XIII Encontro Nacional de Pesquisa em Ciência da Informação - XIII ENANCIB 2012

GT8 - Informação e Tecnologia

O IMPACTO DA VARIAÇÃO TEMÁTICA NA CATEGORIZAÇÃO AUTOMÁTICA DE ARTIGOS CIENTÍFICOS EM PORTUGUÊS DO BRASIL

Modalidade de Apresentação: Comunicação Oral

Alexandre Ribeiro Afonso - UnB

Cláudio Gottschalg Duque – UnB

rafonso.alex@gmail.com

RESUMO

Nesta pesquisa, é verificado o impacto da variação de áreas científicas nos corpora textuais de entrada para um sistema automático de categorização textual. Foi medida a efetividade de três algoritmos de categorização textual, também considerando as características linguísticas do português do Brasil em tais textos. Observou-se que a presença de artigos científicos de uma mesma grande área, em um corpus de teste, causa uma queda de efetividade considerável sobre os algoritmos de categorização, mas para alguns corpora, mesmo contendo artigos de uma mesma grande área, a queda de efetividade não é acentuada. Considerando os experimentos realizados, é possível inferir que específicas combinações de áreas científicas dentro do corpus de teste produzem resultados específicos de categorização. Conclui que a efetiva categorização automática depende de vários fatores, inclusive das características do corpus de entrada, e não somente dos algoritmos de pré-processamento e categorização.

Palavras-chave: Categorização Automática de Textos. Português do Brasil. Efetividade. Artigos Científicos. Bibliotecas Digitais.

1 INTRODUÇÃO

O desenvolvimento de métodos para a classificação de textos científicos, tendo por objetivo a efetiva[1] recuperação da informação por parte do usuário, é um campo de estudos chave da Ciência da Informação desde seu surgimento (BARRETO, 2008).

Considerando o atual cenário tecnológico, onde as publicações científicas cada vez mais se encontram em meio eletrônico e armazenadas em bibliotecas digitais, observa-se que o usuário exige uma efetiva recuperação da informação no acesso on-line a este material, e a classificação prévia destes registros para tal recuperação também deve ser efetiva, afinal, o objetivo da automação é ganhar tempo sem perda de qualidade no serviço.

A proposta de um sistema de classificação automática é ser capaz de classificar textos sem intervenção ou com pouca intervenção humana, reunindo os documentos em classes específicas (por exemplo, por área científica) de acordo com o seu conteúdo. Para tal atividade, algoritmos e técnicas de Inteligência Artificial (Aprendizagem de Máquinas) têm sido desenvolvidas e implementadas em software, sendo este um amplo tópico de interesse na Ciência ou Engenharia da Computação, na Ciência da Informação, Linguística Computacional e nas áreas que trabalham com tecnologia da informação em geral.

Apesar de toda essa revolução tecnológica, deve-se considerar que a Ciência da Informação tem um forte caráter social (SARACEVIC, 1996). Logo, verificar até que ponto esta tecnologia proposta contribui para a organização da informação nas bibliotecas digitais

[1] Recuperação ou Classificação efetiva diz respeito à recuperação ou classificação da informação registrada de forma ágil e correta em termos de resultados, do ponto de vista do usuário.

brasileiras, e para o processo de recuperação da informação por parte do usuário, é tão importante quanto o desenvolvimento do software classificador. Observando que os textos científicos têm características linguísticas peculiares para cada língua e com terminologia científica nacional própria (BIDERMAN, 2006), a experimentação sobre o desempenho dos sistemas classificadores, considerando tais peculiaridades sociolinguísticas, deve ser exercida.

Levando em consideração os aspectos linguísticos do português do Brasil e da terminologia das ciências no Brasil, a ênfase deste trabalho é mensurar a efetividade de classificação com diferentes combinações de áreas científicas nos *corpora* de teste (com os *corpora* de teste contendo áreas científicas de uma mesma grande área e de diferentes grandes áreas).

Para tais experimentos, algoritmos e ferramentas computacionais propostas pela tecnologia atual para processamento de texto (etiquetadores morfossintáticos, sistemas de radicalização e classificadores de textos) foram utilizadas.

2 ESTUDOS NA ÁREA DE CLASSIFICAÇÃO TEXTUAL AUTOMÁTICA E REVISÃO BIBLIOGRÁFICA

A área de Organização e Recuperação Automática da Informação (incluindo sistemas automáticos de classificação de textos) tem sido desenvolvida nacionalmente, para o português do Brasil, em diversas áreas: Linguística, Ciência da Computação e Ciência da Informação.

Pelo fato do processamento computacional linguístico estar altamente relacionado ao processo de organização e recuperação automática da informação (DUQUE, 2005), esta área tem sido desenvolvida no Brasil principalmente após a criação do Núcleo Interinstitucional de Linguística Computacional - NILC em 1993 (NUNES; ALUÍSIO; PARDO, 2010). O NILC tem criado diversos recursos linguístico-computacionais (*corpora*, etiquetadores morfossintáticos, analisadores sintáticos, etc.) e estudos descritivos sobre o português do Brasil, os quais servem como módulos de apoio imprescindíveis para a construção de sistemas automáticos de organização e recuperação da informação textual.

Na Ciência da Informação têm surgido cada vez mais pesquisas, abrangendo Informação e Tecnologia, Organização e Representação do Conhecimento e Bibliotecas Digitais, com estudos voltados ou ligados ao português do Brasil ou para a realidade social brasileira, envolvendo o acesso à informação digital. Para o tema, são notáveis publicações e

workshops realizados em periódicos e eventos brasileiros de Ciência da Informação (Periódicos: Ciência da Informação, Datagramazero, Perspectivas em Ciência da Informação, Eventos: ENANCIB, *Workshop Internacional de Arquitetura da Informação e Multimodalidade*, *Texto e Imagem*, entre outros).

