DESCOBERTA DE CONHECIMENTOS SOBRE A ESQUISTOSSOMOSE A PARTIR DE DOCUMENTOS CIENTÍFICOS UTILIZANDO TÉCNICAS DE MINERAÇÃO DE TEXTOS

Email:
dgsamc@gmail.com
ludmillarsouza@gmail.com
rosangela@inforium.com.br
rene.veloso@unimontes.br

Douglas Andrey de Oliveira Araújo[1]; Ludmilla Regina de Souza David[2]; Rosângela Silqueira Hickson Rios[3]; Renê Rodrigues Veloso[4]

Resumo

O presente estudo objetivou aplicar técnicas de mineração de texto para a descoberta de conhecimentos sobre a esquistossomose a partir de documentos científicos disponíveis no acervo Memórias do Instituto Oswaldo Cruz. Trata-se de um estudo retrospectivo e exploratório da base de dados memorias.ioc.fiocruz.br , a partir da qual foram obtidos 179 resumos de artigos científicos publicados no período de 2005 a 2015, selecionados pelo descritor "schisto". Foram utilizadas técnicas de mineração de textos (text mining) com o auxílio do software Weka 3.6.12 (Universidade de Waikato – Nova Zelândia). A extração de termos mais freqüentes foi realizada pelo método de indexação Frequência do Termo – Frequência Inversa de Documentos (TF-IDF). A categorização de documentos, por sua vez, aconteceu por dois métodos: classificação por aprendizado de máquina, utilizando árvores de decisão, e classificação por agrupamento. Posteriormente, o conhecimento revelado foi avaliado por especialistas da área e foram fornecidos indicadores sobre o desempenho do método desenvolvido.

Palavras-chave: Classificação de documentos. Extração de termos. Schistosoma. Aprendizado de Máquina. Classificação por agrupamento.

Abstract

This research aimed the knowledge discovery of schistosomiasis from scientific documents available on Web Site Memórias do Instituto Oswaldo Cruz. This is a retrospective, exploratory study of database memorias.ioc.fiocruz.br, which has been used to get 179 summaries of scientific documents, published between 2005 and 2015, selected by descriptor "schisto". To get knowledge discovery it has been applied Text Mining techniques supported by software Weka 3.6.12,

[1] Mestrando do Programa Pós-Graduação Stricto sensu em Tecnologia da Informação aplicada à Biologia Computacional da Faculdade Promove de Tecnologia.
[2] Professora Doutora do Programa Pós-Graduação Stricto sensu em Tecnologia da Informação aplicada à Biologia Computacional da Faculdade Promove de Tecnologia.
[3] Professora Doutora do Programa Pós-Graduação Stricto sensu em Tecnologia da Informação aplicada à Biologia Computacional da Faculdade Promove de Tecnologia.
[4] Professor Doutor do curso de Sistemas de Informação da Universidade Estadual de Montes Claros.

(University of Waikato – New Zealand). The extraction of most frequent terms was carried out by indexing method called Term Frequency – Inverse Do-cument Frequency (TF-IDF). The categorizing of documents, in turn, took place through two methods: classification by machine learning using decision tree and classification by clustering. Thereafter, the revealed knowledge was evaluated by domain specialists and were provided indicators regarding performance of methods applied.

Keywords: Classification of documents. Terms extraction. Schistosome. Machine Learning. Classification by clustering.

1 INTRODUÇÃO

A esquistossomose humana é considerada uma Doença Tropical Negligenciada (DTN) causada por vermes trematódeos do gênero *Schistosoma*, popularmente conhecida no Brasil como barriga d'água, xistose ou bilharziose. De acordo com o Ministério da Saúde do Brasil, a magnitude de sua prevalência associada à severidade das formas clínicas e a sua evolução conferem à esquistossomose uma grande relevância enquanto problema de saúde pública (Ministério da Saúde, 2014). Estimativas da Organização Mundial da Sáude (OMS) revelaram que mais de 258 milhões de pessoas têm sido infectadas em 78 países considerados endêmicos localizados na África Subsaariana, Oriente Médio, Caribe e América do Sul, resultando em cerca de 200.000 mortes por ano (NEVES *et al*., 2015; OMS, 2016). Estima-se que no Brasil cerca de 1,5 milhões de pessoas vivem em áreas sob o risco de contrair a doença. As áreas endêmicas e focais brasileiras abrangem 19 unidades federadas e compreendem os estados de Alagoas, Bahia, Pernambuco, Rio Grande do Norte (faixa litorânea), Paraíba, Sergipe, Espírito Santo e Minas Gerais (BRASIL, 2014).

Há muita informação disponível sobre a esquistossomose em formato de textos científicos. O conhecimento da temática e atualizações acerca da doença podem ser eficientemente revelados por meio de técnicas de mineração de textos. A mineração de textos é o processo de descoberta de conhecimento útil em textos. As principais contribuições dos estudos de mineração de textos estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e à melhor compreensão dos textos disponíveis em documentos (MORAIS; AMBROSIO, 2007).

Considerando-se a imensa quantidade de conhecimentos presentes nos diversos documentos científicos publicados nos últimos onze anos no site memorias.ioc.fiocruz.br, e que estes documentos podem conter informações importantes para os estudiosos, o presente estudo objetivou criar uma base de conhecimentos sobre a esquistossomose e classificar a informação utilizando-se de técnicas de mineração de textos.

2 FUNDAMENTAÇÃO TEÓRICA

Mineração de textos (*Text Mining*) é o processo de busca e extração de informação útil em coleções de textos (bases de dados textuais), por meio da identificação e exploração de padrões encontrados nestes textos (FELDMAN; SANGER, 2007). Por motivos históricos, a mineração de textos herdou conceitos e ferramentas da mineração de dados tradicional, e por

isso mantém semelhanças em seus processos com uma peculiaridade essencial: enquanto na mineração de dados as informações são totalmente estruturadas, na mineração de textos as informações são de origem não estruturada ou semi-estruturada, isto é, textos (FELDMAN; SANGER, 2007).

