# Indexação de artigos científicos de Informática em Saúde por meio da competição de técnicas de extração de características

**Fabio Teixeira[1], Fernando S. Sousa[1], Gabriela Denise Araujo[1], Felipe Mancini[3], Luciano V. de Araujo[2], Ivan T. Pisa[1]**

[1]Departamento de Informática em Saúde – Universidade Federal de São Paulo (UNIFESP) – São Paulo – SP – Brasil

[2]Escola de Artes Ciências e Humanidades – Universidade de São Paulo (USP) – São Paulo – SP – Brasil

[3]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – São Paulo – SP - Brasil

buscasaude@unifesp.br

***Abstract.*** *The objective of this study was to develop a automated mechanism for indexing of scientific articles under interdisciplinary domain of Health Informatics. It was contemplated building a database with 10,800 titles and abstracts of papers distributed uniformly between Health Informatics, Computer Science and Health domains. The evaluation was performed by measuring $f_{0,5}$-score, which reached a value of 66%. Although the articles submitted to the task of indexing, belonging to an interdisciplinary scope, the proposed method was able to characterize them according to your area of interest, with satisfactory success rate.*

***Resumo.*** *O objetivo deste estudo foi desenvolver um mecanismo automatizado para a indexação de artigos científicos sob o domínio interdisciplinar da Informática em Saúde. Contemplou a construção de uma base de dados com 10.800 títulos e resumos de artigos científicos distribuídos uniformemente entre os domínios da Informática em Saúde, Ciência da Computação e Saúde. A avaliação foi realizada por meio da medida de desempenho f-score$_{0,5,}$ que alcançou o valor de 66%. Embora os artigos, submetidos à tarefa de indexação, pertencerem a um escopo interdisciplinar, o método proposto foi capaz de caracterizá-los de acordo com sua área de interesse, com taxa de acerto satisfatória.*

## 1. Introdução

A interdisciplinaridade da Informática em Saúde (IS) e a amplitude dos temas abordados em seu contexto, que ultrapassam barreiras previamente definidas por domínios de conhecimento, diversificando fontes de armazenamento e recuperação de informações, dificultam caracterizá-la sob um arcabouço de termos, conceitos e limites de atuação [Bernstam et al. 2010].

A consequência de um domínio interdisciplinar, como o da IS, é a dificuldade para a recuperação de informação pertinente ao seu contexto, uma vez que o conhecimento está diluído sob diversas áreas, tais como Saúde, Ciência da Computação,

Ciência da Informação e Engenharia Biomédica [Knaup and Dickhaus 2009 Pacheco et al. 2009 Van Bemmel 2008].

Um estudo promovido pela International Medical Informatics Association (IMIA) destaca as disciplinas que contribuem para a construção do domínio da IS, conforme mostra a Figura 1 , na qual o compartilhamento de métodos e ferramentas entre elas está presente e contribui para o desenvolvimento da área e definição do seu escopo [Mantas et al. 2010].

