CATEGORIZAÇÃO DE TEXTO UTILIZANDO HEURÍSTICAS E LÓGICA DIFUSA

Luiz F. B. Loja[1], Renato S. Gomide[1], Sirlon D. Carvalho[2], Ricardo A. G. Teixeira[3], Francisco R. Melo[4], Edna L. Flôres[1]

1 – Universidade Federal de Uberlândia, Departamento de Engenharia Elétrica, Uberlândia – MG, luiz@doutorado.ufu.br, renato.s.gomide@gmail.com, edna@ufu.br

2 – Instituo Federal de Educação, Ciência e Tecnologia Goiás, Campus Luziânia, Luziânia – GO, sirlondiniz@gmail.com

3 – Universidade Federal de Goiás, Goiânia – GO, professorricardoteixeira@gmail.com

4 – Universidade Estadual de Goiás, Anápolis – GO, francisco.melo@ueg.br

***Resumo* - O objetivo dos fóruns de discussão existentes na Internet é proporcionar a seus usuários a troca de informações de maneira fácil e rápida. Esses possuem uma estrutura de tópicos e assuntos bem definida. Porém, por simples distração ou até mesmo por falta de conhecimento parte dos usuários não relacionam seus comentários com os devidos assuntos ou tópicos correlatos. Este artigo apresenta uma ferramenta que objetiva categoriza um texto em seu respectivo assunto utilizando métodos heurísticos e lógica difusa. Após vários testes, conclui-se que a categorização de textos pode ser realizada parcialmente utilizando os métodos desenvolvidos.**

***Palavras-Chave* - inteligência artificial, heurística, categorização de textos, lógica difusa, fórum**

TEXT CATEGORIZATION USING HEURISTICS AND FUZZY LOGIC

***Abstract* - The aim of the discussion forums hosted on Internet is provide its users to exchange information quickly and easily. These forums, have a structure of topics and issues well defined. However, because of simple distraction or even lack of knowledge most users do not relate his comments on the appropriate issues or topics. This paper presents a tool that aims to categorize a text on your subject matter using heuristic methods and fuzzy logic. After a lot of tests, it is concluded that text categorization can be partially made by using the developed methods.**

***Keywords* - artificial intelligence, heuristic, defining texts, fuzzy logic, fórum.**

I. INTRODUÇÃO

A estrutura organizacional dos fóruns convencionais é bem definida. A primeira parte dessa estrutura é dividida em mensagens por assunto. Por exemplo: um fórum de uma faculdade pode ser dividido em assuntos tais como computação, direito, odontologia e assim por diante. Para cada assunto uma segunda divisão é realizada por meio de tópicos. E por sua vez esses possuem assuntos específicos que podem ser comentados pelos usuários.

Por exemplo, em um fórum de computação quando um usuário deseja fazer uma pergunta, mais especificamente sobre Java (linguagem de programação), ele deve inserir sua mensagem no assunto linguagem de programação no tópico de Java. Entretanto, muitas vezes os usuários por pura falta de informação ou até mesmo por displicência acabam inserindo uma mensagem em um assunto ou tópico não apropriado.

O fato da mensagem não estar no lugar correto prejudica o próprio usuário, pois os outros integrantes do fórum que interagem com aquele assunto não visualizam a pergunta enviada. Assim, não podem respondê-la.

Com o objetivo minimizar esse problema iniciou-se o desenvolvimento de uma ferramenta que analisa os comentários dos usuários classificando o texto inserido.

Para classificar um texto devem ser considerados vários aspectos da linguagem. Por exemplo, o sentido semântico das palavras e sua relevância para cada assunto. Na classificação do texto, a ferramenta implementada neste trabalho utiliza lógica difusa e métodos heurísticos.

Este artigo está estruturado em cinco seções. A Seção II apresenta o processo de formação da base de conhecimento da ferramenta. A Seção III descreve o processo utilizado na categorização do texto. A Seção IV mostra a ferramenta desenvolvida que auxilia na verificação dos valores obtidos com as técnicas de classificação utilizadas. A Seção V apresenta um breve estudo de caso utilizando a ferramenta desenvolvida. E finalmente, a ultima seção realiza as conclusões.