2.1 Sistemas Automáticos de Classificação Textual: Definições e Estudos na Área

A Categorização Automática de Textos e o Agrupamento Automático de Textos são duas áreas distintas dentro das possibilidades de Classificação Automática de Textos[2].

A categorização utiliza conhecimento formal (ontologias, tesauros, vocabulários controlados ou pré-treinamento) conjuntamente com o sistema inteligente de classificação, ou seja, algum conhecimento formalizado é utilizado pelo sistema para auxiliar a classificação. O agrupamento não utiliza nenhuma entrada auxiliar de conhecimento formal externa para agrupar os textos de entrada, o sistema geralmente utiliza algum mecanismo algorítmico de aprendizagem (meta-heurísticas, redes neurais autoadaptativas, aprendizagem indutiva, etc.) para classificar os documentos on-line (MANNING; RAGHAVAN; SCHÜTZE, 2008), e ainda, pode ser necessário que o sistema descubra o número de grupos de textos a serem criados, ou esse dado numérico pode ser cedido pelo usuário ao sistema agrupador.

Os sistemas de categorização (alvo de estudo da pesquisa aqui descrita) geralmente possuem taxas de acerto de classificação melhores que o agrupamento, pelo fato de utilizarem algum conhecimento formalizado sobre as classes possíveis de classificação, cedido pelo usuário, de forma supervisionada. Neste caso, o usuário conhece a natureza dos textos, as classes em que os textos serão classificados, e logo, pode treinar o sistema ou inserir conhecimento formal para auxiliar o software classificador.

Anteriormente à fase de classificação, seja agrupamento ou categorização, os textos podem passar por um pré-processamento, o objetivo é representar cada texto do *corpus* com palavras-chave ou termos que tenham peso semântico e que contribuam com a identificação do tema para aquele texto. Isso significa que os textos são representados por índices antes de serem classificados. No caso deste trabalho, a indexação é por características linguísticas e algumas possibilidades de escolha de termos são testadas.

[2] Alguns autores utilizam o termo Classificação Automática de Textos como sinônimo do termo Categorização Automática de Textos, aqui foi colocado o termo Classificação como o ato de "colocar em classes", seja Agrupamento ou Categorização.

Como observado na literatura (veja as referências a seguir), a efetividade de classificação depende de vários fatores: a língua, a natureza dos textos a classificar (jornalístico ou científico), a técnica de pré-processamento (ou indexação) utilizada para representação dos textos e os algoritmos[3] de classificação. Nesta pesquisa, foram construídos diferentes *corpora* de teste onde cada *corpus* contém combinações de áreas científicas distintas, e diferentes formas de indexação linguística com diferentes algoritmos de categorização foram aplicados aos *corpora*. A questão principal, portanto, é: saber quais os melhores resultados dessas diferentes combinações de *corpora* com técnicas de indexação e algoritmos de categorização, em relação à efetividade classificatória.

Em relação aos estudos desenvolvidos relacionados à classificação automática de textos, são identificados poucos estudos para o português do Brasil, e há trabalhos realizados no Brasil que analisam a efetividade de categorização de textos em inglês. A grande parte dos trabalhos visa testar os efeitos de diferentes técnicas algorítmicas de pré-processamento ou indexação (preparação dos textos e escolha de termos chave antes da classificação), ou testar o desempenho de diferentes algoritmos de classificação. A influência da variedade científica no *corpus* de teste (o alvo desta pesquisa) para a efetividade de classificação é praticamente inexplorada. Abaixo, listamos os trabalhos sobre classificação relacionados ao estudo aqui descrito.

Maia e Souza (2010) descrevem um estudo sobre agrupamento de textos e outro em categorização de textos em português do Brasil, sendo os mesmos *corpora* utilizados nas duas situações. A pesquisa procurou verificar qual a melhor forma de representação linguística de textos na fase de pré-processamento, antes das realizações de classificação em si. Para tais formas de representação textual, os autores testaram sintagmas nominais e termos simples, verificando se o uso de sintagmas nominais seria mais produtivo para a construção automática de grupos ou categorias com menos erros de classificação. O algoritmo de agrupamento utilizado foi o *Simple K-Means* (SKM) e o algoritmo de categorização com treinamento prévio foi o *Naive Bayes*. Os *corpora* descritos por Maia e Souza (2010) são jornalísticos e científicos. Os resultados apontam que os métodos que envolvem sintagmas nominais na classificação apresentam índices semelhantes ao dos termos sem *stopwords*. Por exemplo, no *corpus* jornalístico os sintagmas nominais e os termos sem *stopwords* atingiram os mesmos

[3] Neste contexto, Algoritmo refere-se a: dada uma entrada ou input uma sequência lógica de passos é executada para atingir um objetivo final, com uma saída ou output.

resultados com o algoritmo de categorização *Naive Bayes*, obtendo 91% de classificação correta. Como o uso de sintagmas nominais exige um processamento bem maior na categorização e agrupamento de documentos, a relação custo *versus* benefício de se utilizar sintagmas nominais não se apresentou interessante.

DaSilva, Vieira, Osório e Quaresma (2004) propõem e avaliam o uso da informação linguística na fase de pré-processamento nas tarefas de mineração de textos (agrupamento e classificação) em português do Brasil. Eles apresentam diversos experimentos, comparando as suas propostas para seleção de termos baseadas em conhecimento linguístico com técnicas usuais aplicadas no campo de agrupamento/categorização. O estudo mostra que o uso de informações linguísticas, no caso, identificando as classes de palavras (*part-of-speech information*), para localizar termos dos textos, é útil durante a fase de pré-processamento, antes da categorização e do agrupamento textual, como alternativa para o simples uso de radicalização (extração dos radicais dos temos) e a retirada de palavras sem peso semântico antes da classificação (preposições, artigos, interjeições, etc.).

