JHI JOURNAL OF HEALTH INFORMATICS

Categorização automática de conteúdos web de saúde em português brasileiro com classificador bayesiano

Automatic categorization of health-related web content in brazilian portuguese language with bayesian classifier

Categorización automática de contenidos web sobre salud en portugués de Brasil con el clasificador bayesiano

Fernando Sequeira Sousa[1],Felipe Mancini[2], Fabio Oliveira Teixeira[1], Alex Esteves Jaccoud Falcão[1], Anderson Diniz Hummel[1], Fátima de Lourdes dos Santos Nunes[3], Daniel Sigulem[4], Ivan Torres Pisa[5]

RESUMO

Descritores: Classificação, Informação de Saúde ao Consumidor, Internet

Objetivo: Investigar aplicação de diferentes métodos de representação de textos por vetores de pesos com classificador bayesiano para classificação automática de conteúdos web de saúde em português. **Métodos:** Foi utilizado conjunto de 3.702 páginas web de saúde em português dividido em 19 categorias provenientes do Open Directory Project. Foram comparados desempenho de 4 métodos de representação de textos por vetores de pesos utilizados com o classificador Naive Bayes, medidos por revocação, precisão e F_2, considerando da primeira à quinta posições dos rankings de relevância de categorias. **Resultados:** A representação dos textos por ocorrência dos termos utilizada com o classificador Naive Bayes (*nb-to*) atingiu 0,91 de revocação, precisão e F_2 para a primeira posição do ranking de relevância de categorias; para a quinta posição os valores foram 0,98; 0,20 e 0,54, respectivamente. Estes valores colocam *nb-to* como o melhor classificador dos investigados para a base de dados utilizada, com diferença estatística entre este e todos os demais classificadores. **Conclusão:** Métodos de recuperação de informação baseados no Naive Bayes podem ser utilizados com sucesso para categorizar conteúdo web de saúde em idioma português, sendo que o classificador *nb-to* atingiu o melhor desempenho na tarefa de classificação.

ABSTRACT

Keywords: Classification, Consumer Health Information, Internet

Objective: To investigate the application of different methods of text representation by weighting vectors with a Bayesian classifier to automatically categorize health-related web pages in Brazilian Portuguese language. **Methods:** We used a set of 3,702 health-related web pages in Brazilian Portuguese language, separated in 19 categories, from Open Directory Project. We compared the effectiveness of 4 different methods of text representation by weighting vectors used with the Naive Bayes classifier, according to recall, precision and F_2 measures and considering from the first to the fifth positions of category relevance ranking produced by the classifiers. **Results:** The text representation by counting the term occurrence with the Naive Bayes classifier (*nb-to*) reached 0.91 of recall, precision and F_2 to the first position of the category relevance ranking, while to the fifth position the values were 0.98, 0.20, and 0.54, respectively. These results place *nb-to* as the best classifier to the database used in this work, with statistical differences between this classifier and all other. **Conclusion:** The information retrieval methods based on Naïve Bayes can be successfully used to categorize health-related web content in Brazilian Portuguese language and the *nb-to* classifier achieved the best performance at classification.

RESUMEN

Descriptores: Clasificación, Información de Salud al Consumidor, Internet

Objetivo: Investigar aplicación de diferentes métodos de representación de textos con un clasificador bayesiano para clasificación automática de contenidos web sobre salud en portugués de Brasil. **Métodos**: Se utilizó un conjunto de 3.702 páginas web sobre salud en portugués de Brasil, en 19 categorías y provenientes del Open Directory Project. Se comparó el desempeño de 4 métodos de representación de textos por vector de pesos utilizados con el clasificador Naive Bayes, medidos por recall, precisión y F_2, considerándose de la primera a la quinta posición en los rankings de relevancia de categorías. **Resultados**: El clasificador Naive Bayes (*nb-to*) alcanzó 0,91 de recall, precisión y F_2 para la primera posición en el ranking de relevancia de categorías, mientras que para la quinta posición, los valores fueron de 0,98, 0,20 y 0,54, respectivamente. Estos valores colocan el nb-to como el mejor clasificador para la base de datos utilizado, con diferencia estadística entre éste y todos los clasificadores testados. **Conclusión**: Los métodos de recuperación de información basados en el Naive Bayes pueden ser utilizados con éxito para categorizar contenidos web sobre salud en el idioma portugués de Brasil, pero el clasificador nb-to alcanzó el mejor desempeño al ejecutar la tarea de clasificación.