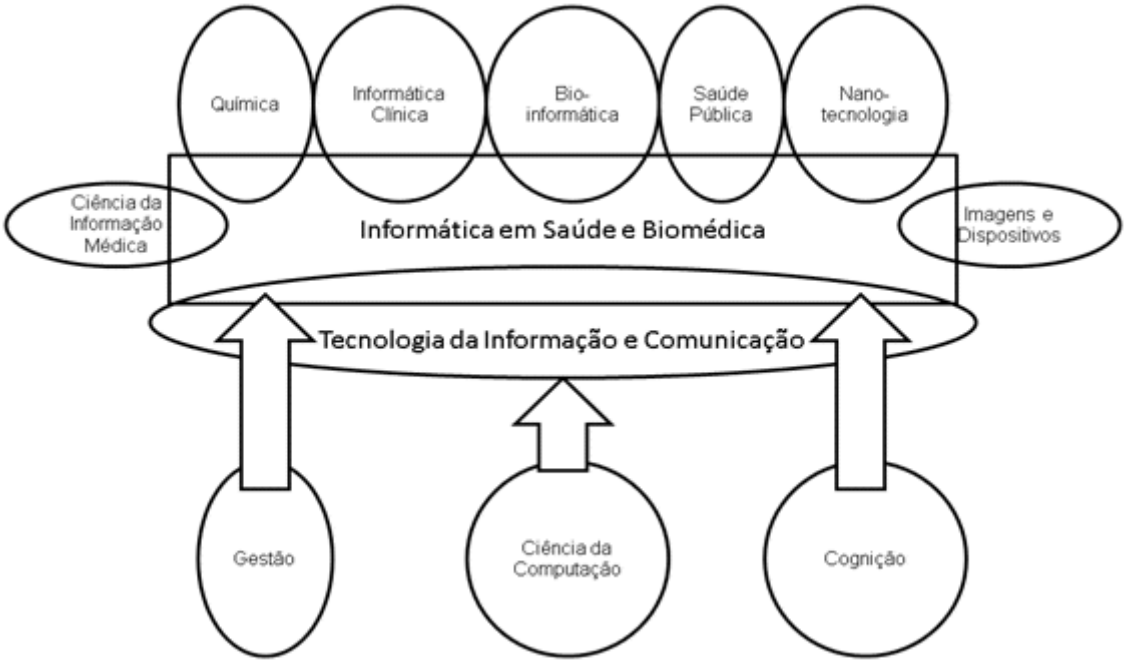


**Figura 1 - Disciplinas que contribuem para a interdisciplinaridade da Informática em Saúde (Figura traduzida e adaptada) [Mantas et al. 2010]**

O volume de artigos armazenados em bibliotecas virtuais e o aumento substancial do mesmo sugerem mecanismos automatizados que auxiliem a tarefa humana de indexação e recuperação de documentos. Especificamente para o domínio da Informática em Saúde, no qual a interdisciplinaridade intrínseca à mesma provoca um aumento da granularidade das fontes de publicação de conteúdo, o desafio atual é oferecer mecanismos capazes de recuperar informações de maneira eficiente neste domínio.

Portanto, os resultados deste trabalho pretendem responder à seguinte questão:

É possível criar um mecanismo automático que indexe conteúdos de artigos científicos de IS em bases que não são originalmente de IS?

## 2. Objetivo

O objetivo deste trabalho foi desenvolver um método capaz de indexar artigos científicos a partir de um conjunto de categorias pré-definidas, delimitadas pelos domínios da Informática em Saúde, Ciência da Computação e Saúde, utilizando técnicas de extração de características como parâmetro do classificador probabilístico Naive Bayes.

## 3. Métodos

Os métodos, inicialmente, concentraram-se na construção da base de dados de treinamento e validação. Uma vez construída a base de dados, esta suportou a criação dos vetores de características dos artigos científicos. O classificador de padrões Naive Bayes, escolhido para realizar a tarefa de indexação dos artigos, recebeu como parâmetro os vetores de características dos mesmos. Finalmente, as análises de desempenho e estatística foram abordadas. Os tópicos seguintes detalharão os eventos descritos resumidamente neste parágrafo.

*Construção da base de dados*

Os dados avaliados no estudo foram coletados a partir do portal ISI Web Of Knowledge (http://apps.isiknowledge.com), que concentra bancos de dados de publicações científicas de diferentes domínios de conhecimento.

O escopo da coleta concentrou-se em títulos e resumos de artigos científicos do idioma inglês, classificados sob um conjunto de categorias, relacionadas no Tabela 1 , associadas às revistas e aos respectivos domínios, disponíveis no portal utilizado.

**Tabela 1 - Categorias que compõem a base de dados.**

| **Domínios** | | |
|---|---|---|
| **Ciência da Computação** | **Informática em Saúde** | **Saúde** |
| Computer Science, Artificial Intelligence | Computer Science, Information Systems | Anatomy & Morphology |
| Computer Science, Information Systems | Computer Science, Interdisciplinary Applications | Biochemistry & Molecular Biology |
| Computer Science, Software Engineering | Engineering, Biomedical | Biology |
| Computer Science, Theory & Methods | Health Care Sciences & Services | Cell Biology |
| Engineering, Electrical & Electronic | Information Science & Library Science | Clinical Neurology |
| Information Science & Library Science | Mathematical & Computational Biology | Infectious Diseases |
| Management | Medical Informatics | Medicine, Research & Experimental |
| | Medicine, Research & Experimental | Microbiology |
| | Public, Environmental & Occupational Health | Neurosciences |
| | Statistics & Probability | Nursing |
| | | Oncology |
| | | Parasitology |
| | | Pediatrics |
| | | Psychiatry |
| | | Psychology, Developmental |
| | | Virology |

Foram selecionados 10.800 artigos científicos dispostos uniformemente entre as os domínios estudados. Portanto, cada domínio contribuiu com 3600 artigos, que posteriormente foram subdivididos em 2 conjuntos, treino e validação, por meio da distribuição de 75% e 25%, respectivamente. Esta subdivisão resultou em 8.100 artigos para a base de treinamento e 2.700 para a base de validação.