Camargo (2007) aborda a categorização automática de textos em português com o uso de informações linguísticas na etapa de pré-processamento dos textos. Para isso, são utilizadas duas coleções de textos, sendo uma composta de artigos de jornal (Folha de São Paulo) e outra composta de textos científicos (Resumo e Título de Teses e Dissertações da COPPE/UFRJ, 2000 a 2006). Os algoritmos de categorização utilizados são o *Naive Bayes*, a Máquina de Vetor Suporte e um sistema baseado em regras de decisão. Os resultados obtidos comprovam que o *Naive Bayes* oferece excelentes resultados na classificação de textos e mostra que os outros dois métodos podem ser usados na classificação em combinação com o primeiro, oferecendo resultados ainda melhores.

Figueiredo (2008) trabalha na etapa de pré-processamento dos textos em Inglês e propõe uma estratégia de tratamento de dados, baseada em extração de características, que precede a tarefa de classificação, a fim de introduzir em documentos características discriminativas de cada classe capazes de melhorar a eficácia da classificação. Resultados experimentais demonstram ganhos em quase todos os cenários testados, desde os algoritmos de classificação mais simples até o algoritmo mais complexo, estado da arte.

3 PROBLEMÁTICA, OBJETIVOS E ESTRATÉGIA DE PESQUISA

Como descrito anteriormente, a efetividade da classificação automática de textos depende de uma série de fatores: a língua, as áreas científicas dos artigos no *corpus* a classificar, a forma de pré-processamento adotada (ou indexação) para representar os textos, e os algoritmos de classificação. No caso desta pesquisa, o interesse maior é verificar a variação do *corpus*, considerando a diversidade de áreas científicas presentes neste *corpus*. Porém, experimentaram-se variações nas escolhas dos algoritmos de categorização e na indexação linguística, visando encontrar as melhores combinações globais em termos de efetividade.

A língua adotada nos testes é o português do Brasil. Com artigos científicos escritos nesta língua somente.

Quatro *corpora* foram testados (categorizados), cada um contendo textos de áreas científicas específicas: *Corpus* 1 (Linguística, Farmácia e Educação Física), *Corpus* 2 (Linguística, Farmácia, História, Geografia, Educação Física), *Corpus* 3 (Linguística, Farmácia, Odontologia, Geografia, Educação Física), *Corpus* 4 (Linguística, Farmácia, Odontologia, Geografia, Educação Física, História), sendo que para cada área científica foram escolhidas as mesmas quantidades de 19 (dezenove) textos para compor os *corpora*.

A indexação, a realizada neste trabalho, é executada considerando características linguísticas dos textos. Nesta pesquisa, foram testados em cada *corpus* quatro conjuntos possíveis para a seleção de termos com base na informação linguística: Somente Substantivos (S), Substantivos, Verbos e Adjetivos (S, V, ADJ), Substantivos e Adjetivos (S, ADJ), Adjetivos e Verbos (ADJ, V). O objetivo é verificar, para cada algoritmo de categorização testado, o melhor conjunto de classes de palavras para representar o texto. Ainda, para diminuir e selecionar ainda mais o conjunto representativo, cada termo escolhido deve aparecer no *corpus* com uma frequência mínima de 5 (cinco) ocorrências, esta restrição numérica contribui para evitar a captura de termos com baixa significância semântica.

Após o procedimento de indexação de cada texto de cada *corpus*, os algoritmos de categorização foram treinados com uma parte do *corpus* (65%) e testados com outra parte restante (35%). Foram utilizados três algoritmos de categorização: as Redes Neurais (RN) do tipo *Multi-Layer Perceptrons* (MLP), as Redes Bayesianas (RB) e a Classificação Arbórea J48. Todos estes algoritmos de categorização já são clássicos na área e estão implementados no pacote (software) de livre acesso *WEKA*[4], também clássico nas pesquisas em classificação de textos.

[4] http://www.cs.waikato.ac.nz/ml/weka/

Como existem quatro *corpora* distintos, cada um contendo áreas científicas diversas, quatro experimentos foram realizados, cada experimento para um *corpus* específico. Após a coleta dos resultados de efetividade para cada experimento (tempo de execução e corretude de classificação), eles foram comparados em termos de efetividade.

Observe que em cada experimento existem as seguintes combinações possíveis: O *corpus*, as escolhas linguísticas para indexação e o algoritmo de classificação. A figura abaixo ilustra dois dos experimentos possíveis para tais combinações:

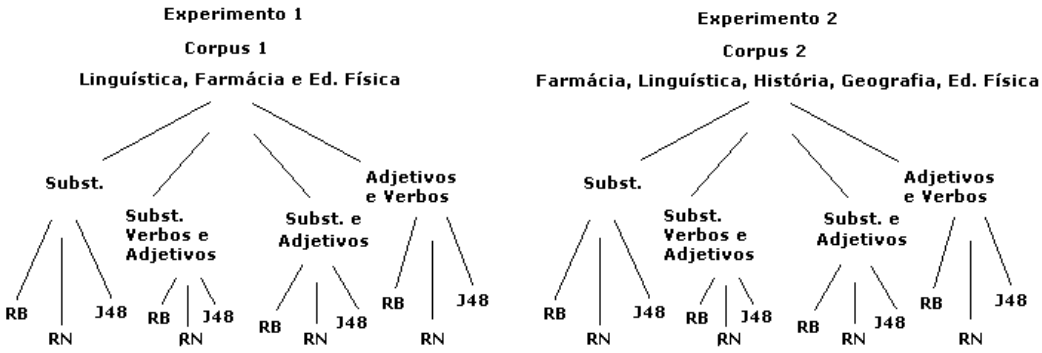

Figura 1-Exemplos de possibilidades combinatórias para cada experimento.