II. PROCESSO DE FORMAÇÃO DA BASE DE CONHECIMENTO

O objetivo desta seção é apresentar o processo de formação da base de conhecimento. As três etapas desse processo são: escolha dos textos para fomentação da base de palavras, pré-processamento dos textos selecionados e formação da base de conhecimento de palavra-categoria.

A. Escolha dos Textos para Fomentação da base de Palavras

O primeiro passo para construção da base de palavras foi definir quais áreas de pesquisa seriam utilizadas pela ferramenta. As áreas de conhecimento escolhidas foram computação, química, odontologia, direito e sociologia, pois estas áreas possuem vocabulários diferenciados.

XI CEEL – ISSN 2178-8308
25 a 29 de novembro de 2013
Universidade Federal de Uberlândia – UFU
Uberlândia – Minas Gerais – Brasil

Com o finalidade de selecionar palavras relacionadas a essas áreas, foram realizadas várias pesquisas em diversos fóruns e sites especializados em assuntos relacionados a elas.

Os textos foram reunidos e inseridos na ferramenta. Para cada texto introduzido foi informado área de conhecimento e realizado um pré-processamento do texto antes de carregá-lo.

B. Pré-processamento do texto

A finalidade de preparar o texto é retirar palavras que não são relevantes na sua categorização, formatar as palavras que possuem diversas formas como plural, gerúndio e sufixos temporais e retirar a acentuação. Esta etapa foi baseada em parte do processo proposto por [6]. Este procedimento facilita a análise textual e sintetiza a lista de palavras pertencentes a cada categoria.

Conforme em [1] a vantagem de usar essas técnicas é a diminuição significativa do tempo de processamento, na análise do texto e no carregamento das palavras em suas determinadas categorias.

A retirada dessas palavras aumenta o poder de expressão do texto, pois as palavras sem relevância são eliminadas e as palavras que possuem mesmo radical são agregadas.

As duas fases da preparação do texto são: retirada de *stopwords* e *stemming* [2].

1) Retirada de Stopwords

A remoção das *stopwords* é o procedimento de retirada de palavras sem contribuição semântica no texto. Estas palavras repetem-se inúmeras vezes e não auxiliam no seu entendimento [1].

Existem diversas categorias de *stopwords*. Conforme [6], algumas categorias de stopwords são advérbios, artigos, pronomes, vogais, consoantes, entre outras.

[3] apresentou uma lista de palavras que obedecem esse padrão. No pré-processamento, neste trabalho foi utilizada essa lista.

Com objetivo igualar o texto antes de aplicar o método de remoção, as palavras são passadas para o diminutivo, são removidos seus acentos e toda a pontuação é eliminada. Após esse processamento o texto está pronto para a próxima fase.

2) Stemming

Na língua portuguesa uma palavra pode assumir diversas formas, tais como gerúndio, plural ou possuir sufixos temporais. A finalidade do processo de *Stemming* é normalizar a palavra. Por exemplo, são consideradas as mesmas palavras **casa** e **casas**, se o **s** for removido da segunda palavra. Com essa retirada em um texto essas duas palavras aparecem duas vezes como **casa**.

A Figura 1 mostra uma lista de exemplos da aplicação do processo de *stemming*. Nessa figura a primeira coluna indica o sufixo que deve ser encontrado para que o processo de *stemming* seja aplicado. A segunda coluna informa o tamanho mínimo da palavra para que esse processo possa ser utilizado. A terceira coluna apresenta o caractere que o sufixo substitui. E finalmente, a quarta coluna mostra um exemplo de conversão realizado pelo processo de *stemming*.

Para retirar essas variações lingüísticas utilizou-se neste trabalho o processo proposto por [4] que consiste em várias regras gramaticais.

Regras de Redução de Plural			
Suf. Remover	**Stem Mínimo**	**Substituição**	**Exemplo**
"ns"	1	"m"	bons → bom
"ões"	3	"ão"	balões → balão
"ães"	1	"ão"	capitães → capitão
"ais"	1	"al"	noarmais → normal
"éis"	2	"el"	papéis → papel
"eis"	2	"el"	amáveis → amável
"óis"	2	"ol"	lençóis → lençol
"is"	2	"il"	barris → barril
"les"	3	"l"	males → mal
"res"	3	"r"	mares → mar
"s"	2	""	casas → casa

Fig. 1. Aplicação do *Stemming*.

