maioria, serviram para identificar nenhum assunto. Entretanto, algumas identificaram assuntos errados, devido a esta economia de palavras.

Uma das principais características das mensagens de um chat é serem concisas, para se ganhar tempo. A conseqüência é que há muita informação subentendida. Por exemplo, quando o grupo estava discutindo sobre "desempenho no chat devido a problemas de conexão" (assunto "redes de computadores"), muitas mensagens utilizavam somente o termo "redes", admitindo que o grupo já entendia seu significado (sem confundir com "redes neurais" ou outro assunto). A conclusão é que métodos que analisem o contexto devem ser utilizados. Neste trabalho, foi comprovado que um método que faz tal análise (mesmo apesar de não ser uma análise tão complexa) tende a melhorar a precisão (lembrando que o segundo método, que fazia a análise de um conjunto de mensagens, foi

melhor que o primeiro). Entretanto, concluiu-se que o esquema de soma dos pesos não é o melhor, uma vez que, quando a soma atinge um limiar, todas as mensagens posteriores sobre o mesmo tema irão gerar o assunto. Uma opção a ser testada futuramente é utilizar grupos de mensagens por janelas ou por tempo, para avaliar a soma (a soma só seria avaliada sobre as últimas N mensagens).

Outra constatação é que mensagens com peso baixo podem indicar a falta de um assunto específico na ontologia. Este foi o caso do uso do termo "recomendações", que indica o assunto "sistemas de informação" com peso baixo. Entretanto, poder-se-ia criar um assunto "filho" (especialização), onde este termo teria um peso maior.

A partir da análise dos conceitos erroneamente identificados, também foi possível notar que havia falhas nos pesos dos termos associados a alguns conceitos na ontologia. Termos de significado muito genérico tendem a induzir a erros se tiverem pesos muito altos (por exemplo: projeto, sistema, técnicas). Os pesos destes termos foram diminuídos manualmente na ontologia. Uma implementação futura deverá analisar os pesos dos termos na ontologia, de forma automática, para diminuir o peso de termos que aparecem em muitos assuntos, conforme sugestão de SALTON & McGILL (1983). Outra maneira de minimizar tal problema é cuidando para que os termos de maior peso em cada conceito estejam na mesma faixa de valor (normalizados).

Outro erro comum foi identificar um assunto "filho" (mais específico) quando deveria ser identificado o conceito "pai" (exemplo: o termo "tabelas" levou ao conceito "projeto de banco de dados" ao invés de "banco de dados"). A solução encontrada foi diminuir nos conceitos "filhos" o peso de termos que são mais genéricos, ou seja, identificam melhor o conceito "pai". Está sendo planejada uma ferramenta que encontra palavras que aparecem em conceitos "pai" e "filho", para serem apresentadas a um especialista que, manualmente, poderá modificar os seus pesos.

Um caso especial encontrado a partir da análise dos erros é o de mensagens com vários assuntos (ex: uma mensagem citava "inteligência artificial", "arquitetura de computadores" e outras sub-áreas). Um termo com peso alto num destes assuntos influencia o assunto final identificado. Uma solução pode ser equiparar os pesos nos assuntos (normalização). Mas a questão principal é que somente um assunto vai ser identificado (resposta final do método), quando na verdade o correto seria indicar os vários assuntos presentes (lembrando que os métodos podem detectar vários assuntos, mas somente identifica como assunto correto o de maior peso). Uma curiosidade é que mensagens com vários assuntos (corretamente identificados) tinham pesos semelhantes para os assuntos detectados. Isto poderia ser utilizado pelos métodos para identificar vários assuntos numa mesma mensagem (quando diversos assuntos forem detectados com peso acima do limiar então é porque realmente existem vários conceitos sendo discutidos na mensagem ou num grupo de mensagens).

Por fim, o sistema descrito neste artigo separa numa lista as palavras que apareceram no chat mas que não estavam presentes na ontologia nem eram "*stopwords*". Analisando-as, notou-se haver expressões que realmente não deveriam estar na ontologia (como "xi", "haha", nomes próprios, gírias, erros ortográficos e números), mas também puderam ser identificadas novas *stopwords* (verbos genéricos, por exemplo) e termos bastante significativos (exemplo "distribuídos"), que ficaram de fora da ontologia. Fica claro que a ontologia deve ser revisada e uma das formas é coletar os termos usados no chat

e que não se encontram na ontologia.

5 Conclusões

Este trabalho avaliou a identificação de assuntos em mensagens de chat. Para tanto, foram implementados, avaliados e comparados dois métodos probabilístico s.

A principal contribuição do artigo (e também sua diferença para outros trabalhos já comentados) está em avaliar o processo de classificação de textos curtos, concisos e escritos rapidamente, como é o caso das mensagens de chat. Só para constar, as mensagens do newsgroup Usenet, utilizadas em alguns trabalhos, tem em média 124,7 palavras, já excluindo as chamadas *stopwords*. Nas mensagens de chat analisadas, encontrou-se uma média de 3,3 palavras por mensagem, sem *stopwords*, e uma média de 6,35 palavras contando as *stopwords*. A concisão das mensagens de chat é um empecilho aos métodos de classificação de texto. Entretanto, foi demonstrado que métodos simples (baseados em estatística e que não utilizam análise sintática) podem consegui r um bom nível de precisão e abrangência (85,7% e 87,5%, respectivamente, no melhor caso).

Este bom desempenho ainda poderia ser melhor se erros ortográficos e de digitação, bem como gírias e abreviaturas pudessem ser identificados e corrigidos ou transformados para termos correspondentes na ontologia. A análise dos erros gerados permitiu identificar algumas características das mensagens de chat que podem influenciar o processo.