4 QUESTÕES DE PESQUISA

Objetivou-se colher os seguintes dados, ao analisar os resultados dos quatro experimentos:

- *Questão 1:* Qual a melhor combinação (em termos de efetividade) dos Algoritmos de Categorização com Indexação Linguística para cada *corpus*?
- *Questão 2:* Em termos de efetividade, é possível identificar uma melhor Indexação Linguística e um melhor Algoritmo de Categorização dentre os melhores resultados da Questão 1?
- *Questão 3:* Considerando os resultados da Questão 1, há alterações de efetividade variando a natureza científica dos textos em cada copus?

5 METODOLOGIA E DESCRIÇÃO DAS ETAPAS DOS EXPERIMENTOS

Os experimentos seguem sempre o mesmo modelo, ou seja, os mesmos passos sequenciais foram realizados igualmente para todos os quatro experimentos. Tal modelo segue em volta dos trabalhos relacionados descritos na literatura. A figura 2, a seguir, ilustra esses passos sequenciais.

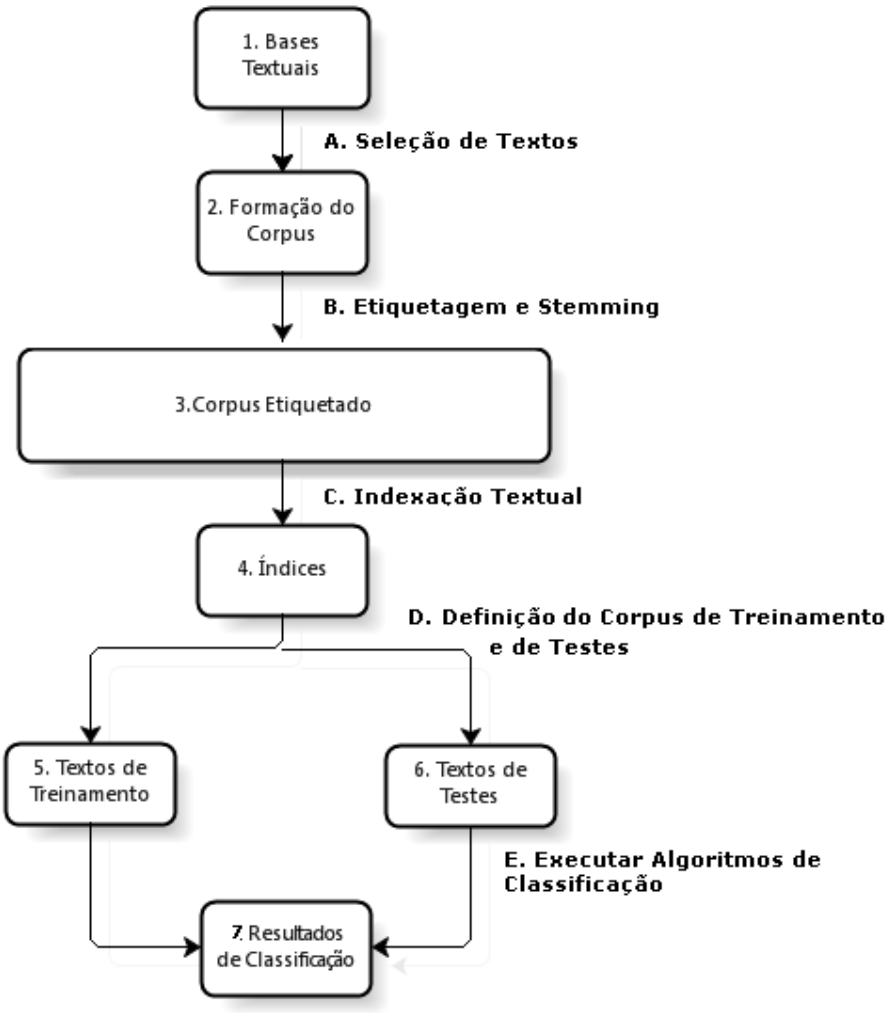

Figura 2- Modelo dos experimentos realizados durante a pesquisa.

As anotações ao lado das setas com uma marca alfabética (A. B. C. D. E.) indicam um processo manual ou algorítmico, os retângulos com marca numérica (1. 2. 3. 4. 5. 6. 7.) indicam um conjunto de dados recebidos ou produzidos pelos processos (A. B. C. D. E.).

5.1 Descrições dos Processos Experimentais

A seguir, as descrições dos processos (A,B,C,D,E) esquematizados na figura 2:

A. Seleção de Textos:

É um processo manual onde os textos são selecionados das bases textuais. As bases textuais são bibliotecas digitais das universidades brasileiras que mantêm periódicos científicos publicados pela própria universidade. Para este trabalho, foram retirados artigos das bases das universidades brasileiras: UFG, PUC Minas, UFOP, UFES, UPF, UNESP. Todas estas universidades permitem o *download* dos periódicos científicos utilizados, livremente.

Após a seleção dos artigos, que é totalmente aleatória e sem repetições de escolha de textos, os quatro *corpora* foram reunidos, cada um contendo uma combinação diferente de áreas científicas. A escolha dos periódicos foi realizada objetivando a seleção por área de conhecimento, verificando a descrição dos objetivos informativos dos periódicos. Portanto, evitou-se a escolha de periódicos que publicassem artigos envolvendo grandes áreas, segundo a Tabela de Áreas de Conhecimento do CNPQ. Por exemplo, alguns periódicos publicam artigos da grande área: Ciências da Saúde, e aceitam artigos de várias áreas (Farmácia, Odontologia, Medicina, etc.), tais periódicos foram evitados, porém, há o fato que a interdisciplinaridade existe nas áreas de conhecimento. Quando os artigos foram colhidos de um mesmo periódico, procurou-se variar o número da publicação no ato de escolha de artigos, pois alguns deles trabalham seus números de forma temática (em cada número um tema específico é abordado) e isso poderia levar a resultados tendenciosos na categorização automática.