Para obter a forma primitiva de uma palavra geralmente é necessária a utilização de mais de uma regra. Portanto, neste artigo aplicou-se várias regras diferentes até obter a palavra normalizada.

Após o processo de *stemming* o texto obtido é utilizado para formar a base de conhecimento palavra-categoria.

C. Formação da Base de Conhecimento Palavra-Categoria

Após o pré-processamento, conforme sua categoria o texto é carregado na base de dados. Para isso, é contado no texto a quantidade de repetições da palavra. A cada palavra repetida é acumulado um ponto de relevância para à categoria que o texto pertence. A quantidade total de pontos acumulados por categoria constitui o grau de relevância da palavra em relação ela. A relevância é utilizada na fase classificação do texto.

Neste trabalho para cada assunto foram carregados mais de 50 textos. Assim, cada categoria possui uma quantidade média de 10 a 15 mil pontos.

As palavras em que a relevância foi menor do que 10, ou seja, aquelas que repetiram menos de 10 vezes considerando todos os textos, foram retiradas da lista de relevância. A finalidade disso foi diminuir as palavras que não possuem relevância tão alta, pois essas palavras influenciariam negativamente na categorização do texto.

III. MÉTODOS PROPOSTOS

De acordo com as categorias pré-definidas, esta seção apresenta as técnicas utilizadas na classificação do texto. A Subseção A justifica o uso dos métodos heurísticos na categorização de texto. A Subseção B apresenta a primeira regra heurística de análise baseada no número de repetições da palavra no texto. A Subseção C mostra o segundo método heurístico utilizado na classificação dos textos. Finalmente, a Subseção D apresenta a lógica difusa utilizada para fazer o balanceamento entre os resultados obtidos pelos dois métodos heurísticos. E a Subseção E apresenta as conclusões finais desse processo.

A. Uso de Métodos Heurísticos para a Classificação de Textos

Não existe um algoritmo bem definido para a classificação de textos. Assim, o uso de métodos heurísticos é aconselhável, pois eles servem justamente para solucionar problemas que não possuem uma solução bem estabelecida [5].

Portanto, para categorizar um texto foram definidos dois métodos heurísticos e um método que utiliza lógica difusa.

B. Método de Pontuação Geral

Após o processo de pré-processamento do texto ele é analisado em termos de pontuação geral. Esta pontuação é calculada multiplicando o número de repetições da palavra no texto, pela sua relevância em cada categoria.

Inicialmente o valor da pontuação geral de cada categoria é igual a zero. Este valor é alterado pelo somatório da quantidade de palavras pertencentes a categoria multiplicado pela sua relevância. Quanto maior é a pontuação de uma categoria nessa etapa, maior a probabilidade do texto pertencer a ela. Essa etapa possui maior grau de pertinência na análise final do texto e conclui a primeira etapa do processo classificação.

C. Método de Relevância

A segunda fase do processo de análise classifica a palavra em uma determinada categoria. Para cada palavra existente no texto é verificado seu grau de relevância em relação a todas as categorias selecionadas. A palavra é classificada conforme a categoria que possui o maior grau de relevância.

Por exemplo: para área de computação a palavra “computador” possui um grau de relevância igual a 350, porém para as demais áreas o grau de relevância dessa palavra é menor do que 200. Dessa maneira, a palavra computador é classificada como pertencente a área de computação.

Após as palavras serem classificadas por categoria, o texto é classificado pelo maior número de palavras relacionadas a ele.

A etapa do método de relevância é a segunda fase na análise do texto com grau de pertinência menor do que o método de pontuação geral.

D. Método de lógica difusa

Os dados de pontuação e relevância das palavras auxiliam a categorização do texto. E a possibilidade de exibição dos dados em gráficos de barra e de pizza facilita a interpretação humana da categorização.

A classificação estatística não é suficiente para categorização do texto [7]. Com o objetivo de melhorar a classificação do texto foi adicionado outro método de classificação utilizando lógica difusa. Este sistema foi modelado com a finalidade de levar em consideração o grau de diferença entre a quantidade de palavras obtidas em cada categoria e associar tal diferença com a relevância de palavras obtidas na mesma categoria.