Conforme Morais e Ambrósio (2007), a mineração de textos pode ser considerada sinônimo de Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts*). Essa afirmação é corroborada pelos trabalhos de Feldman e Dagan (1995) e Lopes (2004), que sugerem as utilizações das terminologias Mineração de Dados em Textos (*Text Data Mining*) e Descoberta de Conhecimentos a partir de Bancos de Dados Textuais (*Knowledge Discovery from Textual Databases*).

Em geral, o processo de mineração de textos é composto pelas seguintes etapas: 1) definição do objetivo, 2) levantamento dos textos, 3) pré-processamento, 4) indexação enormalização, 5) cálculo da relevância dos termos, 6) seleção dos termos, 7) pós-processamento ou análise dos resultados (MORAIS; AMBRÓSIO, 2007; BHUMIKA; SUKHJIT; NAYYAR, 2013).

Existem duas abordagens típicas para a execução da mineração de textos: a análise semântica, baseada na funcionalidade dos termos encontrados nos textos, e a análise estatística, baseada na freqüência dos termos encontrados nos textos. Estas abordagens podem ser utilizadas separadamente ou em conjunto (EBECKEN; LOPES; COSTA, 2003). Para estudos que utilizam coleções de documentos, a abordagem estatística é a mais apropriada, pois provê informações necessárias sobre associações entre palavras e documentos, visto que essa é condição para as tarefas de categorização (FELDMAN; SANGER, 2007).

3 METODOLOGIA

A metodologia utilizada neste trabalho respeitou as etapas comuns observadas em processos de Mineração de Textos segundo Morais e Ambrósio (2007) e Bhumika, Sukhjit e Nayyar (2013). O objetivo foi descobrir conhecimentos sobre esquistossomose em textos científicos, por meio de coleções de termos extraídos e categorização dos documentos avaliados.

A fonte do *corpora* foi o site memoriasioc.fiocruz.br devido ao seu foco em doenças tropicais negligenciadas, sobretudo a esquistossomose, com vasto e respeitado acervo de publicações nacionais e internacionais. O levantamento dos resumos de artigos científicos ocorreu no sistema de busca do próprio site, utilizando-se como descritor o termo "*schisto*", perfazendo um total de 179 resumos de artigos científicos publicados no período de 2005 a 2015. Todos os artigos científicos analisados apresentaram resumo em língua inglesa e foram analisados nesse idioma.

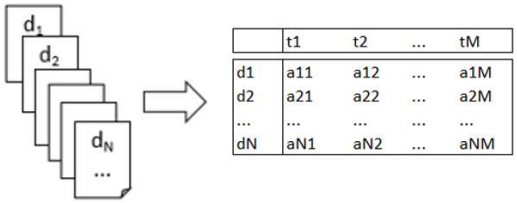
Para o pré-processamento foram selecionados resumos dos artigos científicos no site memoriasioc.fiocruz.br em formato *Portable Document Format* (PDF), os quais foram convertidos no formato Arquivo Texto Plano (TXT), utilizando-se a ferramenta "*Free PDF to Text Converter* – 3.0 ®". Como a ferramenta principal de mineração de textos foi o Weka 3,6.12, os arquivos em formato TXT foram convertidos para o formato do Weka, o *Attribute-Relation File Format (ARFF)*. Através do filtro TextDirectoryToArff, todos os arquivos de documentos foram convertidos em um único arquivo *dataset*.

A identificação de termos (*tokenization*) foi utilizada para a definição do tipo de *Token*. Nesta pesquisa foram escolhidos termos compostos, em especial os bigramas. Posteriormente realizou-se a remoção de termos considerados irrelevantes (*Stop-words)*. As *stopwords* fazem parte de uma lista controlada conhecida como *stoplist*. Esta pesquisa utilizou

stoplists baixadas do site *Google Code Archive* (https://code.google.com/p/stop-words) e posteriormente atualizadas conforme necessidade de filtragem. Recursos de normalização morfológica das palavras (*stemming*) não foram utilizados neste estudo, uma vez que, na fase de testes, não contribuiram efetivamente para a formação de bigramas úteis.

Após a criação dos arquivos em formato ARFF, criou-se a matriz documento-termo (Figura 1), onde cada documento foi representado e indexado por um vetor multidimensional, e cada dimensão foi representada por um termo da coleção (Feldman e Sanger, 2007). Essa matriz documento-termo seguiu a estrutura conhecida como *bag-of-words*, onde o documento original se tornou um conjunto de palavras independentes extraídas do texto, mas sem manter a ordem original de onde foram retiradas.

Figura 1 – Representação esquemática da coleção de documentos transformada em uma matriz documento-termo.



Legenda: d → documento, t → termo.
Fonte: Elaborada pelos autores.

As limitações da matriz documento-termo, relacionadas à alta dimensionalidade da matriz (*overfitting*) e a perda do relacionamento semântico entre os termos foram minimizadas por meio da realização de cálculos do peso dos termos (MARCACINI, 2011). Cada documento na matriz documento-termo é representado por um vetor di=(ai1, ai2, ... aim), onde o valor de aij pode indicar a presença/ausência do termo, ou pode indicar a importância do termo dentro da coleção de documentos.

O cálculo da frequência do termo (*Term Frequency* – TF) foi baseado nas seguintes abordagens: frequência relativa do termo e frequência inversa do termo nos documentos, conhecido como TF-IDF. Na frequência relativa (Frel), o cálculo da frequência do termo considerou a frequência absoluta (Fabs) do termo no documento, adotando-se a normalização da quantidade de palavras do documento. A normalização no Weka é dada pela razão entre o tamanho médio dos documentos na coleção e o tamanho de cada documento. No Weka 3.6.12, a frequência do termo foi dada pela fórmula: TF = log (1+Frel).

A frequência inversa nos documentos (*Inverse Document Frequency* – IDF) foi definida considerando-se também a frequência inversa do termo em vários documentos da coleção. Dessa forma, foi possível aumentar a relevância dos termos que apareceram em poucos documentos e diminuir a importância dos termos que aparecem em muitos documentos. No software Weka 3.6.12, a frequência inversa dos documentos foi estabelecida da seguinte forma: IDF = TF x log (TF/N), onde TF corresponde à frequência normalizada do termo e N representa o número de documentos onde o termo apareceu.