Seguinte a esta etapa, manualmente, para cada artigo, foram extraídos somente: Título, Resumo, Palavras-Chave e a primeira página da introdução e, assim, os quatro *Corpora* de Testes foram formados, como está esquematizado na figura 2 descrita.

B. Etiquetagem e Stemming:

A Etiquetagem consiste no processo de etiquetar cada palavra[5] de cada artigo do *corpus*, colocando suas etiquetas morfossintáticas (Substantivo, Adjetivo, Verbo, Preposição,

[5] Neste caso, o conceito de Palavra considerado é qualquer sequência de símbolos entre espaços em branco no texto.

Advérbio, etc.). Foi utilizado o software *MXPOST* para tal tarefa, na sua versão para o português do Brasil, descrito por Aires (2000). Feito isso, ocorre o processo de *stemming* que retira o sufixo de cada palavra (por exemplo, as palavras *aluno, aluna e alunos* tornam-se o radical *alun*). Para o processo de *stemming*, foi utilizado o software *STEMMER* específico para o português do Brasil (CALDAS; IMAMURA; REZENDE, 2001).

C. Indexação Textual:

Para os testes realizados, optou-se trabalhar apenas com as classes de palavras (substantivo, verbo ou adjetivo) dos textos de um *corpus*, pelo fato destas classes terem maior valor semântico e contribuiriam mais para o processo de categorização. Além disso, cada termo (substantivo, verbo ou adjetivo) de cada texto deve aparecer com frequência mínima de valor 5 (cinco) no *corpus* para ser incluído no Índice de Representação de cada texto. O objetivo também é capturar termos com peso semântico considerável, com maior frequência no *corpus*, e evitar termos que inserem ruído na representação. A seleção dos termos por etiqueta morfossintática e frequência no *corpus* foi realizada pelo pacote *WEKA*.

Após esta seleção dos termos e criação do Índice de Representação para cada texto, é possível adicionar um peso a cada termo do Índice de Representação para auxiliar a categorização a ser realizada pelos algoritmos de categorização. Neste caso, foi utilizado um fator numérico que indica o valor semântico do termo e o quanto ele é significativo para aquele texto a que pertence, considerando ainda, a significância deste termo para o texto em relação a todo o *corpus* de teste. Este fator é denominado *IDF* (*Inverse Document Frequency)-Transform* (MARKOV; LAROSE, 2007):

IDF-T = fij*log (número de documentos no corpus/número de documentos com o termo i no corpus) (I)

fij é a frequência do termo *i* no documento *j*.

D. Definição do Corpus de Treinamento e de Testes:

Para cada experimento, o *corpus* deve ser dividido em duas partes, uma parte para treinar o algoritmo de classificação, e outra parte para verificar a efetividade do algoritmo pós-treino. No caso desta pesquisa, optou-se pelo valor de 65% do *corpus* para treinamento e o restante 35% para teste, nos quatro experimentos.

E. Executar Algoritmos de Classificação:

Foram utilizados três algoritmos de categorização em cada experimento (Redes Neurais MLP, Árvores de Decisão - J48, Redes Bayesianas). Após o treinamento, o algoritmo é testado com a parte de testes do *corpus*. Após os testes, a coleta de resultados de efetividade (corretude de classificação e tempo de execução) é realizada. O software *WEKA* foi utilizado no processo de treinamento e testes.

Como o objetivo desta pesquisa não é verificar as estratégias matemáticas e computacionais dos algoritmos de classificação, mas sim verificar a efetividade destes algoritmos para um conjunto de textos com características linguísticas e de formato específicas, não entraremos em detalhes sobre tais peculiaridades computacionais de funcionamento para cada algoritmo. Para uma descrição sobre a implementação das Redes Bayesianas veja (BOUCKAERT, 2008), para uma descrição das Redes Neurais MLP veja (JAIN; MAO; MOHIUDDIN, 1996) e para o classificador J48 (ROOS, 1993).

Para os algoritmos de categorização: Redes Bayesianas e J48, no treinamento, foram utilizados os valores *default* de configuração do software *WEKA*, apenas nas Redes Neurais MLP o número de camadas escondidas foi estabelecido com 40 camadas e o número de ciclos de treinamento foi estabelecido para 20.000 iterações. Estes valores foram escolhidos empiricamente, por apresentarem melhores resultados. Verificou-se no treinamento que quanto maiores estes dois valores: número de camadas escondidas e ciclos de treinamento, maior era a taxa de acertos da rede neural na classificação, porém, estes valores escolhidos já ultrapassavam o tempo em horas para treinamento da rede neural, logo, este limite numérico descrito de configuração foi estabelecido. Os demais atributos de configuração das redes neurais foram deixados *default*.

6 RESULTADOS ESTATÍSTICOS DE CATEGORIZAÇÃO TEXTUAL

6.1 Resultados de Corretude na Categorização Textual

Para cada um dos quatro *corpora* (quatro experimentos) foram testadas quatro formas de Indexação Linguística para cada texto: Substantivos (S), Substantivos, Verbos e Adjetivos (S,V,ADJ), Substantivos e Adjetivos (S,ADJ) e Verbos e Adjetivo (ADJ,V). Para cada *corpus* são testados três algoritmos de categorização: Redes Bayesianas (RB), Redes Neurais MLP (RN) e a Categorização por Árvores de Decisão J48. Logo, para cada experimento temos 12 possibilidades combinatórias de Indexação Linguística com Algoritmo de Categorização,

totalizando 48 execuções possíveis nos quatro experimentos. Abaixo, para cada experimento, são listadas as combinações que obtiveram *melhores* resultados percentuais de corretude na categorização.