Outro fator que influencia o desempenho dos métodos de classificação é a qualidade da ontologia. Na ontologia utilizada neste trabalho, havia uma média 145 palavras em cada vetor que definia um assunto, gerando um vocabulário com 2874 palavras (mais 443 termos considerados *stopwords*), em português e inglês.

Se comparada aos vocabulários utilizados em outros trabalhos, este quantia é pequena. Por exemplo, BAKER & McCALLUM (1998) utilizaram um vocabulário de 62258 palavras, sem *stopwords*, sem tratamento de *stemming* e excluindo palavras que apareciam uma vez só. Já McCALLUM,& NIGAM (1998) utilizaram um total de 22958 palavras, sem *stemming* e sem as que apareciam só uma vez. McCALLUM ET AL. (1998) usaram um vocabulário de 52309 palavras, sem *stemming*, com lista de *stopwords*, mas removendo palavras que apareciam só uma vez. Entretanto, os trabalhos citados usaram parte das próprias mensagens como treino. A ontologia utilizada nos experimentos apresentados no presente artigo foi construída a partir de artigos científicos e não de mensagens. Somente parte do vocabulário da ontologia se fez presente nas mensagens analisadas. Para se ter uma idéia, a primeira sessão teve 534 palavras diferentes, sem contar *stopwords*. Já na terceira sessão, apareceram 374 palavras diferente s, sem considerar *stopwords*. Por curiosidade, vale salientar que as palavras mais freqüentes apareceram no máximo 11 vezes numa sessão.

A aplicação dos resultados deste trabalho poderá melhorar o processo de identificação de temas em mensagens de chat. Isto tem conseqüências diretas sobre sistemas que analisam discussões em chats, tais como:

- sistemas de identificação de especialistas ou análise de *expertise*: para encontrar pessoas autoridades em determinado assunto ou simplesmente identificar quem conhece algo sobre algum assunto;
- sistemas de recomendação: para indicar itens de forma sensível ao contexto

(ofertas personalizadas conforme interesse de cada pessoa que participa do chat);

- sistemas de publicidade online (*adverstising*): para apresentar informações personalizadas de acordo com assuntos sendo discutidos no chat.

6 Agradecimentos

O presente trabalho foi realizado com o apoio do CNPq, uma entid ade do Governo Brasileiro voltada ao desenvolvimento científico e tecnológico.

7 Referências Bibliográficas

BAKER, L. D. & McCALLUM, A. K. 1998. Distributional clustering of words for text classification. IN: Proceedings ACM International Conference on Re search and Development in Information Retrieval, SIGIR-98, 21., Melbourne, 1998. p.96-103.

GUARINO, Nicola. 1998. Formal Ontology and Information Systems. In: International Conference on Formal Ontologies in Information Systems - FOIS'98, Trento, Itália, Junho de 1998. p. 3-15

JOACHIMS, T. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. IN: Proceedings International Conference on Machine Learning, ICML-97, Nashville, 1997. p.143-151.

KHAN, Faisal M. ET AL. 2002. Mining chat-room conversations for social and semantic interactions. Technical Report, LU-CSE-02-011, Lehigh University.

KNIGHT, Kevin. 1999. Mining online text. Communications of the ACM, v.42, n.11, p.58-61.

LANG, K. 1995. NewsWeeder: learning to filter netnews. IN: Proceedings International Conference on Machine Learning, ICML-95, 12., Lake Tahoe, 1995. p.331-339.

LEWIS, David D. 1998. Naive (bayes) at forty: The independence assumption in information retrieval. In: European Conference on Machine Learning, Chemnitz, Alemanha, 1998. p.4-15. (Lecture Notes in Computer Science, v.1398).

LOH, S.; WIVES, L. K.; OLIVEIRA, J. P. M. 2000. Concept-based knowledge discovery in texts extracted from the Web. ACM SIGKDD Explorations, v.2, n.1, Julho de 2000, p. 29-39.

McCALLUM, A. K. & NIGAM, K. 1998. Employing EM in pool-based active learning for text classification. IN: Proceedings International Conference on Machine Learning, ICML-98, Madison, 1998. p.350-358.

McCALLUM, A. K.; ROSENFELD, R.; MITCHELL, T. M.; NG, A. Y. 1998. Improving text classification by shrinkage in a hierarchy of classes. IN: Proceedings International Conference on Machine Learning, ICML-98, Madison, 1998. p.359-367.

NIGAM, K.; McCALLUM, A. K.; THRUN, S.; MITCHELL, T. M. 2000. Text classification from labeled and unlabeled documens using EM. Machine Learning, v. 39, n.2/3, p.103-134.

RAGAS, Hein & KOSTER, Cornelis H. A. 1998. Four text classification algorithms compared on a Dutch corpus. In: International ACM-SIGIR Conference on Research and Development in Information Retrieval, Melbourne, 1998, p.369-370.

RILOFF, Ellen & LEHNERT, Wendy. 1994. Information extraction as a basis for high-precision text classification. ACM Transactions on Information Systems, v.12, n.3, Julho de 1994, p.296-333.

ROCCHIO, J. J. 1966. Document retrieval systems - optimization and evaluation. Tese (Doutorado)- Harvard University, Cambridge.

SALTON, G. & McGILL, M. J. 1983. Introduction to modern information retrieval. New York: McGraw-Hill, 1983.

SEBASTIANI, Fabrizio. 2002. Machine learning in automated text categorization. ACM Computing Surveys, v.34, n.1, Março de 2002.

SCHAPIRE, R. E. & SINGER, Y. 2000. BoosTexter: a boosting-based system for text categorization. Machine Learning, v. 39, n.2/3, p.135-168.

SOWA, John F. 2002. Building, sharing, and merging ontologies. Disponível em http://www.jfsowa.com/ontology