As revistas com maior fator de impacto e associadas às categorias definidas na Tabela 1 foram escolhidas, no entanto, deveriam possuir a quantidade mínima de 400 artigos publicados.

*Extração de características dos artigos científicos*

O processo de indexação de artigos foi composto, além da construção da base de dados a ser avaliada, pela transformação dos documentos textuais em vetores

numéricos, capazes de representá-los de maneira unívoca [Zhang et al. 2011]. O modelo de espaço vetorial é um dos métodos amplamente utilizados pela comunidade científica para tal representação [Salton, G. et al. 1975].

Este estudo utilizou 35.484 termos para compor a dimensão dos vetores numéricos que identificaram os artigos. A origem dos termos se deu a partir das palavras únicas presentes nos títulos e resumos dos artigos que compuseram a base de dados. Para alcançar o número total de termos utilizados foram executados processamentos preliminares, como: remoção de palavras presentes em uma lista de *stopwords* [Baeza-Yates and Ribeiro-Neto 2011] e aplicação de *stemming* [Porter 1997] para cada palavra.

O modelo de espaço vetorial utilizado neste estudo fez uso das técnicas não-supervisionadas *term frequency* (tf), *binary occurrence* (bo), *term occurrence* (to) e *term frequency inverse document frequency* (tf.idf) [Salton, Gerard and Buckley 1988].

*Classificador de padrões probabilístico*

Este estudo concentrou-se na aplicação do classificador *Naive Bayes e* na sua variação denominada *Multinomial* [Nigam and McCallum 1998], que permite capturar o cálculo da relevância dos termos. A escolha baseou-se na simplicidade do método e na sua eficiência para a tarefa supervisionada de classificação e indexação de textos, comprovada, ao longo dos anos, por meio de estudos científicos [Sohn et al. 2008].

Uma vez criadas as matrizes de características dos documentos que compõem a base de treino, as mesmas foram apresentadas como parâmetro de entrada do classificador de padrões probabilístico Naive Bayes, a fim de realizar a tarefa de indexação dos artigos que compuseram a porção de validação da base de dados utilizada nos experimentos deste estudo.

*Indexação de artigos científicos*

A indexação de artigos científicos utilizou as categorias listadas na Tabela 1 para indexar, de acordo com sua relevância, cada artigo científico presente na base de dados de validação. Conforme mostra a Tabela 2, o classificador de padrões treinado, por meio dos vetores de características dos artigos da base de treinamento, foi capaz de associar a cada artigo científico, do subconjunto de validação composto por 2.700 documentos, 30 elementos do vetor $R$, que armazenaram as relevâncias das Categorias $C_j$ em relação aos artigos $A_i$. Foram comparadas 4 estratégias de indexação, de acordo com as técnicas de extração de características abordadas anteriormente.

**Tabela 2 - Matriz das relevâncias entre artigos e categorias.**

| | $C_{j=1}$ | $C_{j=2}$ | ... | $C_{j=29}$ | $C_{j=30}$ |
|---|---|---|---|---|---|
| $A_{i=1}$ | $R_{i,j}$ | | | | $R_{1,30}$ |
| $A_{i=2}$ | $R_{2,1}$ | | | | $R_{2,30}$ |
| ⋮ | ⋮ | | | | ⋮ |
| $A_{i=2.699}$ | $R_{2.699,1}$ | | | | $R_{2.699,30}$ |
| $A_{i=2.700}$ | $R_{2.700,1}$ | | | | $R_{2.700,30}$ |

Também foi explorada a eficiência da indexação de artigos científicos quando diferentes técnicas de extração de características foram utilizadas, em conjunto, para realizar tal tarefa.

Foram considerados os vetores $R$ de cada artigo científico, resultantes da aplicação do classificador de acordo com as técnicas de extração de características abordadas, para propor a indexação final do mesmo.

Esta estratégia, neste estudo, foi denominada como "competição de técnicas" e utilizou a maior pontuação de relevância atribuída à posição do vetor e sua respectiva categoria para compor a indexação final do artigo.