[1] *Mestre em Ciências. Programa de Pós-graduação em Gestão e Informática em Saúde, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil*
[2] *Doutor em Ciências – Gestão e Informática em Saúde. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Guarulhos (SP), Brasil*
[3] *Professora Livre-docente. Escola de Artes, Ciências e Humanidades, Universidade de São paulo - USP, São Paulo (SP), Brasil*
[4] *Professor Titular em Informática em Saúde. Departamento de Informática em Saúde, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil*
[5] *Professor Adjunto. Departamento de Informática em Saúde, Universidade Federal de São Paulo - UNIFESP, São Paulo (SP), Brasil*

Autor Coorespondente: **Ivan Torres Pisa**
e-mail: **buscasaude@unifesp.br**

Artigo recebido: **28/09/2011**
Aprovado: **01/12/2011**

INTRODUÇÃO

Hoje se estima que existam pelo menos 17,5 bilhões de páginas web[1] e este número continua crescendo. A grande quantidade e o constante crescimento de documentos disponíveis na web podem ser atribuídos, em grande parte, à facilidade em se criar textos e outros conteúdos (imagens, filmes, documentos) web a partir de ferramentas que dispensam conhecimentos específicos de programação[2]. Se por um lado este universo de informação leva conhecimento a mais pessoas, por outro lado apresenta desvantagens[3], em especial quanto à dificuldade do usuário em avaliar se a informação encontrada é relevante para o seu propósito e se apresenta um considerável nível de confiança.

Segundo o Centro de Estudos sobre as Tecnologias da Informação e Comunicação (CETIC)[4], em 2010, 35% das atividades desenvolvidas na internet no Brasil foram relativas à procura de informação relacionada à saúde. Ainda assim, os usuários chegam a conclusões erradas quando buscam por informações sobre este tema[5], principalmente devido ao baixo conhecimento dos mesmos sobre assuntos de saúde e à apresentação de resultados pouco relevantes por parte das ferramentas de busca. A estratégia de busca por vezes também não é satisfatória. Usualmente os usuários utilizam poucos termos, observam poucos resultados e raramente utilizam mecanismos avançados de buscav[6]. Existe, portanto, a necessidade de se estudar estratégias e criar ferramentas que auxiliem o usuário da internet a encontrar e organizar o conteúdo disponível na web, de acordo com seus interesses[7-8]. Uma abordagem que pode ser adotada é a classificação automática de textos[9].

O interesse de grande parte dos trabalhos que aplicam métodos de classificação de textos a conteúdos de saúde é para a classificação ou indexação de textos científicos[10-11]. Entretanto, algumas pesquisas focam na classificação de textos web de saúde voltados para o público leigo.

Como exemplo, o trabalho de Bangalore e colaboradores[12] aplicou um classificador baseado em relevância de categorias para classificar retornos de uma busca no Google em categorias pertencentes ao Medical Subject Headings (MeSH) e avaliou o resultado da classificação com a opinião de especialistas, conseguindo resultados satisfatórios. Em outra aplicação para classificação de conteúdo web de saúde, Himmel e colaboradores[13] desenvolveram um classificador automático para mensagens de um fórum médico em categorias pré-definidas. Destaca-se deste trabalho a aplicação de métodos já conhecidos e bem estabelecidos na área de classificação em uma base de dados específica da web, cujos textos são voltados para o público leigo e escritos em outro idioma (alemão), mostrando que estes métodos podem ser aplicados em idiomas diferentes do inglês com resultados igualmente satisfatórios.

Recentemente um esforço vem sendo feito para classificar conteúdos web no idioma português brasileiro em conteúdos de saúde e não saúde[7,14]. O foco desses estudos é desenvolver um classificador automático de textos que, a partir do retorno de uma pesquisa no Google, selecione somente páginas com conteúdo de saúde, sempre pensando no melhor resultado para o público leigo.

Dada a ainda pequena quantidade de pesquisas que exploram a categorização automática de conteúdos web em português brasileiro, com foco na disponibilização da informação para o público leigo, este trabalho teve como objetivo investigar a aplicação de métodos de recuperação de informação e classificação de textos para esse propósito, aplicando diferentes métodos de representação de textos por vetores de pesos com um classificador bayesiano.