Na etapa de seleção, os termos foram selecionados conforme o grau de importância, ordenando-se os termos de forma decrescente conforme a sua frequência. Adicionalmente, a lista de termos foi avaliada por um dos autores da pesquisa, Souza LR., e assim os termos menos relevantes foram removidos. Em seguida, foram selecionados os termos mais representativos por ano e os termos representativos de todo o período de 2005-2015.

Para criar a árvore de decisão, foram utilizados documentos previamente classificados pelo usuário como insumo para o aprendizado de máquina do algoritmo C4.5. Os documentos

fornecidos pelo especialista como base de treinamento do algoritmo foram classificados em: 1) Epidemiologia da esquistossomose, 2) Biologia molecular do *Schistosoma*, 3) Diagnóstico da esquistossomose, 4) Outros. O classificador J48 do software Weka foi utilizado para aplicar o C4.5 e assim a classificação dos documentos foi realizada por predição.

A classificação dos documentos por agrupamento foi realizada pelo algoritmo *Kmeans*. No Weka 3.6.12 o Kmeans foi aplicado via agrupador *SimpleKmeans*. Os autores do estudo realizaram simulações para encontrar os grupos mais adequados de documentos utilizando a lista de termos selecionados. Todos os termos foram comparados entre si e alocados em cada grupo conforme a distância (similaridade) entre eles.

O conhecimento descoberto foi avaliado de forma subjetiva por especialistas da área e de forma objetiva por meio de índices estatísticos que indicam a eficiência dos métodos de mineração de textos e a qualidade dos resultados. A avaliação de especialistas foi realizada por meio de um questionário estruturado, que considerou aspectos referentes à novidade e utilidade das informações extraídas. Os termos que mais agregaram valor ao conhecimento bem como os termos desconhecidos dentro do contexto da esquistossomose foram apontados pelos avaliadores. Além disso, o conhecimento revelado pela mineração de textos foi avaliado quanto à capacidade de indicar a evolução das pesquisas no período de estudo e a utilidade como referência em buscas exploratórias. As análises foram realizadas com as opiniões de seis especialistas: 4 doutores e 2 mestres, todos atuantes na área da esquistossomose.

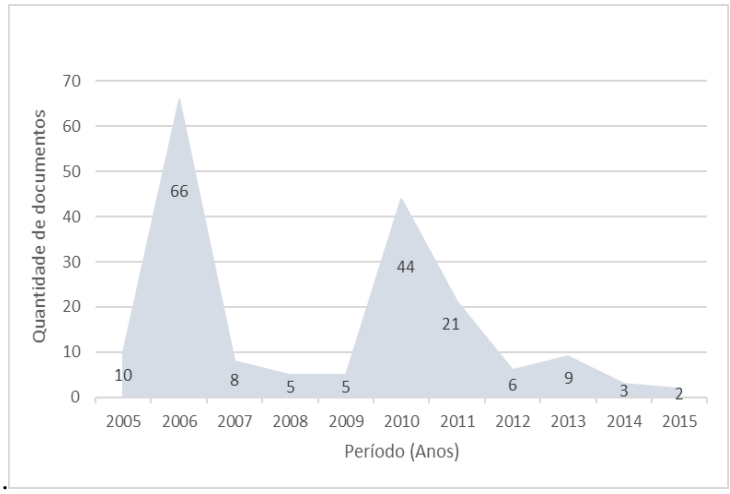
A classificação de documentos por árvore de decisão foi avaliada de duas formas: 1) os documentos classificados foram avaliados pela doutora participante do estudo: Souza, LR., que comparou a classificação do sistema com a sua própria classificação e indicou correções no modelo de classificação e 2) os seguintes indicadores foram utilizados para avaliação das classificações: o percentual geral de instâncias classificadas correta e incorretamente, a matriz de confusão (*confusion matrix*) que exibe a distribuição das classificações (correta e incorreta) pelas classes, os indicadores de acuracidade *Precision* e *Recall* e o indicador ROC Area, que indica a precisão/sensibilidade da classificação quanto à taxa de positivos verdadeiros e de positivos falsos.

A classificação de documentos por agrupamento considerou o conhecimento de Souza, LR.(participante do projeto), por meio da leitura dos documentos de cada grupo e observação das características comuns para identificação os padrões existentes.

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

O presente estudo analisou 179 resumos de artigos científicos do acervo Memórias do Instituto Oswaldo Cruz publicados entre os anos de 2005-2015. Observou-se um predomínio de artigos publicados no ano de 2006 (n=66, 37%), seguido pelos anos de 2010 (n=44, 25%) e 2011 (n=21, 12%). A quantidade de artigos avaliados por ano de publicação pode ser observado na figura 2.

Figura 2 – Quantidade de artigos científicos sobre *Schistosoma* publicados no acervo Memórias do Instituto Oswaldo Cruz entre os anos de 2005-2015



Fonte: Dados da pesquisa.

Os termos extraídos dos documentos foram gerados em bigramas e a relação de termos representativos de cada ano é apresentada na tabela 1. Os termos representativos de todo período avaliado são apresentados na tabela 2. Ressalta-se que os termos que não agregaram valor ao contexto do estudo como, por exemplo, "*This paper*", "*among the*", "*associated with*", entre outros, foram removidos das listas de termos. As listas de termos estão representadas em ordem decrescente de acordo com sua frequência (TF-IDF).