A Tabela 3 foi utilizada para exemplificar a aplicação da estratégia a um artigo científico exemplo. Nesta tabela, as letras de A a G foram utilizadas para representar as categorias, os valores à direita de cada letra indicam a pontuação de relevância da categoria em relação ao artigo científico, de acordo com a técnica utilizada pelo classificador Naive Bayes.

Portanto, a indexação por meio da competição de técnicas atribuiu ao artigo exemplo, de acordo com a tabela, as categorias C, D, F, G e E como sugestão de indexação final.

**Tabela 3 - Categorias e respectivos valores de relevância associadas a um artigo, de acordo com o método utilizado pelo classificador Naive Bayes.**

| Posição do vetor | NB/tf.idf | NB/tf | NB/to | NB/bo |
|---|---|---|---|---|
| 1 | C = 0,99 | E = 0,23 | A = 0,75 | A = 0,55 |
| 2 | D = 0,82 | A = 0,18 | D = 0,74 | D = 0,50 |
| 3 | G = 0,60 | F = 0,15 | F = 0,62 | F = 0,42 |
| 4 | A = 0,50 | B = 0,14 | G = 0,63 | C = 0,30 |
| 5 | E = 0,44 | C = 0,12 | B = 0,43 | B = 0,18 |

*Medidas de desempenho aplicadas*

A medida convencional de desempenho *f-score* [Gehanno et al. 2009 Magdy and Jones 2010] foi aplicada aos resultados obtidos por meio do classificador de padrões construído para os experimentos e suas respectivas variações nos parâmetros de entrada.

O *f-score* é uma medida harmônica entre precisão e revocação, regida pela Equação 1 , na qual $P$ e $R$ representam os valores de Precisão e Revocação, respectivamente, e $\beta$ é um parâmetro de ponderação da Revocação em relação à Precisão, determinando a importância da mesma para o sistema de recuperação de informação avaliado. Neste estudo, os experimentos utilizaram o parâmetro $\beta$ igual a 0.5, que determinou maior importância à precisão.

$$F - score_{\beta} = \frac{(1 + \beta^2) * (P * R)}{\beta^2 * P + R}$$

**Equação 1 - Medida *f-score*.**

Além do parâmetro $\beta$, uma variação da medida de desempenho mencionada no parágrafo anterior foi utilizada para avaliar o classificador construído quanto à indexação de artigos científicos, uma vez que a pontuação de relevância da categoria em relação aos mesmos foi considerada.

A Equação 2 , Equação 3 e Equação 4 mostram os cálculos da precisão, revocação e *f-score*, respectivamente, utilizados quando um determinado intervalo $k$ de categorias, associadas pelo classificador a cada artigo $A_i$ de acordo com sua pontuação

de relevância, representado pelo parâmetro $R^{c}_{j\leq k,i}$, foi utilizado para determinar se a indexação sugerida pelo classificador foi correta.

Neste estudo, o intervalo de categorias foi avaliado apenas para o valor de $k$ igual a 5. A escolha deste valor respeitou a quantidade máxima de categorias associadas pelo Portal ISI Web of Knowledge aos artigos que compuseram a base de dados utilizada nos experimentos.

A indexação de textos por meio do cálculo da relevância de índices também foi explorado por Radlinski e Craswell [Radlinski and Craswell 2010], quando os mesmos avaliaram páginas web retornadas a partir de consultas submetidas a um buscador.

$$P@k = \frac{1}{|C|}\sum_{c=1}^{C}\left[\frac{\sum_{i=1}^{A} Relevantes\ e\ Retornados\left(R^{c}_{j\leq k,i}\right)}{\sum_{i=1}^{A} Retornados\left(R^{c}_{j\leq k,i}\right)}\right]$$

**Equação 2 - Precisão baseada em intervalo de avaliação.**

$$R@k = \frac{1}{|C|}\sum_{c=1}^{C}\left[\frac{\sum_{i=1}^{A} Relevantes\ e\ Retornados\left(R^{c}_{j\leq k,i}\right)}{\sum_{i=1}^{A} Relevantes\left(R^{c}_{j\leq k,i}\right)}\right]$$