MÉTODOS

As páginas web utilizadas neste trabalho foram coletadas de uma lista de páginas de saúde em português brasileiro provenientes do diretório web Open Directory Project (ODP - http://www.dmoz.org), mantido por voluntários responsáveis pela manutenção e edição de uma ou mais categorias. O conjunto utilizado foi composto por 3.702 páginas web de 19 categorias relacionadas à saúde: Acidentes; Associações e Entidades; Boa Forma; Clínicas e Hospitais; Cuidados Pessoais; Distúrbios; Homeopatia; Medicina Preventiva; Odontologia; Órgãos Públicos; Planos de Saúde e Seguro; Produtos e Serviços de Apoio; Profissionais; Saúde; Saúde da Criança; Saúde da Mulher; Saúde do Homem; Saúde Ocupacional e Ambiental; e Terapias Alternativas.

Após a coleta, as páginas web foram pré-processadas para eliminar palavras e termos irrelevantes e diminuir a complexidade da representação das páginas web. Por irrelevantes, entendem-se elementos que não fazem diferença no processo de classificação. Em ordem, foram removidos os marcadores HTML, as *stopwords* presentes no SnowBall e as palavras foram reduzidas à sua raiz (*stemming* PTStemmer)[15]. Estes textos foram então transformados em vetores de características, representando a contagem das palavras presentes nos textos[16]. Foram utilizados quatro métodos provenientes da área de recuperação de informação, frequentemente empregados para representação de textos[17]: frequência do termo (*tf*); frequência do termo ponderada pelo inverso de sua frequência nos documentos (*tf.idf*); ocorrência do termo (*to*); e ocorrência binária (*bo*). *To* é a abordagem em que o vetor de características é construído contando quantas vezes as palavras ocorrem. *Bo* não considera o total de vezes, mas sim a ocorrência ou não do termo. *Tf* considera a frequência das palavras nos textos, ou seja, quantas vezes uma palavra aparece em relação ao total de palavras. *Tf.idf* também considera a frequência das palavras em um texto, porém este valor é ponderado pelo inverso da frequência de um termo em todos os documentos, considerando assim a importância do documento no conjunto.

Os vetores de características contendo a contagem das palavras presentes nos textos coletados, realizada por cada um dos métodos citados anteriormente, foram então submetidos, separadamente, ao classificador de padrões Naive Bayes[18], formando, portanto, o grupo de 4 métodos de classificação utilizados neste trabalho. A regra de classificação é bastante simples, o que não implica no seu

desempenho; pelo contrário, a regra bayesiana garante que o desempenho do classificador seja ótimo[18]. Optou-se em utilizar este classificador pela sua simplicidade e por geralmente atingir os melhores resultados em problemas de classificação, além de ser amplamente utilizado com sucesso como classificadores automáticos de textos[7,19].

Neste trabalho foi utilizada uma implementação do Naive Bayes Multinomial[19], na qual são calculadas as probabilidades dos textos pertencerem a cada uma das categorias. A categoria com maior probabilidade torna-se àquela associada ao texto. A partir dessas probabilidades também é possível gerar o ranking de relevância de categorias para um determinado texto. Esta abordagem foi utilizada para medir a melhora no desempenho dos classificadores quando se considera além da primeira posição como a categoria correta. Este ranking também pode ser utilizado para a sugestão de multirrótulos para as páginas web.

Avaliação de Desempenho

A partir do conjunto de 3.702 páginas web coletadas foram criados 50 conjuntos, cada um deles composto por uma amostra aleatória de 70% do total para treinamento os 30% restantes para os testes, utilizando o processo conhecido como *holdout* repetitivo[20]. O desempenho dos classificadores foi medido a partir de medidas clássicas do campo de recuperação de informação e classificação de textos: revocação, precisão e F-measure[16].

A revocação e a precisão foram calculadas a partir da estratégia de *microaveraging*[16]. Devido à adoção do ranking de relevância e mensuração do desempenho além da categoria de maior relevância, ou seja, se a classe correta ocorre além da primeira posição do ranking, o cálculo dessas medidas foi generalizado para qualquer posição do ranking de categorias[10,21].

A microrrevocação média do classificador até a *k*-ésima posição do vetor de relevância é dada pela Equação (1):

Na Equação (1), o numerador de μRec_k é dado pela soma de categorias encontradas e corretas para o documento *i* até a posição *k* do ranking de relevância, dividido pelo total de categorias corretas. O denominador de μRec_k será a quantidade de documentos no conjunto de testes ($|P|$).