Tabela 1 – Relação de bigramas extraídos dos documentos organizados por frequência e representados por ano (2005 - 2010)

Pos	Termos/2005	F	Termos/2006	F	Termos/2007	F	Termos/2008	F	Termos/2009	F	Termos/2010	F
1	'hepatic stellate'	39%	'schistosomiasis mansoni'	29%	'acute schistosomiasis'	38%	'DNA extraction'	57%	'PCR assay'	58%	'Schistosoma mansoni'	31%
2	'granuloma formation'	37%	'mansoni infection'	28%	'oxamniquine treatment'	30%	'SWAP stimulation'	39%	'mansoni DNA'	46%	'mansoni infection'	28%
3	'schistosomal granuloma'	33%	'Biomphalaria glabrata'	28%	'aspartic protease'	30%	'Plasmodium berghei'	34%	'urinary schistosomiasis'	35%	'liver fibrosis'	24%
4	'mansoni infection'	33%	'Schistosoma mansoni'	28%	'B tenagophila'	29%	'antischistosomal activity'	30%	'urine reagent'	35%	'Biomphalaria glabrata'	24%
5	'worm burden'	32%	'health education'	19%	'vaccine formulation'	29%	'S mansoni'	29%	'Kato-Katz slides'	33%	'Biomphalaria tenagophila'	22%
6	'Biomphalaria glabrata'	30%	'Schistosomiasis Control'	17%	'Schistosoma mansoni'	26%	'schistosomiasis epidemiology'	29%	'chronic schistosomiasis'	31%	'with praziquantel'	19%
7	'freshwater snail'	30%	'a vaccine'	16%	'Schistosoma haematobium'	23%	'schistosomiasis mansoni'	25%	'serum cholesterol'	31%	'potato apyrase'	18%
8	'snail Biomphalaria'	30%	'control programs'	16%	'Biomphalaria glabrata'	23%	'interferon-gamma IFN-Î³'	25%	'Schistosoma mansoni'	30%	'hepatic fibrosis'	16%
9	'Schistosoma mansoni'	30%	'endemic area'	14%	'Computerized morphometric'	20%	'interleukin IL'	25%	'Biomphalaria oligoza'	30%	'magnetic resonance'	15%
10	'Citrus reticulata'	29%	'S haematobium'	13%	'anti-angiogenesis drug'	20%	'stimulation hepatosplenic'	25%	'Biomphalaria peregrina'	30%	'portal hypertension'	14%
11	'enzyme activities'	29%	'hepatosplenic schistosomiasis'	13%	'hepatic schistosomiasis'	20%	'Schistosomiasis mansoni'	22%	'Biomphalaria species'	30%	'population movement'	14%
12	'IgG antibodies'	27%	'Biomphalaria tenagophila'	13%	'schistosomiasis Computerized'	20%	'Schistosoma co-infection'	21%	'Biomphalaria tenagophila'	30%	'hepatosplenic schistosomiasis'	14%
13	'IgM antibodies'	27%	'portal vein'	12%	'manifestations eosinophilia'	19%	'berghei immunization'	21%	'Gastropoda Planorbidae'	30%	'chemokine receptors'	14%
14	'antioxidant enzymes'	24%	'circulating hemocytes'	12%	'Schistosoma japonicum'	19%	'experimental malaria'	21%	'LE Schistosoma'	30%	'ATP diphosphohydrolase'	11%
15	'glucose metabolism'	24%	'Sm22 6'	12%	'hemoglobin-degrading proteases'	19%	'malarial parasitaemia'	21%	'characterising Biomphalaria'	30%		
16	'sativa seeds'	24%	'Rainforest Zone'	11%	'mammalian hemoglobins'	19%	'and lovastatin'	20%	'guaibensis Biomphalaria'	30%		
17	'Biomphalaria spp'	19%	'schistosomal myeloradiculopathy'	11%	'Schistosomiasis mansoni'	18%	'lovastatin action'	20%	'mansoni strain'	30%		
18	'Melanoides tuberculatus'	19%	'cercarial shedding'	11%	'DNA-based vaccine'	18%	'and clonazepam'	19%	'molecular PCR-RFLP'	30%		
19	'environmental factors'	19%	'tenagophila Taim'	10%	'IgG2a Protamine'	18%	'and praziquantel'	19%	'molluscs Gastropoda'	30%		
20	'habitat preference'	19%	'human infection'	10%	'Protamine sulphate/DNA'	18%	'oxamniquine praziquantel'	19%	'peregrina Specimens'	30%		
21	'snail Melanoides'	19%	'mortality rate'	10%	'IgG1 IgG3'	15%	'treatment clonazepam'	19%	'tenagophila guaibensis'	30%		
22	'Biomphalaria tenagophila'	19%	'potato apyrase'	10%	'IgG2 IgG3'	15%	'Schistosoma mansoni'	16%	'DNA extraction'	29%		
23	'Anisakis sensitisation'	18%	'anti-potato apyrase'	9%	'anti-malaria IgG'	15%			'sensitive PCR'	29%		
24	'antigen purified'	18%	'mansoni DNA'	9%	'anti-merozoite surface'	15%			'Twelve Kato-Katz'	21%		
25	'simplex antigen'	18%	'chain reaction'	8%	'antigen IgG2'	15%			'cercariae Their'	20%		
26	'immunosorbent assay'	18%	'polymerase chain'	8%	'falciparum antigens'	15%			'chronic murine'	20%		
27	'Anisakis simplex'	15%	'B amazonica'	7%	'haematobium Antibody'	15%			'mansoni cercariae'	20%		
28	'PAK antigen'	15%	'time polymerase'	7%	'malaria antigens'	15%			'murine schistosomiasis'	20%		
29	'PAS antigens'	15%			'protein Glurp-R0'	15%			'Nigeria schistosomiasis'	18%		
30	'crude extract'	15%			'protein-3 MSP3b'	15%			'Schistosoma haematobium'	18%		
31	'helminthic infections'	15%							'Urine samples'	18%		
32	'simplex crude'	15%							'macrohaematuria microhaematuria'	18%		

Fonte: Dados da pesquisa.
*P – posição; F – frequência.