**Equação 3 - Revocação baseada em intervalo de avaliação.**

$$F@k_{\beta} = \frac{(1+\beta^{2}) * (P@k * R@k)}{\beta^{2} * P@k + R@k}$$

**Equação 4 - *F-score* baseado em intervalo de avaliação.**

A Tabela 4 foi utilizada para exemplificar os cálculos da precisão, revocação e *f-score*, quando um intervalo $k$ de categorias foi considerado. Na primeira coluna deste quadro, os artigos A, B, C e E exemplificam a indexação proposta pelo Portal ISI Web of Knowledge, por meio da categoria “Statistics & Probability”. A coluna central reúne os artigos nos quais a categoria “Statistics & Probability” foi associada pelo classificador automatizado, representado pelo parâmetro $\sum_{i=1}^{A} Retornados\left(R^{c}_{j\leq k,i}\right)$ na Equação 2 , considerando que a relevância da categoria em relação aos artigos pode variar entre a primeira e quinta posição. De acordo com a indexação sugerida no exemplo pelo Portal, disposta na primeira coluna da Tabela 4 , apenas 2 artigos desta associação automatizada foram categorizados corretamente, os quais foram identificados como “verdadeiros positivos” na coluna central e alimentam o parâmetro $\sum_{i=1}^{A} Relevantes\ e\ Retornados\left(R^{c}_{j\leq k,i}\right)$ das equações Equação 2 e Equação 3 . Portanto, obtemos uma precisão para esta categoria de 0,67.

Para a Revocação, o parâmetro $\sum_{i=1}^{A} Relevantes\left(R^{c}_{j\leq k,i}\right)$ expressa a quantidade de artigos que deveriam ser retornados pelo classificador, no exemplo, estes foram representados pelas letras A, B, C e E. Assumindo o intervalo $k$ de categorias associadas aos artigos, esta medida assumiu o valor de 0,50. Uma vez calculadas as médias aritméticas da precisão e revocação das categorias avaliadas no estudo, é possível, portanto, realizar o cálculo da medida *f-score* demonstrada na Equação 4 .

**Tabela 4 - Exemplificação das medidas de precisão, revocação e *f-score* para a indexação de artigos.**

| Categoria: Statistics & Probability | | |
|---|---|---|
| **Portal ISI Web of Knowledge** | **Classificador Automatizado Relevância 1ª a 5ª posição** | **Classificador Automatizado Relevância 6ª a 30ª posição** |
| Artigo A → | Artigo A (verdadeiro positivo) (1º posição) | |
| Artigo B → | | Artigo B (falso negativo) (8º posição) |
| Artigo C → | Artigo C (verdadeiro positivo) (3º posição) | |
| | Artigo D (falso positivo) (5º posição) | |
| Artigo E → | | Artigo E (falso negativo) (15º posição) |
| | | Artigo F (verdadeiro negativo) (6º posição) |

*Análises estatísticas*

A análise da independência dos diferentes resultados alcançados pelas combinações de parâmetros apresentadas ao classificador foi realizada por meio dos testes Wilcoxon signed-rank [Bauer 1972] e T-pareado [Altman 1990]. A restrição da distribuição normal das variáveis avaliadas, exigida pelo teste T-pareado, foi constatada pelo teste estatístico Shapiro-Wilk [Royston 1982].

Os testes T pareado e Wilcoxon signed-rank verificaram valores médios de acerto entre os grupos avaliados quanto à indexação de artigos científicos. A opção pareada de tais testes foi considerada a fim de realizar a correspondência de exemplos entre as estratégias propostas. A hipótese nula dos testes considerou que a diferença média entre os grupos foi igual a zero. Portanto, valores de $p < 0,05$ (95%) rejeitaram tal hipótese.

## 4. Resultados e Discussão

As medidas de desempenho consideraram a posição de relevância da categoria em relação ao artigo científico, atribuída pelo classificador de padrões. Também foi comparado o desempenho do classificador quando diferentes métodos de extração de características foram apresentados como parâmetro de entrada para o mesmo, bem como a utilização da competição de técnicas.

A Tabela 5 mostra a média aritmética e o desvio padrão da medida de desempenho aplicada às técnicas utilizadas, representada pela coluna $f_{0,5}$*score*, quando a atribuição de categorias realizada pelo classificador variou entre a primeira e a quinta posição de relevância. O maior valor alcançado foi destacado com o caractere "*". Nesta tabela também é apresentada a análise estatística referente à distribuição normal dos resultados em relação a esta medida, representada pela coluna "p-values $f_{0,5}$-*score*".