No sistema difuso foram implementadas cinco variáveis de entrada, 27 regras e seis variáveis de saída. A Figura 2 mostra o sistema de categorização nebuloso.

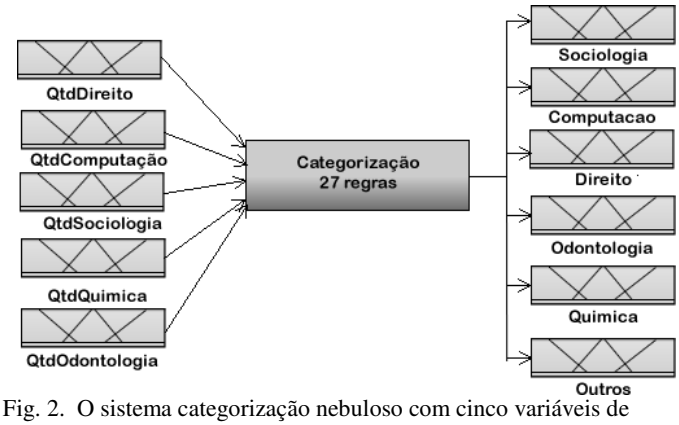

Fig. 2. O sistema categorização nebuloso com cinco variáveis de entrada, 27 regras e seis variáveis de saída

As variáveis de entrada do sistema de categorização são: QtdSociologia, QtdComputacao, QtdDireito, QtdOdontologia e QtdQuimica. Essas variáveis recebem as quantidades de palavras por assuntos e possuem três termos nebulosos, "pouco", "médio" e "muito". A Figura 3 mostra as variáveis de entrada utilizadas em sociologia.

Por exemplo, os parâmetros das variáveis de entrada da variável QtdSociologia, são:

- pouco [0 0 6];
- médio [3 7 12];
- muito [9 15 500];

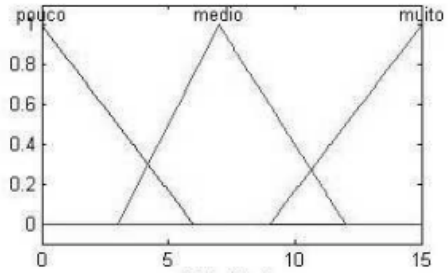

Fig. 3. Exemplo da variável de entrada QtdSociologia.

As regras do sistema nebuloso levam em consideração somente as variáveis de entrada e a combinação entre essas variáveis é mostrada na Tabela 1.

Um exemplo de regra definida na Tabela 1 é: **Se** é muito a quantidade de termos sociologia **então** o grau de pertinência dessa área é alto e a outros assuntos é baixo.

Sociologia, Computacao, Direito, Odontologia, Quimica e Outros, são as variáveis de saída. Estas variáveis recebem o grau de pertinência do documento categorizado em cada um dos assuntos. Se o documento não é considerado pertencente a nenhum dos cinco assuntos ele é classificado como pertencente a "outros assuntos".

A Figura 4 mostra que todas as variáveis de saída possuem três termos nebulosos, "baixo", "medio" e "alto".

Exemplo dos parâmetros da variável de saída Sociologia são:

- baixo = [0 0 0,4]
- medio = [0,3 0,5 0,7]
- alto = [0,6 1 1]

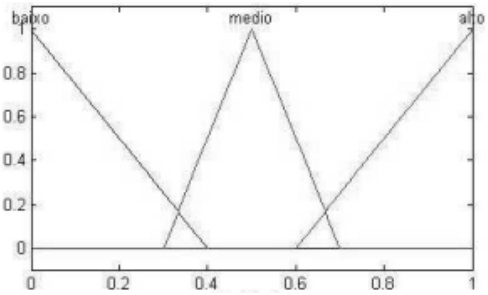

Fig. 4. Exemplo da variável de saída de Sociologia.

Tabela. 1. Composição da Base de Regras usando as variáveis de entrada (Termos de Sociologia - QtSoc, Termos de Computação - QtCom, Termos de Direito - QtDir, Termos de Odontologia - QtOdo e Termos de Química -QtQui) e de saída (Pertencente a Sociologia - Socio, Pertencente a Computação - Comp, Pertencente a Direito - Dirt, Pertencente a Odontologia - Odont, Pertencente a Química - Quim e Pertencente a Outros Assuntos - Outros).