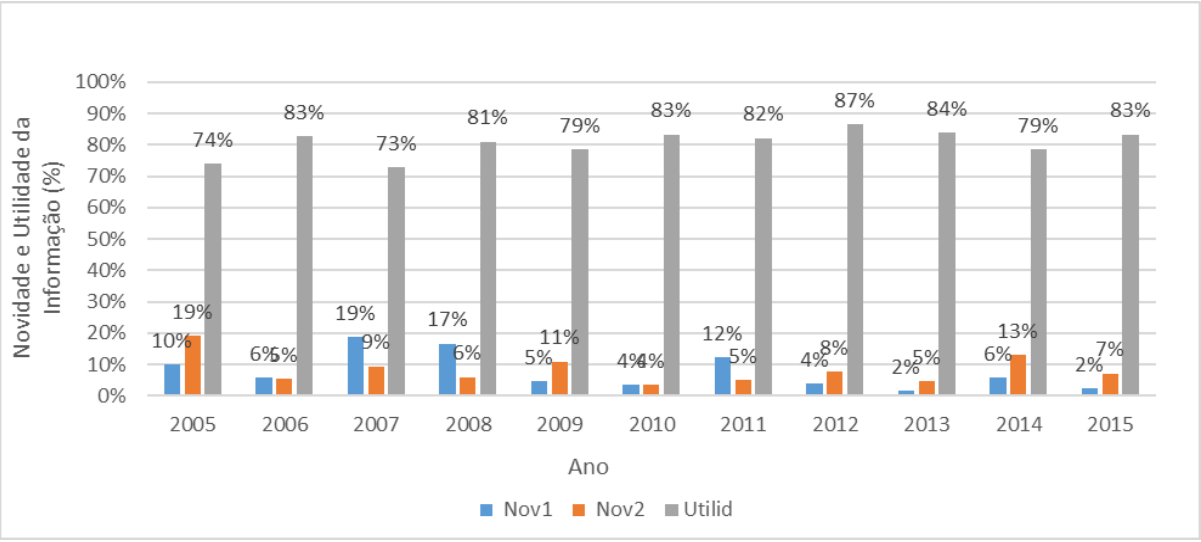
Tabela 1 (continuação) – Relação de bigramas extraídos dos documentos organizados por frequência e representados por ano (2011 - 2015)

Pos	Termos/2011	F	Termos/2012	F	Termos/2013	F	Termos/2014	F	Termos/2015	F
1	'SmATPDase 2'	27%	'low-intensity transmission'	60%	'KK method'	52%	'Schistosoma mansoni'	39%	'KK slides'	48%
2	'peptidase families'	26%	'B glabrata'	52%	'S mansoni'	41%	'collagen synthesis'	39%	'Biomphalaria alexandrina'	38%
3	'mansoni antigens'	24%	'schistosomiasis mansoni'	49%	'Ca2+ channels'	39%	'hepatic periportal'	39%	'Schistosoma mansoni'	38%
4	'Biomphalaria glabrata'	24%	'Kato-Katz technique'	44%	'mansoni worms'	39%	'periportal fibrosis'	39%	'alexandrina snails'	38%
5	'secondary sporocysts'	24%	'B alexandrina'	44%	'tract excretions'	34%	'disease screening'	26%	'molecular techniques'	38%
6	'parasite antigens'	23%	'Kato-Katz method'	42%	'the Kato-Katz'	33%	'disease serology'	26%	'schistosomiasis mansoni'	38%
7	'colonic carcinoma'	23%	'amoebocyte-producing organ'	41%	'immunological assays'	31%	'health programme'	26%	'Biomphalaria control'	24%
8	'periportal fibrosis'	23%	'the amoebocyte-producing'	41%	'female worms'	28%	'migration Chagas'	26%	'Biomphalaria species'	24%
9	'endemic areas'	22%	'Biomphalaria glabrata'	35%	'SL trans-splicing'	28%	'Alhandra Faecal'	26%	'S mansoni'	24%
10	'immunosorbent assay'	22%	'alexandrina snails'	34%	'life stages'	28%	'Biomphalaria glabrata'	26%	'SOD1 enzyme'	24%
11	'linked immunosorbent'	22%	'HelmintexÃ® method'	33%	'calcium channel'	27%	'Kato-Katz method'	26%	'alleles function'	24%
12	'Schistosoma japonicum'	22%	'diagnosing schistosomiasis'	33%	'channel antagonist'	27%	'comparative cross-sectional'	26%	'cercarial production'	24%
13	'endopeptidase families'	21%	'gradient method'	33%	'schistosomiasis control'	27%	'excreted cercariae'	26%	'genetic background'	24%
14	'histone modifying'	20%	'identification methods'	33%	'Kappa index'	25%	'glabrata specimens'	26%	'genetic variability'	24%
15	'mansoni sporocysts'	19%	'parasite identification'	33%	'Kato-Katz analysis'	25%	'southern coastal'	26%	'higher SOD1'	24%
16	'during lactation'	19%	'saline gradient'	33%	'antischistosomal drugs'	25%	'specimens caught'	26%	'mansoni infection'	24%
17	'Egyptian patients'	18%	'chain reaction'	32%	'pre-patent phase'	25%	'TGF -Î²1'	25%	'mansoni transmission'	24%
18	'Schistosoma mansoni'	18%	'parasitological methods'	31%	'Microtus fortis'	23%	'Undernourished mice'	25%	'snail control'	24%
19	'mansoni infection'	17%	'Kato-Katz KK'	28%	'Schistosoma japonicum'	23%	'biology collagen'	25%	'species Biomphalaria'	24%
20	'mansoni SmATPDase'	17%	'Biomphalaria alexandrina'	27%	'serum albumin'	23%	'curves biology'	25%	'transmission modulation'	24%
21	'schistosome chromosomes'	17%	'genetic variability'	27%	'ELISA reactive'	22%	'eggshell glycoproteins'	25%	'Kato-KatzÂ® KK'	24%
22	'inflammatory response'	16%	'HPJ Willis'	24%	'a low-endemicity'	22%	'factor TGF'	25%	'PCR -ELISA'	24%
23	'Biomphalaria tenagophila'	16%	'coproscopic techniques'	24%	'low-endemicity area'	22%	'hepatic stellate'	25%	'chain reaction'	24%
24	'cellular adhesion'	16%	'polymerase chain'	24%	'parasitological examination'	22%	'mansoni Undernourished'	25%	'parasitological exams'	24%
25	'tenagophila Taim'	16%	'reaction PCR'	24%	'polymerase chain'	22%	'periovular granulomas'	25%	'polymerase chain'	24%
26	'cutaneous leishmaniasis'	15%	'Schistosoma mansoni'	20%	'Schistosoma mansoni'	21%	'protein synthesis'	25%	'population-based study'	24%
27	'PCR assay'	15%			'mansoni infection'	20%	'repair mechanisms'	25%	'reaction PCR'	24%
28	'carbohydrate moieties'	13%			'morphological alterations'	17%	'undernourished mice'	25%	techniques KK+TFTestÂ®'	24%
29	'Biomphalaria amazonica'	13%								
30	'Biomphalaria cousini'	13%								

Fonte: Dados da pesquisa.
*P – posição; F – frequência.