De maneira similar, a microprecisão média do classificador até a *k*-ésima posição do vetor de relevância é dada pela Equação (2):

Na Equação (2) o numerador será a soma das categorias corretas e encontradas para o documento *i* até a *k*-ésima posição do ranking de relevância pelo total de categorias encontradas para o mesmo documento e posição. O denominador de μPre_k será também a quantidade de documentos de testes ($|P|$).

Por fim, foi utilizado o F-measure como a medida de balanceamento entre revocação e precisão. Como o foco dessa pesquisa foi encontrar a maior parte dos documentos que pertençam a uma categoria, ou seja, foi priorizada a revocação do classificador ao invés de sua precisão, foi empregada a derivação do F-measure conhecida como F_2, que atribui uma importância maior para a revocação em relação à precisão. O F_2 da *k*-ésima posição do ranking de relevância é calculado segundo Equação (3):

Experimento

O experimento teve como objetivo investigar diferenças no desempenho entre os diferentes classificadores de padrões utilizados neste trabalho na tarefa de categorizar páginas web de saúde. Os classificadores foram treinados e testados com a base de dados coletada e pré-processada, tendo seus desempenhos medidos de acordo com revocação, precisão e F_2. Após a extração dessas medidas em 50 diferentes conjuntos aleatórios de treinamento e testes (*holdout* repetitivo), foi aplicado o teste t de Student para comparação de médias[22] para verificar diferenças estatísticas significantes entre os classificadores.

RESULTADOS E DISCUSSÃO

Considerando o ranking de relevância de cada classificador, a Figura 1 ilustra a evolução de suas revocações médias para as 5 primeiras posições do ranking de relevância (μRec_k). A Tabela 1 apresenta os respectivos valores. Conforme esperado, observou-se um comportamento padrão para todos os classificadores testados, com crescimento no valor da revocação, tendendo a 1 conforme aumentaram-se as posições do ranking de relevância.

No geral, observou-se que o classificador *nb-to* inicia a curva com um desempenho superior ao dos outros classificadores. Por outro lado, os classificadores *nb-tfidf* e *nb-tf* demoram mais para alcançar um desempenho similar aos dos demais. O *tf.idf*, mesmo destacado como uma

$$\mu Rec_k = \frac{\sum_{i=1}^{|P|} \frac{(categorias\ encontradas\ e\ corretas)_{i,k}}{(categorias\ corretas)_{i,k}}}{|P|} \quad (1)$$

$$\mu Pre_k = \frac{\sum_{i=1}^{|P|} \frac{(categorias\ encontradas\ e\ corretas)_{i,k}}{(categorias\ encontradas)_{i,k}}}{|P|} \quad (2)$$

$$F_{2,k} = 5 \cdot \frac{\mu Pre_k \cdot \mu Rec_k}{(4 * \mu Pre_k) + \mu Rec_k} \quad (3)$$

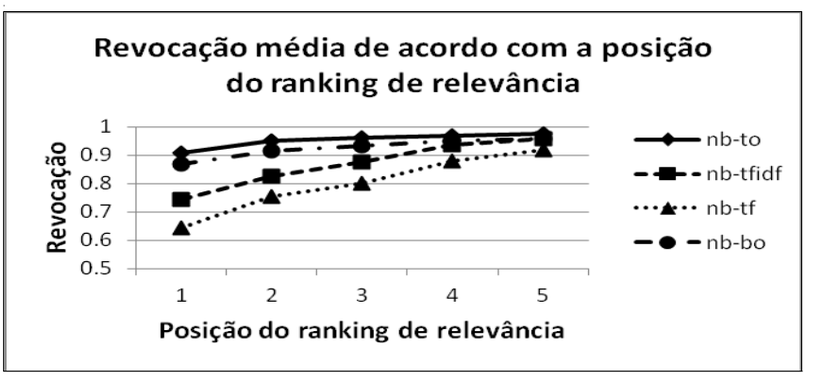

Figura 1 – Valores de revocação para as 5 primeiras categorias doranking de relevância

boa e bem estabelecida técnica para tal tarefa[23] alcançou os melhores classificadores apenas na quarta posição. Ou seja, de todos os classificadores testados, o *nb-to* sempre apresentou o melhor desempenho até quinta posição do ranking de relevância, enquanto que o *nb-tf* apresentou o pior desempenho, ficando mais distante dos outros classificadores na quinta posição.