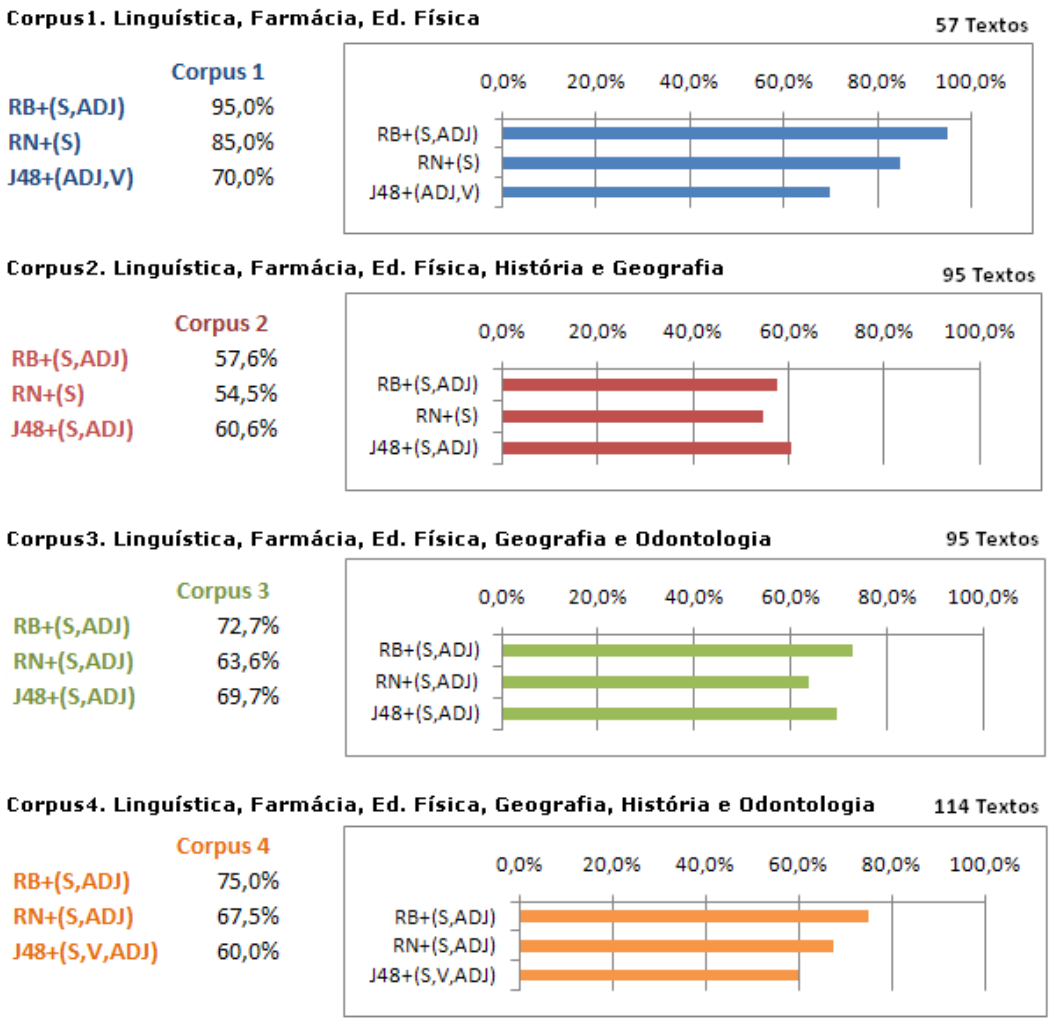

Figura 3-Melhores resultados estatísticos de corretude, em porcentagem de acertos na categorização, colhidos para os quatro corpora de teste e considerando os três algoritmos de classificação.

6.2 Resultados de Medição Temporal na Categorização Textual

Em relação à observação temporal, ou quanto tempo um algoritmo gasta para realizar o treinamento e efetuar os testes, somente foi medido o tempo das combinações listadas na figura 3, ou seja, somente as combinações (Indexação Linguística + Algoritmo) com melhores resultados de corretude para cada *corpus*. Em *todos* os testes de todos os quatro experimentos da figura 3, verificou-se que as combinações com os algoritmos Redes Bayesianas (RB) e

Árvores de Decisão J48 são as mais rápidas, tomando menos de um minuto para treinamento e teste. Já as combinações (Algoritmo + Indexação Linguística) com Redes Neurais MLP (RN) são lentas demais tanto no treinamento quanto no teste, como descrito a seguir: (*Corpus* 1: 26 minutos), (*Corpus* 2: 1 hora e 2 min.), (*Corpus* 3: 1 hora e 25 min.), (*Corpus* 4: 1 hora e 54 min.).

Como o tempo de execução da categorização depende muito da configuração do hardware, do sistema operacional, da versão do software classificador e de outras especificações de arquitetura computacional, são listadas as principais características do sistema computacional utilizado. Algumas variáveis, porém, não podem ser controladas nos experimentos pelo pesquisador e também podem afetar o tempo de computação de forma considerável, como por exemplo, o número de processos em execução, que são controlados automaticamente pelo sistema operacional da máquina.

- **Processador**: Pentium Dual Core T4500 Intel processor
- **Memória Principal**: 2GB
- **Sistema Operacional**: Microsoft Windows XP Professional 2002 SP 3
- **Versão *Weka***: 3.6.3
- **Java Development Kit (JDK)**: jdk1.6.0_04

7 ANÁLISE DE RESULTADOS DOS EXPERIMENTOS DE CATEGORIZAÇÃO TEXTUAL

O critério para escolher a melhor combinação de indexação com algoritmo de classificação, ou seja, a combinação mais efetiva, foi sempre o mesmo: escolhe-se a combinação que produz a mais alta taxa de acertos na categorização em cada experimento, quando há um empate escolhe-se a combinação que gasta menos tempo de computação, caso haja um novo empate a escolha da combinação mais efetiva é aleatória.