A avaliação das listas feita pelos especialistas de domínio considerou aspectos subjetivos como a Novidade e a Utilidade da informação. Abaixo é exibido o gráfico com a opinião dos especialistas para cada conjunto de dados anual. O indicador Nov1 exibe o percentual de novidade dos termos que já eram conhecidos, mas que surpreendeu o pesquisador ao ser encontrado naquela coleção. O indicador Nov2 exibe o percentual dos termos considerados completamente novos para o especialista. O indicador Utilid informa o percentual de utilidade dos termos conforme opinião do especialista.

Figura 3 – Avaliação dos aspectos da Novidade e Utilidade da informação pelos especialistas de domínio.



Fonte: Dados da pesquisa.

O resultado da avaliação indicou que, em geral, os termos já eram conhecidos pelos especialistas com um percentual médio de novidade igual a 8% (+/-0,06) para ambos indicadores (termos desconhecidos no contexto e termos completamente desconhecidos). Os termos que mais se destacaram foram 'PAK antigen' e 'PAS antigens' no ano de 2005 e *'Alhandra Faecal'* no ano de 2014. Eles foram identificados por 5 dos 6 especialistas como uma informação completamente nova para o especialista. Quanto ao aspecto da utilidade da informação, os especialistas julgaram que, em geral, o conjunto de termos é útil e capaz de despertar o interesse pela leitura dos documentos originários de onde se extraiu os termos – percentual médio de utilidade igual a 81% (+/-0,04). Os termos que mais agregaram valor ao conhecimento dos especialistas podem ser vistos na Tabela 4:

Tabela 4 – Termos que mais agregaram conhecimento segundo especialistas de domínio.

Termos 2005	Termos 2006	Termos 2007	Termos 2008	Termos 2009	Termos 2010
environmental factors	a vaccine	acute schistosomiasis	and praziquantel	LE Schistosoma	ATP diphosphohydrolase
enzyme activities	cercarial shedding	anti-angiogenesis drug	antischistosomal activity	Biomphalaria oligoza	Biomphalaria glabrata
Granuloma formation	control programs	B tenagophila	experimental malária	Biomphalaria peregrina	hepatic fibrosis
habitat preference	endemic área	Computerized morphometric	interferon-gamma IFN-Î	cercariae Their	liver fibrosis
IgG antibodies	health education	DNA-based vaccine	interleukin IL	chronic murine	mansoni infection
immunosorbent assay	hepatosplenic schistosomiasis	hepatic schistosomiasis	lovastatin action	chronic schistosomiasis	portal hypertension
mansoni infection	human infection	IgG1 IgG3	Oxamniquine praziquantel	DNA extraction	Schistosoma mansoni
Schistosoma mansoni	mansoni infection	IgG2 IgG3	Schistosoma co-infection	macrohaematuria microhaem	with praziquantel
schistosomal granuloma	potato apyrase	manifestations eosinophilia	schistosomiasis epidemiology	mansoni DNA	
snail Biomphalaria	Schistosomiasis Control	oxamniquine treatment	schistosomiasis mansoni	mansoni strain	
worm burden	schistosomiasis mansoni	Schistosoma mansoni	SWAP stimulation	molecular PCR-RFLP	
	Sm22 6	Schistosomiasis mansoni	treatment clonazepam	Nigeria schistosomiasis	
				Schistosoma mansoni	

Fonte: Dados da pesquisa

Tabela 4 (continuação) – Termos que mais agregaram conhecimento segundo especialistas de domínio.

Termos 2011	Termos 2012	Termos 2013	Termos 2014	Termos 2015
Biomphalaria amazonica	Biomphalaria amazonica	antischistosomal drugs	disease screening	Biomphalaria control
Biomphalaria cousini	Biomphalaria cousini	Ca2+ channels	disease serology	cercarial production
Biomphalaria glabrata	Biomphalaria glabrata	channel antagonist	eggshell glycoproteins	genetic background
carbohydrate moieties	carbohydrate moieties	ELISA reactive	excreted cercariae	mansoni infection
cellular adhesion	cellular adhesion	immunological assays	glabrata specimens	mansoni trasmission
immunosorbent assay	immunosorbent assay	Kappa index	health programme	molecular techniques
inflammatory response	inflammatory response	mansoni worms	hepatic periportal	parasitological exams
linked immunosorbent	linked immunosorbent	S mansoni	mansoni Undernourished	PCR -ELISA
mansoni antigens	mansoni antigens	schistosomiasis control	periovular granulomas	reaction PCR
mansoni infection	mansoni infection	SL trans-splicing	protein synthesis	Schistosoma mansoni
mansoni sporocysts	mansoni sporocysts		Schistosoma mansoni	schistosomiasis mansoni
parasite antigens	parasite antigens		TGF -Î²1	species Biomphalaria
PCR assay	PCR assay			techniquesKK+TFTestÂ
secondary sporocysts	secondary sporocysts			

Fonte: Dados da pesquisa

Quanto à capacidade dos termos indicarem a evolução das pesquisas no período de 2005-2015, a maioria dos especialistas (5 de 6) considerou que o conjunto de termos exibido é efetivo para este propósito. Os termos selecionados que indicam esta evolução são: *'Control programs'*, 'DNA *extraction'*, *'endemic areas'*, *'low-endemicity areas'*, *'magnetic resonance'*, *'molecular* PCR-RFLP', 'oxamniquine *treatment'*, 'PCR-ELISA', *'polymerase chain'*, *'population movement'*, *'worm burden'*. Finalmente, a maioria dos especialistas (5 em 6) considerou que o conhecimento extraído sobre a esquistossomose é capaz de guiar o processo de busca exploratória.