Tabela 1 – Valores de revocação na primeira e na quinta posição do ranking de relevância (ordenado pela primeira posição)

Classificador	Primeira posição	Quinta posição
nb-tf	0,64	0,92
nb-tfidf	0,74	0,96
nb-bo	0,87	0,96
nb-to	0,91	0,98

Foi possível notar também uma grande diferença no desempenho quando comparadas a primeira e a quinta posição (Tabela 1). Na quinta posição todos os classificadores alcançaram uma revocação média entre as categorias maior que 0,90. O maior aumento foi observado para o classificador *nb-tf* (de 0,64 para 0,92). Já o classificador *nb-to* teve o menor aumento (de 0,91 para 0,96). Entretanto, seu desempenho foi melhor que os outros três classificadores tanto na primeira quanto na quinta posição, já que foram encontradas diferenças estatísticas significantes entre todos os classificadores nestas duas posições. Ou seja, como o *nb-to* apresentou o maior valor numérico para a primeira e quinta posições, e houve diferença significante entre este e os outros classificadores, o *nb-to* mostrou-se a melhor alternativa para classificar conteúdos web de saúde em português brasileiro.

De maneira oposta ao comportamento de crescimento da revocação, a precisão média em cada posição do ranking de relevância (μPre_k) apresentou tendência de queda, conforme apresentado na Figura 2. Respectivos valores são apresentados na Tabela 2.

A queda no valor de precisão foi mais acentuada que o aumento da revocação. Logo da primeira para a segunda posição, observou-se um grande decremento, quando os valores para todos os classificadores eram maiores que 0,6 e diminuíram para valores menores que 0,5. Além disso, a partir da segunda posição o desempenho da precisão para todos os quatro classificadores começa a se aproximar. Assim como a revocação, o classificador *nb-to* obteve o melhor desempenho nas primeiras posições do ranking de relevância, sendo que o classificador *nb-bo* é aquele que se aproxima mais rapidamente. Já os classificadores *nb-tf* e *nb-tfidf* foram os que obtiveram os piores desempenhos.

Tabela 2 – Valores de precisão para cada classificador na primeira e na quinta posição do ranking de relevância (ordenado pela primeira posição)

Classificador	Primeira posição	Quinta posição
nb-tf	0,64	0,18
nb-tfidf	0,74	0,19
nb-bo	0,87	0,19
nb-to	0,91	0,20

Avaliando a diferença entre a primeira e quinta posições (Tabela 2), observa-se que existe uma grande diferença

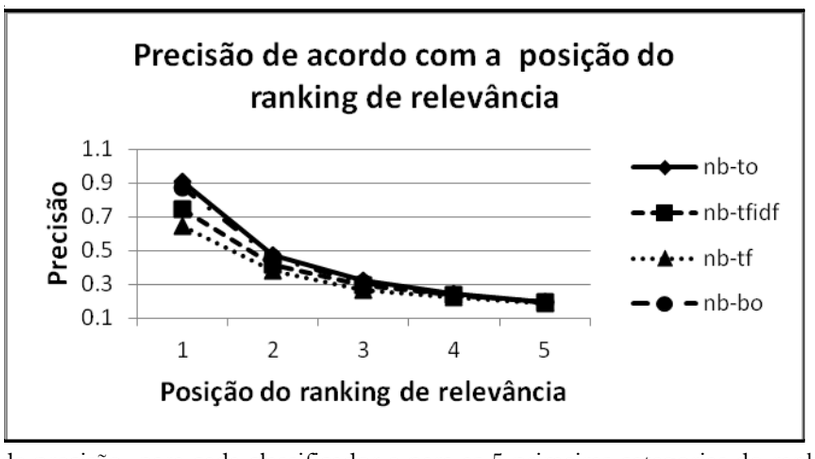

Figura 2 – Valores de precisão para cada classificador e para as 5 primeiras categorias do ranking de relevância

no desempenho da precisão. Mais uma vez, o classificador *nb-to* obteve um desempenho numericamente melhor que os outros, ainda que essa diferença seja mínima para a quinta posição (apenas 0,01). Para a primeira posição, os valores se igualam ao da revocação, já que a quantidade de falsos positivos e falsos negativos é igual quando se utiliza *microaveraging*.

Finalmente, a Figura 3 ilustra a evolução do F_2 médio em cada posição do ranking de relevância ($F2_k$). Respectivos valores são apresentados na Tabela 3. Nesta medida observou-se também uma tendência de queda para todos os classificadores, visivelmente menos acentuada que a queda da precisão. É justamente devido à queda acentuada na precisão que F_2 também diminui. Já a contribuição do aumento da revocação na curva do F_2 é a maior suavidade na queda dos valores ao longo das posições de relevância.