**Tabela 5 - Resultados obtidos utilizando os métodos de extração de características e Competição de técnicas.**

| **Método** | ***$f_{0,5}$-score*** | **p-values $F_{0,5}$-score** |
|---|---|---|
| tf.idf | 0,60 ± 0,21 | 0,57* |
| tf | 0,52 ± 0,17 | 0,28* |
| to | 0,55 ± 0,16 | 0,01 |
| bo | 0,58 ± 0,17 | 0,01 |
| Competição de técnicas | 0,66 ± 0,17* | 0,57* |

Legenda: * Distribuição normal

A diferença estatística entre os resultados encontrados para a medida $f_{0,5}$-*score* pode ser visualizada na Tabela 6 , que considerou significativo os valores de $p < 0,05$.

**Tabela 6 - Diferença estatística entre os resultados quando avaliado o valor $f_{0,5}$-*score*.**

| | $f_{0.5}$-*score* | | | |
|---|---|---|---|---|
| **Métodos** | **tf.idf** | **Tf** | **to** | **bo** |
| **Competição** | 0.0001 | 0.0001 | 0.0101 | 0.0455 |
| **tf.idf** | | 0.0001 | 0.06884 | 0.04255 |
| **tf** | | | 0.0045 | 0.0006 |
| **to** | | | | 0.0486 |

Os experimentos contaram com a variação nos parâmetros de entrada do classificador, mediante a utilização de diferentes técnicas de extração de características dos documentos oriundos da base de dados construída para o estudo e, também, explorou-se a competição de técnicas.

Os desvios-padrão encontrados nas avaliações de desempenho das técnicas tf.idf, tf, to e bo motivaram a utilização da competição de técnicas.

Estudos como o de Lan et al [Lan et al. 2006] mostraram que diferentes técnicas de atribuição de pesos aos termos que compõem os vetores de características utilizados pelos classificadores podem ter um grau de influência maior nos resultados do que a escolha do próprio classificador. O método que utilizou a competição de técnicas apresentou melhores resultados, na qual a pontuação $f_{0,5}$-*score* alcançou o valor de 0,66.

A distribuição dos artigos científicos e categorias em relação à indexação sugerida às revistas pelo Portal Web of Knowledge, de acordo com a divisão dos domínios construída neste estudo, mostra que somente as categorias “Computer Science, Information Systems” e “Information Science & Library Science” compartilham artigos publicados em revistas dos domínios da Ciência da Computação e Informática em Saúde, enquanto apenas a categoria “Medicine, Research & Experimental” foi compartilhada entre os artigos publicados em revistas dos domínios da Saúde e Informática em Saúde.

A distribuição após a aplicação do método de indexação por meio de competição de técnicas mostrou um maior compartilhamento entre as categorias e domínios, o que sugere uma incompatibilidade entre a categorização original das revistas sugeridas pelo portal ISI Web of Knowledge e a proposta deste estudo. De acordo com os resultados, o método proposto indica que uma parte dos artigos reflete parcialmente a categorização atribuída às revistas pelo portal ISI Web of Knowledge.

A indexação incorreta e/ou incompleta de revistas ou artigos científicos pode prejudicar a recuperação de informação, uma vez que as categorias são utilizadas como parâmetros em sistemas de buscas construídos pelos Portais. Spreckelsen [Spreckelsen et al. 2011], em seu trabalho, destacou a importância do corpo de conhecimento de Informática em Saúde, disponível nas bibliotecas virtuais, ser cuidadosamente delimitado por meio das revistas e artigos publicados, pois os índices que medem o fator de impacto da área são amparados nos mesmos, sendo que uma fraca indexação comprometeria tais índices.

## 5. Conclusão

A indexação de artigos científicos sob uma lista pré-definida de categorias caracterizou o objetivo proposto no trabalho. Este objetivo também contou com a exploração de métodos para a construção dos vetores de características e posterior utilização como parâmetro de entrada do classificador utilizado no estudo, além disso, utilizou a competição de técnicas, que apresentou o melhor resultado. O desempenho alcançado foi satisfatório ao indexar artigos de um domínio interdisciplinar como o da Informática em Saúde.