Questão 1: Qual a melhor combinação (em termos de efetividade) dos Algoritmos de Categorização com Indexação Linguística para cada *corpus*?

Como visto na figura 3, de cada copus testado, os *corpora* 1, 3 e 4 apresentam o mesmo resultado de combinação para o melhor resultado de efetividade (corretude e tempo). Os melhores resultados, para cada *corpus*, retirados da figura 3 são:

- *Corpus* 1: (RB+(S,ADJ)) – 95% de acertos em 1 minuto

- *Corpus* 2: (J48+(S,ADJ)) – 60,6% de acertos em 1 minuto
- *Corpus* 3: (RB+(S,ADJ)) – 72,7% de acertos 1 minuto
- *Corpus* 4: (RB+(S,ADJ)) – 75% de acertos em 1 minuto

Este resultado mostra que não houve um padrão combinatório único melhor de (Algoritmo de Classificação + Indexação Linguística) para todos os *corpora*, mas a partir da figura 3 vemos que o padrão (RB+(S,ADJ)) foi dominante como melhor resultado entre os quatro experimentos e a diferença entre os padrões: (RB+(S,ADJ)) e (J48+(S,ADJ)) no *corpus* 2 é de apenas 3 pontos percentuais.

Questão 2: Em termos de efetividade, é possível identificar uma melhor Indexação Linguística e um melhor Algoritmo de Categorização dentre os melhores resultados da Questão 1?

Da listagem acima (da Questão 1, a qual mostra os melhores resultados de categorização em cada *corpus*) identificamos que o algoritmo de categorização RB é o mais presente. No *corpus* 2 o algoritmo mais efetivo foi o J48, porém a diferença de corretude é de apenas 3 pontos percentuais em relação ao RB. Observa-se que há um padrão único para Indexação Linguística para todos os *corpora* testados, todos os testes apresentam o padrão (S,ADJ), ou substantivos e adjetivos, como escolha mais efetiva. Observa-se que as Redes Neurais MLP não aparecem nenhuma vez, o padrão de indexação linguística contendo substantivos somente (S) também não aparece, o padrão de indexação linguística (V,ADJ) e (S,ADJ,V) também não aparecem nos resultados com melhor medida de efetividade.

Questão 3: Considerando os resultados da Questão 1, há alterações de efetividade variando a natureza científica dos textos em cada copus ?

Existe uma variação de taxa de corretude de categorização notável nos resultados da Questão 1. Para o *corpus* 1 consegue-se uma taxa de acertos de 95%, porém ao inserir mais duas áreas científicas no *corpus* 1 e formar o *corpus* 2 essa taxa cai para 60,6% ou 34,4 pontos percentuais a menos. Já para o *corpus* 3, que também insere mais artigos de duas áreas científicas no *corpus* 1, a queda de acertos é de 22,3 pontos percentuais.

A diferença é que no *corpus* 2 são inseridos textos de História e Geografia e no *corpus* 3 são inseridos textos de Geografia e Odontologia. Isso nos leva a considerar a possibilidade

de que os textos de História e Geografia geram uma confusão maior para o Algoritmo de Categorização. Observe que no *corpus* 3 existem três áreas que pertencem à mesma grande área (Ed. Física, Farmácia e Odontologia), porém a queda percentual é menor que no *corpus* 2. O *corpus* 4 foi formado adicionando textos da área de Odontologia ao *corpus* 2, formando um *corpus* com seis áreas científicas, houve um aumento de acertos na corretude do *corpus* 4 em relação ao *corpus* 2, provavelmente pelo fato dos textos de Odontologia serem altamente identificáveis devido ao vocabulário técnico, o que aumentou a porcentagem de acertos em 14,4 pontos percentuais do copus 2 para o *corpus* 4.

Deve-se notar, porém, que no experimento do *corpus* 1 existem artigos da mesma grande área científica (Ed. Física e Farmácia) que geram um bom resultado de corretude: 95% de acertos. Ou seja, mesmo sendo os textos da mesma grande área, eles foram (na sua maior parte) categorizados pelo algoritmo corretamente.

8 CONCLUSÃO

Algumas pesquisas têm sido realizadas na área de categorização de textos para o português do Brasil. O enfoque da pesquisa descrita neste artigo é diferenciado.

Os estudos em categorização têm sido realizados objetivando medir a corretude dos algoritmos de categorização combinados às diversas formas de pré-processamento dos textos ou indexação. Porém, geralmente, o *corpus* é o mesmo para todos os algoritmos de testes e o tempo de treinamento e categorização não é formalmente medido. Optou-se em verificar os efeitos quando são inseridos novos textos de diversas áreas em um *corpus* de entrada. O tempo de execução do algoritmo de categorização, além da corretude da categorização, é uma variável que também foi considerada ao medir a qualidade dos resultados, a estas duas variáveis juntas chamamos “Efetividade de Categorização”.

De acordo com a hipótese inicial, a inserção de textos de áreas afins (ou da mesma grande área) confirma a diminuição da porcentagem de acertos na categorização de forma considerável, em alguns casos. A inserção de algumas áreas pertencentes a uma mesma grande área, como História e Geografia, gera uma diminuição de acertos na categorização mais acentuada do que outras áreas de uma mesma grande área, como: Educação Física, Farmácia e Odontologia, ao coexistirem no mesmo *corpus* de teste. A hipotética razão linguístico-terminológica para tal fato ocorrer (e que deve ser investigada futuramente) seria que algumas áreas científicas não possuem uma terminologia técnica e regular nos textos que

as caracterizem fortemente, ao ponto que os algoritmos de categorização possam identificar tais diferenças terminológicas e identificar a que classe tais textos pertencem.