As listas de termos extraídos dos documentos funcionaram como um índice, capaz de representar o conteúdo e despertar o interesse no conteúdo dos documentos entre 2005 e 2015. Observou-se que o conjunto de termos é um mecanismo útil para se avaliar a evolução das pesquisas sobre a esquistossomose no período avaliado. De acordo com tais resultados pode-se perceber termos como 'S Mansoni' e 'schistosoma mansoni' que tem o mesmo significado. A adição de um vocabulário no processo de extração dos termos que considere palavras diferentes com o mesmo significado poderia reduzir o tamanho da lista. Há de se considerar que a percepção da utilidade da informação extraída varia quanto à experiência e expectativa do usuário (Silberchatz e Tuzhilin, 1995). Além disso, deve-se levar em conta que a seleção dos termos mais relevantes é algo crítico para o processo, pois consiste em definir a quantidade adequada de termos que representará os documentos sem que suas características mais importantes sejam perdidas. No corrente estudo, a seleção dos termos foi baseada no método de Luhn (1958), onde os termos de alta frequência são classificados como não relevantes por aparecerem na grande maioria dos textos e, em geral, não agregar valor para discriminar o texto. Da mesma forma, os termos de baixa frequência são considerados conteúdos pouco discriminatórios e também são eliminados. Assim, os termos de frequência intermediária são julgados mais significativos e considerados nas análises.

Importante ressaltar a influência da escolha do *Token* (MANNING *et. al.*, 2008) , que determina se os termos serão simples ou compostos. Neste estudo, a escolha de termos compostos por duas palavras (bigramas) foi definitiva para melhorar a clareza e a representatividade do termo dentro da coleção de documentos. Nos testes realizados observou-se que termos com somente uma palavra foram pouco informativos e contextualizados como por exemplo 'DNA' ao invés de 'mansoni DNA', assim como termos com mais de duas palavras foram inconsistentes e confusos como 'B *glabrata* was'. O *stemmer,* comumente utilizado para a normalização morfológica (BHUMIKA; SUKHJIT; NAYYAR, 2013; WIEVES, 2002), não contribuiu positivamente para a formação dos termos no presente estudo, porque tornou difícil seu reconhecimento quando exposto na relação final de termos, e por isso não foi utilizado, ex: "*schistosoma* in" ao invés de "*schistosoma* infection".

Além da lista de termos gerou-se a tabela documento-termo do período 2005-2015, onde cada documento foi representado pelos termos mais frequentes. A tabela 5 demonstra a relação entre o documento e os termos que o representam por meio de suas frequências. Ressalta-se que a tabela documento-termo é fundamental para a execução dos processos posteriores de classificação por árvore de decisão ou por agrupamento (FELDMAN; SANGER, 2007).

Tabela 5 – Amostras da matriz documento-termo

No	Doc	Biomphalaria glabrata	Biomphalaria tenagophila	Schistosoma mansoni	chain reaction	mansoni DNA	mansoni infection	polymerase chain	reaction PCR	tenagophila Taim	worm burden
1	Biomphalaria tenagophila: dominant character of the resistance to Schistosoma mansoni in descendants of crossbreedings between resistant (Taim, RS) and susceptible (Joinville, SC) strains		1,753901	0,463846			1,10387				
5	Isolation and characterization of the full-length cDNA encoding a member of a novel cytochrome p450 family (CYP320A1) from the tropical freshwater snail, Biomphalaria glabrata, intermediate host for Schistosoma mansoni	2,370764	2,010421	0,532937			1,268294				
10	Efficacy of Citrus reticulata and Mirazid in treatment of Schistosoma mansoni			0,49593			1,87061				3,780574
28	Protein tyrosine kinases in Schistosoma mansoni			0,356284	2,505451	4,018931		2,505451	1,64303		
34	A bacterial artificial chromosome library for Biomphalaria glabrata, intermediate snail host of Schistosoma mansoni	2,952719		0,526016							
36	Differential lectin labelling of circulating hemocytes from Biomphalaria glabrata and Biomphalaria tenagophila resistant or susceptible to Schistosoma mansoni infection	1,106454	1,490635	0,394221			1,876352			1,986259	