Se					Então					
Entrada					Saída					
QtSoc	QtCom	QtDir	QtOdo	QtQui	Socio	Comp	Dirt	Odont	Quim	Outros
muito	-	-	-	-	alto	-	-	-	-	baixo
-	muito	-	-	-	-	alto	-	-	-	baixo
-	-	muito	-	-	-	-	alto	-	-	baixo
-	-	-	muito	-	-	-	-	alto	-	baixo
-	-	-	-	muito	-	-	-	-	alto	baixo
muito	muito	-	-	-	medio	medio	-	-	-	baixo
muito	-	muito		-	medio	-	medio	-	-	baixo
muito	-	-	muito	-	medio	-	-	medio	-	baixo
muito	-	-	-	muito	medio	-	-	-	medio	baixo
-	muito	muito	-	-	-	medio	medio	-	-	baixo
-	muito	-	muito	-	-	medio	-	medio	-	baixo
-	muito	-	-	muito	-	medio	-	-	medio	baixo
-	-	muito	muito	-	-	-	medio	medio	-	baixo
-	-	muito	-	muito	-	-	medio	-	medio	baixo
-	-	-	muito	muito	-	-	-	medio	medio	baixo
medio	-	-	-	-	medio	-	-	-	-	baixo
-	medio	-	-	-	-	medio	-	-	-	baixo
-	-	medio	-	-	-	-	medio	-	-	baixo
-	-	-	medio	-	-	-	-	medio	-	baixo
-	-	-	-	medio	-	-	-	-	medio	baixo
pouco	-	-	-	-	baixo	-	-	-	-	medio
-	pouco	-	-	-	-	baixo	-	-	-	medio
-	-	pouco	-	-	-	-	baixo	-	-	medio
-	-	-	pouco	-	-	-	-	baixo	-	medio
-	-	-	-	pouco	-	-	-	-	baixo	medio
muito	muito	muito	muito	muito	medio	medio	medio	medio	medio	baixo
pouco	pouco	pouco	pouco	pouco	baixo	baixo	baixo	baixo	baixo	alto

O conjunto de pertinência das variáveis de entrada foram definidos considerando que os textos a serem categorizados tenham até 500 caracteres. E as regras foram definidas considerando somente a quantidade de termos de cada categoria no texto.

E. Conclusão do Processo

Ao final da categorização do método de pontuação geral e do método de relevância é aplicado o método de lógica difusa para determinar a qual categoria o texto realmente pertence.

IV. PROCEDIMENTO EXPERIMENTAL

Com a finalidade de mostrar o funcionamento do algoritmo de análise textual, foi implementada uma ferramenta que ilustra graficamente os resultados obtidos com a aplicação dos métodos desenvolvidos. O objetivo desta seção é explicar essa ferramenta.

A Subseção A apresenta como o texto é formatado após a análise. A Subseção B mostra os dois gráficos obtidos pela ferramenta baseados nos métodos citadas nas Subseções III A e III B. A Subseção C apresenta explicitamente como o texto foi analisado mostrando as palavras e sua pontuação em cada uma das categorias. E finalmente, a Subseção D constrói uma árvore de raízes de palavras analisadas após a fase de preparação do texto.

A. Entrada de Texto

A Figura 5 ilustra como os textos formatados ficam após a preparação. Esta caixa de texto é apresentada após seu pré-processamento.

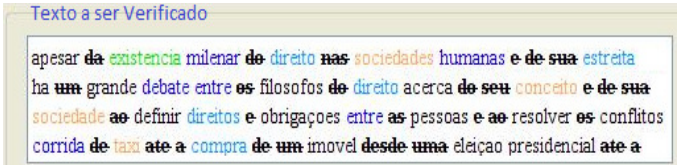

Fig. 5. Reconhecimento do texto.