Algumas combinações de algoritmos de categorização com a escolha de indexação linguística são melhores, em termos de efetividade, que outras, para quaisquer testes. Nesta pesquisa, os resultados apontam que o uso de Redes Bayesianas como algoritmo de categorização e o uso de (Substantivos e Adjetivos) é uma escolha genérica melhor (quando só uma combinação deva ser escolhida) tanto para melhores taxas de acerto como de tempo de categorização. O fato das Redes Bayesianas aqui apresentarem melhores resultados reforça as descrições da literatura sobre a potencialidade das Redes Bayesianas.

Apesar das pesquisas em classificação automática de textos enfatizarem o estudo de algoritmos de classificação e pré-processsamento, foi possível verificar que variações temáticas no *corpus* de teste também causam alto impacto no resultado de classificação, além de outros fatores.

Deve-se observar também, que foi utilizado um conjunto de combinações específicas, ou seja, utilizou-se uma amostra das diversas possibilidades combinatórias entre *corpora*, formas de indexação e algoritmos de categorização. Este campo de possibilidades combinatórias é amplo e, portanto, outros resultados diferentes dos aqui descritos ainda podem ser obtidos.

A investigação de particularidades terminológicas por área de conhecimento e elaboração de métodos computacionais, para melhor identificar automaticamente a temática dos textos científicos, podem prover a melhoria do nível de efetividade.

ABSTRACT

This study verified the impact of the area variations in scientific corpora used as input to text categorization systems. We measured the effectiveness of three text categorization algorithms, also considering the linguistic features of the Brazilian Portuguese language throughout the texts. We noticed that when the corpus has articles from the same major area, the effectiveness of the categorization algorithms decreases significantly, but for some corpora, even when the scientific texts belong to the same major area, the drop in effectiveness is not as high. When analyzing the experiments performed, we infer that specific combinations of scientific areas inside a test corpus produce specific categorization results. We conclude that the effectiveness of automated text categorization depends on several factors, including the characteristics of the corpus. The effectiveness does not depend on the pre-processing algorithms and categorization algorithms only.

Keywords: Automated Text Clustering. Brazilian Portuguese. Effectiveness. Scientific Articles. Digital Libraries.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BARRETO, A. A. Uma quase história da Ciência da Informação. **DataGramaZero - Revista de Ciência da Informação**, v.9 n.2. abr/08. Disponível em: http://www.dgz.org.br/abr08/F_I_art.htm. Acesso em: 11 jun. 2012.

2. SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspec. Ci. Inf.**, Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

3. BIDERMAN, Maria Tereza Camargo. O conhecimento, a terminologia e o dicionário. **Cienc. Cult.**, São Paulo, v. 58, n. 2, Junho 2006.

4. DUQUE, Cláudio Gottschalg. **SIRILICO - Uma proposta para um Sistema de Recuperação de Informação baseado em Teorias da Linguística computacional e Ontologia**. Belo Horizonte, MG. UFMG, D.Sc., Ciência da Informação. Tese – Universidade Federal de Minas Gerais, 2005.

5. NUNES, M. G. V.; ALUÍSIO, S. M.; PARDO, T. S. Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade. **Linguamática**, v. 2, n. 2, Junho 2010.

6. MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi: Cambridge University Press, 2008.

7. MAIA, L. C.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspect. Ciênc. Inf.**, **v.**15, n.1, 2010.

8. DASILVA, C. F.; VIEIRA, R.; OSÓRIO, F. S.; QUARESMA, P. Mining Linguistically Interpreted Texts. In: Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora. 5., 2004, Geneva, Switzerland. **Anais do LINC-04.** Geneva, Switzerland, 2004.

9. CAMARGO, Y. B. L. **Abordagem linguística na classificação de textos em português.** Rio de Janeiro, RJ. COPPE/UFRJ, M.Sc., Engenharia Elétrica. Dissertação – Universidade Federal do Rio de Janeiro, COPPE, 2007.

10. FIGUEIREDO, F. S. **Construção de Evidências para Classificação Automática de Textos.** Belo Horizonte, MG. UFMG, Msc., Ciência da Computação. Dissertação – Universidade Federal de Minas Gerais, 2008.

11. BOUCKAERT, R. R. **Bayesian Network Classifiers in Weka for Version 3-5-7**. Waikato, Nova Zelândia. University of Waikato, 2008.

12. JAIN, A. K.; MAO, J.; MOHIUDDIN K. M**.** Artificial Neural Networks: a tutorial**. IEEE – Computer**., v. 29, n. 3, Março de 1996.

13. ROSS Q. **C4.5: Programs for Machine Learning**. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

14. AIRES, R. V. X. **Implementação, Adaptação, Combinação e Avaliação de Etiquetadores para o Português do Brasil.** São Paulo, SP. USP, Msc. Ciência da Computação. Dissertação - Universidade de São Paulo, 2000.

15. CALDAS JUNIOR, J.; IMAMURA, C.Y.M.; REZENDE, S.O. Avaliação de um Algoritmo de Stemming para a Língua Portuguesa. In: Segundo Congresso de Lógica Aplicada à Tecnologia. 2, 2001, São Paulo-SP. **Anais do Segundo Congresso de Lógica Aplicada à Tecnologia**. São Paulo: SENAC/Plêiade, 2001.

16. MARKOV, Z.; LAROSE, D. T. **Data Mining the Web: Uncovering Patterns in Web Content, Structure and Usage**. Hoboken, New Jersey: John Wiley & Sons, Inc, 2007.