Acompanhando os resultados da revocação e da precisão, o F_2 do classificador *nb-to* obteve um melhor desempenho até a terceira posição, quando começa a ficar muito próximo dos outros classificadores, principalmente do *nb-bo*. Já os classificadores *nb-tf* e *nb-tfidf* começaram mais distantes e com desempenho pior em relação aos outros, e foram aos poucos se aproximando do desempenho dos demais.

Tabela 3 – Valores de F_2 para cada classificador na primeira e na quinta posição do ranking de relevância (ordenado pela primeira posição)

Classificadores	Primeira posição	Quinta posição
nb-tf	0,64	0,51
nb-tfidf	0,74	0,53
nb-bo	0,87	0,53
nb-to	0,91	0,54

Na comparação de F_2 entre a primeira e quinta posições do ranking de relevância dos classificadores (Tabela 3) o classificador *nb-to* manteve seu melhor desempenho entre todos os quatro classificadores utilizados, apesar de que na quinta posição os valores de todos foram muito próximos. Os valores de F_2 para a primeira posição, mais uma vez foram iguais aos obtidos na revocação e precisão, devido às características da medida utilizada. Nota-se também que até a quinta posição todos os classificadores atingiram um desempenho maior que 0,5, considerados um bom resultado dada à quantidade de categoria considerada. Assim como a análise estatística realizada para a revocação, foram encontradas diferenças estatísticas significantes entre todos os classificadores para o F_2, tanto na primeira quanto na quinta posição do ranking de relevância de categorias, evidenciando o melhor desempenho do *nb-to* para a tarefa de classificação proposta.

Para a tarefa de classificação de páginas web em saúde em diferentes categorias, o desempenho médio tanto da revocação quanto do F_2 pode ser considerado satisfatório quando foi considerada a primeira posição do ranking de relevância, uma vez que para cinco dos seis classificadores utilizados a revocação atingiu valores maiores que 0,70, conseguindo um aumento considerável logo na segunda posição e atingindo resultados ainda melhores na quinta. Apenas o classificador *nb-tf* obteve um desempenho abaixo de 0,70 na primeira posição, e juntamente com o *nb-tfidf* foram os que atingiram valores de revocação maiores que 0,90 somente após quatro posições do ranking de relevância.

Os gráficos da Figura 1 e da Figura 2 evidenciam que o comportamento da revocação é contrário ao da precisão. Ao passo que a revocação melhora com o aumento da posição de relevância, a precisão sofre uma queda, em maior proporção que o aumento da revocação. Para a primeira posição do ranking de relevância, foram alcançados valores altos de precisão e de revocação, já que estas medidas se igualam quando utilizamos o *microaveraging*[24]. A partir da segunda posição do ranking, observa-se um crescimento da revocação e um decrescimento da precisão. Este comportamento, de diminuir a precisão enquanto ajusta-se o classificador para obter uma maior revocação (ou diminuir a revocação quando se aumenta a precisão) é o comportamento padrão dos classificadores[21]. Este comportamento também é observado quando é medido o desempenho do classificador nas posições do ranking de relevância[10,21], uma vez que se aumenta a chance do classificador acertar (aumento da revocação) ao passo que também se associa a categoria errada para outras páginas (queda na precisão). O fato de haver apenas uma categoria associada às páginas do conjunto de testes contribui para queda na precisão, já

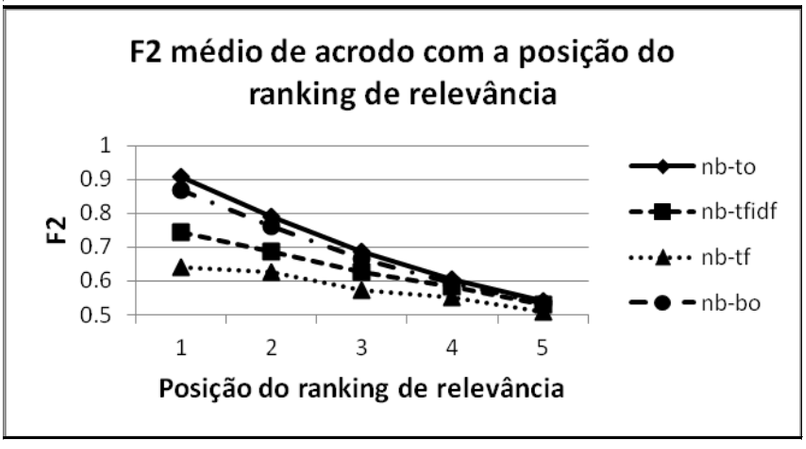

Figura 3 – Valores de F_2 para cada classificador e para as 19 categorias, de acordo com ranking de relevância

que existia apenas uma categoria correta até a *k*-ésima posição do ranking de relevância.