As *stopwords* são mostradas riscadas e com a cor preta. Cada categoria possui uma cor pré-definida: computação é azul escuro, química é vermelho, direito é azul claro, odontologia é verde e finalmente sociologia é laranja. As palavras que possuem relevância para determinada categoria aparecem com a cor definida pela categoria. As palavras que não possuem nenhuma categoria são apresentadas na cor preta. Assim, é fácil identificar a categoria de cada palavra e quais palavras são *stopwords*.

B. Gráficos

Os gráficos de pizza e de barras são obtidos a partir dos métodos aplicados. A Figura 6 ilustra o gráfico de pizza. Este gráfico mostra a porcentagem das palavras pertencentes a cada categoria.

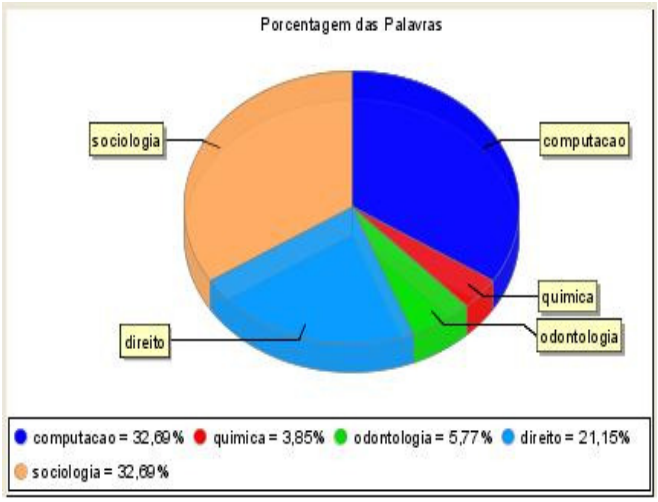

Fig. 6. Porcentagem de palavras no texto.

A Figura 7 ilustra o gráfico de barras. Este gráfico mostra quantos pontos cada categoria obteve aplicando o método de pontuação geral em um texto.

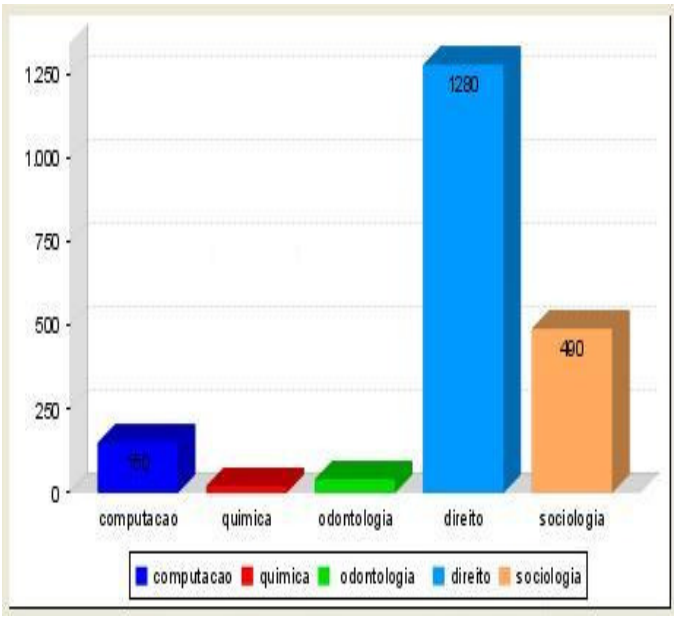

Fig. 7. Pontuação das categorias conforme o texto classificado

C. Pontuação dos Grupos

A Figura 8 apresenta a pontuação por categorias e mostra todas as palavras analisadas e qual foi a pontuação atingida por categoria.