Embora os resultados tenham sido satisfatórios, a dimensão (35 mil) dos vetores de características utilizados para representarem os documentos merece atenção. A literatura expõe alternativas para a redução da dimensionalidade de tais vetores, que não foram contempladas neste estudo, como o trabalho de Yang [Yang, Y. and Pedersen 1997], que explora e compara técnicas capazes de selecionar características de documentos para a tarefa de classificação automática.

Trabalhos futuros serão destinados à disponibilização dos mecanismos automatizados criados à comunidade científica, por meio de serviços que auxiliem profissionais que atuam na indexação de conteúdos em bibliotecas virtuais, pesquisadores que conduzem trabalhos científicos a encontrar informação relevante e demais aplicações aplicadas à mineração de textos nos domínios abordados neste estudo.

## 7. Referências


Altman, D. G. (1990). *Practical Statistics for Medical Research*. 1st ed ed. Chapman and Hall/CRC.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)*. 2. ed. in Text Operations: Addison-Wesley Professional.

Bauer, D. F. (1972). Constructing Confidence Sets Using Rank Statistics. *Journal of the American Statistical Association*, v. 67, n. 339, p. 687–690.

Bernstam, E. V., Smith, J. W. and Johnson, T. R. (Fevereiro 2010). What is biomedical informatics? *Journal of Biomedical Informatics*, v. 43, n. 1, p. 104–110.

Gehanno, J.-F., Rollin, L., Jean, T., et al. (apr 2009). Precision and Recall of Search Strategies for Identifying Studies on Return-To-Work in Medline. *Journal of Occupational Rehabilitation*, v. 19, n. 3, p. 223–230.

Knaup, P. and Dickhaus, H. (2009). Perspectives of medical informatics: advancing health care requires interdisciplinarity and interoperability. Special topic on the occasion of the 35th anniversary of the Heidelberg/Heilbronn curriculum of medical informatics. *Methods of Information in Medicine*, v. 48, n. 1, p. 1–3.

Lan, M., Tan, C.-L. and Low, H.-B. (2006). Proposing a new term weighting scheme for text categorization. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*. . AAAI Press. http://portal.acm.org/citation.cfm?id=1597538.1597660, [accessed on Apr 29].

Magdy, W. and Jones, G. (2010). PRES: a score metric for evaluating recall-oriented information retrieval applications. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. . ACM. http://dx.doi.org/10.1145/1835449.1835551, [accessed on May 5].

Mantas, J., Ammenwerth, E., Demiris, G., et al. (7 jan 2010). Recommendations of the International Medical Informatics Association (IMIA) on Education in Biomedical and Health Informatics. First Revision. *Methods of Information in Medicine*, v. 49, n. 2, p. 105–120.

Nigam, K. and McCallum (1998). A comparison of event models for Naive Bayes text classification.

Pacheco, E. J., Nohama, P. and Schulz, S. (2009). Mapping of Clinical Documentation to Ontology. In *IX Workshop de Informática Médica*.

Porter, M. F. (1997). An algorithm for suffix stripping. Morgan Kaufmann Publishers Inc. p. 313–316.

Radlinski, F. and Craswell, N. (2010). Comparing the sensitivity of information retrieval metrics. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. , SIGIR '10. ACM.

Royston, J. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, v. 31, n. 2.

Salton, G., Wong, A. and Yang, C. S. (nov 1975). A vector space model for automatic indexing. *Communications of the ACM*, v. 18, p. 613–620.

Salton, Gerard and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *INFORMATION PROCESSING AND MANAGEMENT*, v. 24, p. 513–523.

Sohn, S., Kim, W., Comeau, D. C. and Wilbur, W. J. (aug 2008). Optimal training sets for Bayesian prediction of MeSH assignment. *Journal of the American Medical Informatics Association: JAMIA*, v. 15, n. 4, p. 546–553.

Spreckelsen, C., Deserno, T. and Spitzer, K. (2011). Visibility of medical informatics regarding bibliometric indices and databases. *BMC Medical Informatics and Decision Making*, v. 11, n. 1, p. 24.

Van Bemmel, J. H. (2008). Medical Informatics Is Interdisciplinary avant la Lettre. *Methods of Information in Medicine*,

Yang, Y. and Pedersen, J. (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*. . Morgan Kaufmann Publishers, San Francisco, US. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.9956, [accessed on Aug 20].

Zhang, W., Yoshida, T. and Tang, X. (mar 2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, v. 38, n. 3, p. 2758–2765.