Entretanto, a queda no valor da precisão não é um fator problemático para a aplicação proposta. A baixa precisão informa que o classificador associou a categoria errada a uma determinada página, classificada originalmente com apenas uma categoria. Porém, devido à alta variabilidade dos conteúdos da web, nada impede que uma determinada página possa ser associada a mais de um conteúdo específico de saúde. Priorizando o melhor desempenho da revocação é possível encontrar mais páginas que possam ser de uma categoria, mesmo que esta não tenha sido inferida pelo classificador como a mais relevante.

Apesar dos bons resultados encontrados, os classificadores utilizados lidam apenas com os conteúdos textuais das páginas web. Conteúdos não textuais, como imagens e conteúdos em *flash* não foram tratados, mas podem trazer informações ricas acerca do conteúdo e, consequentemente, da categoria que uma página web pertence[25], principalmente do ponto de vista da percepção do usuário.

Outro ponto que pode ser considerado e avaliado é a utilização de ontologias ou dicionários específicos, uma vez que o objetivo final é classificar textos web de um idioma e domínio específico: páginas web de saúde em português brasileiro. Em um trabalho similar a este[7], para classificar conteúdos web em saúde ou não saúde, os autores conseguiram uma melhora significativa no desempenho de um classificador bayesiano ao considerar apenas as palavras e termos presentes no vocabulário controlado MeSH (Medical Subject Headings - http://www.ncbi.nlm.nih.gov/mesh), que contém apenas palavras e termos específicos da saúde. Entretanto, a não utilização de um dicionário não invalida os resultados obtidos neste trabalho, visto os bons resultados atingidos com métodos clássicos e a aplicação destes em um buscador, a fim de direcionar usuários leigos a assuntos de saúde em suas buscas na web.

CONCLUSÃO

Este trabalho avaliou a aplicação do classificador de padrões Naive Bayes para categorizar automaticamente conteúdos web de saúde em idioma português brasileiro e comparou o desempenho de quatro diferentes métodos de representação de textos por vetores de pesos utilizados com este classificador. A avaliação foi realizada sobre uma base de dados composta por 3.702 páginas web de saúde em português brasileiro, divididos em 19 categorias.

Os resultados obtidos neste trabalho mostraram que o classificador Naive Bayes pode ser utilizado com sucesso para a tarefa proposta. Entretanto, este classificador utilizado com o método de representação de textos por *to* (ocorrência do termo) obteve um desempenho significativamente melhor em relação aos outros classificadores, com revocação de 0,91 e 0,98 para a primeira e quinta posições do ranking de relevância de categorias, respectivamente. A representação dos textos por *tf.idf,* apesar de ser amplamente utilizada, apresentou um desempenho inferior a maioria dos classificadores. Além do mais, todas classificadores utilizados apresentaram diferença estatística significante quanto ao desempenho da revocação e de F_2.

Este trabalho contribui para uma área ainda pouco explorada, aplicando métodos de classificação de textos e recuperação de informação para classificar automaticamente conteúdos web de saúde, com o foco de prover a melhor informação para o público leigo. Especificamente para o idioma português brasileiro, a quantidade de trabalhos é ainda menor, e o presente trabalho mostrou que é possível classificar com efetividade conteúdos web de saúde em português brasileiro utilizando métodos já consagrados. Outra contribuição importante deste trabalho é em relação à adoção do ranking de relevância, a partir das probabilidades de categorias do Naive Bayes para melhorarmos o desempenho da revocação dos classificadores. Mesmo com a certeza do aumento da revocação, já que se aumenta a tolerância, o aumento foi significativo, com valores de revocação maiores que 0,90 quando consideradas apenas as 5 primeiras categorias do ranking de relevância.

AGRADECIMENTOS

Os autores agradecem a CAPES-DS e a Grant NIH/Fogarty 5D43TW007015-07 (PI: Dra Lucila Ohno Machado, Diretora Brasileira: Profa. Dra. Heimar de Fátima Marin) pelo apoio financeiro.