Classificação de Dados | Pontuação dos Grupos | Raizes e relações de Palavras

Nº	Raiz	Computação	Direito	Biologia	Filosofia	Sociologia
2	cham	0	0	0	0	0
4	tenh	0	0	0	0	0
6	condut	0	0	0	10	10
8	sent	0	0	0	0	0
10	norm	0	0	0	30	30
12	instituiç	0	0	0	0	10
14	poligam	0	0	0	10	0
16	jurist	0	0	0	20	0
18	fav	0	0	0	0	0
20	signific	0	0	0	0	10
22	fal	0	0	0	0	0
24	tinh	0	0	0	0	0
26	determin	0	0	0	20	40
28	ramificaç	0	0	0	0	0
30	ord	0	0	0	80	20
32	diz	0	0	0	0	0
34	subj	0	0	0	0	0
36	portugu	0	0	0	0	0
38	sistem	1560	0	0	20	60
40	terr	0	0	0	0	0
42	pal	0	0	0	0	0
44	penal	0	0	0	10	0
46	conced	0	0	0	20	0
48	pode	0	0	0	0	0
50	impost	0	0	0	10	0
1	conflit	0	0	0	20	40
2	defin	0	0	0	0	0
3	taxi	0	0	0	0	0
4	simpl	10	0	0	0	0
5	ultim	0	0	0	0	0
6	essenc	0	0	0	10	0

Fig. 8. Pontuação por Categoria.

D. Raízes e Relação de Palavras

Na ferramenta desenvolvida, a aba "Raízes e Relação de Palavras" mostra as palavras após o processo de *stemming*. A Figura 9 ilustra essa aba, mostrando a forma primitiva da palavra e a sua forma natural contida no texto.

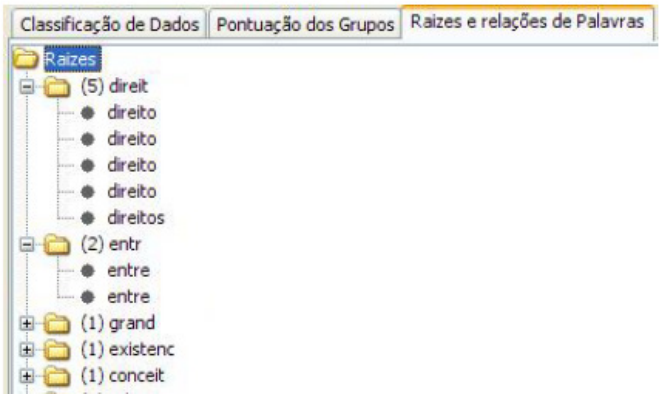

Fig. 9. Raízes e relação de palavras

V. ESTUDO DE CASO

Nesta seção é apresentado um estudo de caso realizado com a ferramenta desenvolvida. A Subseção A descreve como os textos foram escolhidos e como a ferramenta desenvolvida foi aplicada. A Subseção B ilustra os resultados obtidos com a utilização dessa ferramenta. A Subseção C apresenta a conclusão sobre os resultados obtidos nos testes realizados.

A. Aplicação da Ferramenta Desenvolvida

Para utilizar a ferramenta foram selecionados 10 textos com mais de 500 caracteres e 10 com menos de 500 caracteres, totalizando 20 textos de cada categoria. Esses textos foram retirados de fóruns de discussões, artigos e vários sites. Com isso pode-se verificar a usabilidade dos métodos desenvolvidos na classificação dos textos.

B. Resultados Obtidos

Primeiro, a ferramenta desenvolvida foi aplicada a textos com mais de 500 caracteres. O resultado obtido foi de 80% de acerto, ou seja, a cada 10 textos analisados essa ferramenta identificou corretamente 8 textos. A Figura 10 mostra o resultado dessa análise.

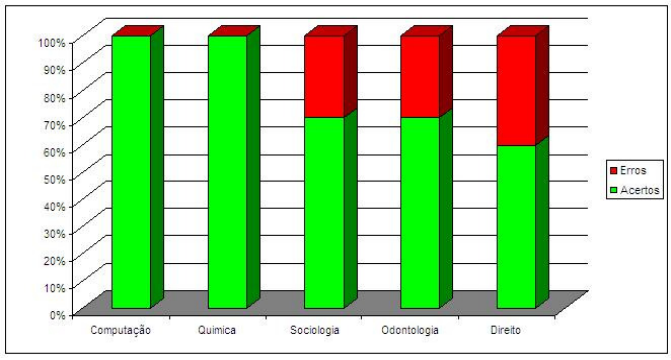

Fig. 10. Análise dos textos com mais de 500 Caracteres.

Na Figura 10 e 11 a barra verde e vermelha representam a quantidade de acertos e a quantidade de erros, respectivamente.

Em seguida, a ferramenta desenvolvida foi aplicada a textos com menos de 500 caracteres. O resultado obtido foi de 82% de acerto, ou seja, a cada dez textos analisados, oito textos foram classificados corretamente. A Figura 11 ilustra o resultado obtido nesta análise.

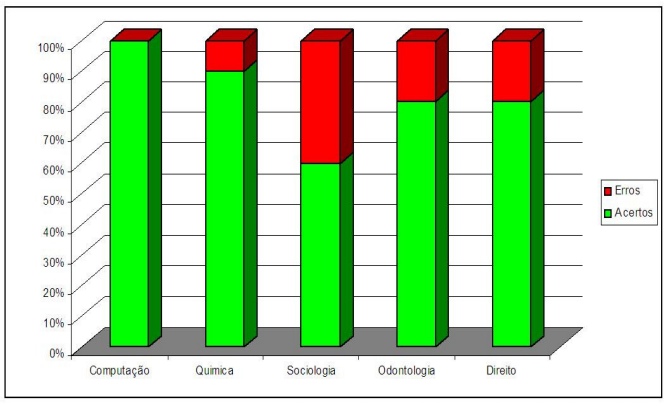

Fig. 11. Análise dos Textos com menos de 500 Caracteres.

C. Conclusão dos Resultados

Os resultados obtidos mostram que a porcentagem de acertos da ferramenta desenvolvida é de 80% . Este resultado mostra que ela pode ser utilizada na classificação de textos considerando as categorias determinadas e a base de dados construída.

VI. CONSIDERAÇÕES FINAIS

Esta seção apresenta as considerações finais sobre as dificuldades de implementação e os projetos futuros. A Subseção A apresenta quais foram as maiores dificuldades na implementação dos métodos e da ferramenta desenvolvida. E A Subseção B analisa algumas implementações futuras dessa ferramenta.

A. Dificuldades de Implementação

Durante a implementação da ferramenta desenvolvida a parte mais difícil foi encontrar os textos corretos para construir a base de dados. Isto porque, quanto mais específicos são os textos melhor é a base de conhecimento formada e maior é o grau de precisão de sua classificação.

B. Projetos Futuros

Com o objetivo de aumentar a precisão da ferramenta pode-se elaborar novos métodos heurísticos.

Com a finalidade de validar a qualidade de classificação da ferramenta desenvolvida neste trabalho, planeja-se utilizá-la em uma maior quantidade de textos. E finalmente aplicar os métodos de classificação diretamente em algum fórum universitário.

VII. CONCLUSÕES

Ao aplicar os métodos de classificação, desenvolvidos neste trabalho, os textos do estudo de caso foram classificados corretamente. Portanto, heurísticas simples combinadas com lógica difusa foi uma solução viável para categorização de textos em um ambiente controlado.

Apesar do índice de acerto ser de 80%, é importante notar que a ferramenta foi aplicada a uma quantidade baixa de textos. Assim, deve-se levar em consideração uma margem de erro que pode ser gerada pela ferramenta. Porém esses erros podem ser minimizados com a melhoria dos métodos de classificação e com a adição de novos métodos.

REFERÊNCIAS

[1] T. S. Galho, S. M. W. Moraes, "Categorização Automática de Documentos de Texto utilizando Lógica Difusa", *In: WORKCOMP SUL*, 1., 2004, p. 91-104.

[2] V. M. Orengo, C. R. Huyck, "A Stemming Algorithm for The Portuguese Language", *In: CONFERENCIA LATINOAMERICANA DE INFORMÁTICA*, 30., 2001, p. 13-15.

[3] R. Balinski, *Filtragem de Informações no Ambiente do Direito*. 87f. Dissertação (Mestrado em Informática) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

[4] M. C. S. Lopes, *Mineração de Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português*. Tese (Doutorado em Informática) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

[5] V. N. Puchkin, *A Ciência do Pensamento Criador*, São Paulo, Zahar Editores, 1969.

[6] Baharudin, B., Lee, L. H., & Khan, K. *A review of machine learning algorithms for text-documents classification.* Journal of advances in information technology,2010 ,p. 4-20.

[7] Yeh, A. S., Hirschman, L., & Morgan, A. A. *Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup, Bioinformatics*, Vol. 19 Suppl, 2003, i331-i339.