REFERÊNCIAS

1. Kunder M de. The size of the world wide web [Internet]. 2011 [citado 2011 jun 14]. Disponível em: http://www.worldwidewebsize.com/index.php?lang=EN
2. Breitman K, Casanova MA, Truszkowski W. Semantic web: concepts, technologies and applications. London: Springer; 2007.
3. Fogg BJ, Soohoo C, Danielson DR, Marable L, Stanford J, Tauber ER. How do users evaluate the credibility of Web sites? A study with over 2.500 participants [Internet]. In: Proceedings of the 2003 conference on Designing for user experiences. San Francisco, California: ACM; 2003. p. 1-15. Available from: http://portal.acm.org/citation.cfm?id=997097
4. Pesquisa sobre o uso das Tecnologias da Informação e da Comunicação no Brasil - TIC Domicílios e Empresas 2010 [Internet]. Centro de Estudos sobre as Tecnologias da Informação e Comunicação (CETIC); 2010 [citado 2011 set 6]. Disponível em: http://www.cetic.br/publicacoes/
5. Keselman A, Browne AC, Kaufman DR. Consumer health information seeking as hypothesis testing. J Am Med Inform Assoc. 2008;15(4):484-95.
6. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods Inf Med. 2002;41(4):289-98.
7. Mancini F, Sousa FS, Teixeira FO, Falcão AEJ, Hummel AD, da Costa TM, et al. Use of medical subject headings (MeSH) in portuguese for categorizing web-based healthcare content. J Biomed Inform. 2010;44(2):299-309.
8. Stvilia B, Mon L, Yi Y. A model for online consumer health information quality. J Am Soc Inf Sci Technol. 2009;60(9):1781-91.
9. Jackson P, Moulinier I. Text categorization. In: Natural language processing for online applications: text retrieval, extraction and categorization. Philadelphia: John Benjamins Publishing Company; 2007. p.119-71.
10. Humphrey SM, Névéol A, Gobeil J, Ruch P, Darmoni SJ,

Browne A. Comparing a rule based vs. statistical system for automatic categorization of MEDLINE documents according to biomedical specialty. J Am Soc Inf Sci Technol. 2009;60(12):2530-9.
11. Teixeira F, Falcão AEJ, Sousa FS, Hummel AD, Costa TM, Mancini F, et al. Similarity-based scoring method for classification of health informatics content. J. Health Inform. 2011;3(2):35-42.
12. Bangalore AK, Divita G, Humphrey S, Browne A, Thorn KE. Automatic Categorization of Google Search Results for Medical Queries Using JDI [Internet]. In: Medinfo 2007. Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems. Amsterdam: IOS Press; 2007. p. 2253-4. Available from: http://search.informit.com.au/documentSummary;dn=793354056979143;res=IELHEA
13. Himmel W, Reincke U, Michelmann HW. Text mining and natural language processing approaches for automatic categorization of lay requests to web-based expert forums. J Med Internet Res. 2009;11(3):e25.
14. Falcão AEJ, Mancini F, Costa TM, Hummel AD, Teixeira FO, Sigulem D, et al. InDeCS: Método automatizado de classificação de páginas web de saúde usando mineração de texto e descritores em ciências da saúde (DeCS). J Health Inform. 2009;1(1):18-24.
15. Ziviani N, Baeza-Yates R. Text operations. In: Baeza-Yates R, Ribeiro-Neto B, editors. Modern Information Retrieval. New York: Addison Wesley; 1999. p.163-90.
16. Sebastiani F. Machine learning in automated text categorization. ACM Comput. Surv. 2002;34(1):1-47.
17. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inform Processing and Management. 1988;24(5):513-23.
18. Duda RO, Hart PE, Stork DG. Pattern classification 2a. ed. Wiley-Interscience; 2000.
19. McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification. Dimension Contemporary German Arts and Letters. 1998;752(1):41-8.
20. Kim J-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics & Data Analysis. 2009;53(11):3735-45.
21. Yang Y. An evaluation of statistical approaches to text categorization. Inf Retr. 1999;1(1-2):69–90.
22. Altman DG. Practical statistics for medical research. London. Chapman & Hall/CRC; 1991.
23. Mladenic D. Text-Learning and Related Intelligent Agents: A Survey. IEEE IntelSyst. 1999;14:44–54.
24. Yang Y, Slattery S, Ghani R. A study of approaches to hypertext categorization. J Intell Inf Syst. 2002;18(2-3):219-41.
25. Qi X, Davison BD. Web page classification: Features and algorithms. ACM Comput. Surv. 2009;41(2):1-31.