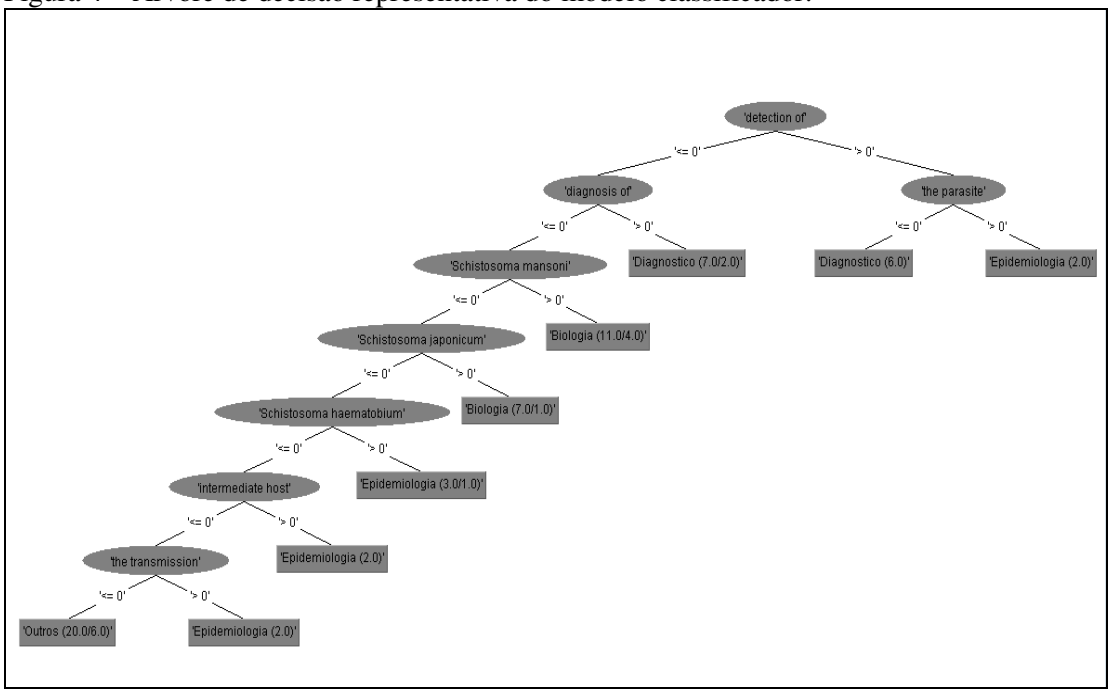
Fonte: Dados da pesquisa

Em média os documentos foram representados por 3 termos, variando entre 1 e 10 termos no total. Não há métrica ou referência para a quantidade ideal de termos por documento e cabe ao pesquisador encontrar a dimensionalidade ideal para a matriz, e a posterior descoberta de padrões pode contribuir para o refinamento na seleção dos termos (CORRÊA; MARCACINI; REZENDE, 2012).

A classificação dos documentos por árvore de decisão utilizou 60 documentos para treinamento do algoritmo, sendo 15 documentos para cada classe previamente definida: 1) Epidemiologia da esquistossomose, 2) Biologia do *Schistosoma*, 3) Diagnóstico da esquistossomose, 4) Outros. A quantidade de documentos classificados por esse método foi: epidemiologia da esquistossomose (n=6), biologia molecular do *Schistosoma* (n=109), Diagnóstico da esquistossomose (n=19) e outros (n=45).

O modelo extraído do conjunto de documentos que possibilitou a sua classificação ficou composto por 8 galhos (incluindo a raiz), 13 nós de decisão e 9 folhas (classificações) (Figura 4).

Figura 4 – Árvore de decisão representativa do modelo classificador:



Fonte: Dados da pesquisa

Figura 5 - Modelo classificador dos documentos:

```
detection of <= 0
|   diagnosis of <= 0
|   |   Schistosoma mansoni <= 0
|   |   |   Schistosoma japonicum <= 0
|   |   |   |   Schistosoma haematobium <= 0
|   |   |   |   |   intermediate host <= 0
|   |   |   |   |   |   the transmission <= 0: Outros (20.0/6.0)
|   |   |   |   |   |   the transmission > 0: Epidemiologia (2.0)
|   |   |   |   |   intermediate host > 0: Epidemiologia (2.0)
|   |   |   |   Schistosoma haematobium > 0: Epidemiologia (3.0/1.0)
|   |   |   Schistosoma japonicum > 0: Biologia (7.0/1.0)
|   |   Schistosoma mansoni > 0: Biologia (11.0/4.0)
|   diagnosis of > 0: Diagnostico (7.0/2.0)
detection of > 0
|   the parasite <= 0: Diagnostico (6.0)
|   the parasite > 0: Epidemiologia (2.0)

Number of Leaves  :     9

Size of the tree :      17
```

Fonte: Dados da pesquisa.

A descoberta de conhecimento por meio de classificações é uma estratégia de exposição rápida e concisa sobre um determinado conteúdo existente em uma coleção de documentos. Na etapa de classificação dos documentos utilizando Árvore de Decisão foi necessário a criação de um novo conjunto de termos, uma vez que uma nova base de documentos previamente classificada foi adicionada à coleção de documentos existentes. Durante os testes observou-se que termos comuns, como por exemplo "detection of", considerados na etapa anterior pouco representativos quanto ao significado, revelaram-se de grande contribuição na formação da árvore de decisão, porque geralmente possuiam boa frequência e distribuição dentro da coleção. Apesar dos esforços em melhorar o nível de acertos na classificação, a árvore ou o modelo classificador não atingiu mais que 59% de acerto considerando a opinião do especialista. Não há medida exata para determinação da quantidade de amostras necessárias para o treinamento do algoritmo, tal que, Feldman e Sanger (2007) sugerem como regra geral e empírica, por volta de 30 amostras. Faz-se necessário, em trabalhos futuros, a criação de uma base de treinamentos capaz de suportar as generalidades necessárias num processo de classificação de vários e tão diversos documentos.

O modelo classificador foi validado considerando indicadores estatísticos (Wives, 2002) e foi ajustado até que se alcançasse os seguintes resultados: a) Percentual de classificações corretas das amostras de treinamento = 100%, b) Matrix de confusão: distribuição correta entre as classes, conforme definido na base de treinamento, c) Média de Acuracidade: Precisão(Precision) = 1 e Cobertura(Recall) = 1 (ideal = 1) e d)Média de ROC area = 0,845 (ideal = 1). Nota-se que o indicador ROC area confirma a necessidade de aperfeiçoamento no classificador, o que impactaria no índice de acertos atingido.

Diferentemente da classificação por árvore de decisão, na classificação por agrupamento o número de grupos (*clusters*) não era conhecido inicialmente e sua identificação ocorreu por meio de simulações para que se visualizasse quais os clusters eram mais expressivos em termos de padrões em sua organização e conteúdo.

Tabela 6 – Simulações na formação dos clusters.

	2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters	9 clusters	10 clusters
Cluster name	Quantidade documentos	Quantidade documentos	Quantidade documentos	Quantidade documentos	Quantidade documentos	Quantidade documentos	Quantidade documentos	Quantidade documentos	Quantidade documentos
Cluster 0	6	6	6	6	6	6	6	6	6
Cluster 1	173	109	40	39	83	41	41	53	40
Cluster 2		64	110	102	26	5	5	5	5
Cluster 3			23	22	22	23	21	21	21
Cluster 4				10	9	8	8	8	8
Cluster 5					33	33	32	30	28
Cluster 6						63	57	39	50
Cluster 7							9	9	7
Cluster 8								8	8
Cluster 9									6
Total	179	179	179	179	179	179	179	179	179

Fonte: Dados da pesquisa.

Durante as simulações foram produzidos vários conjuntos de agrupamentos, variando de 2 a 10 *clusters* - a partir do 11º cluster ocorreu muita fragmentação apresentando grupos de 1 ou 2 documentos, tornando-se inviável o agrupamento. Observou-se que a 10ª simulação apresentou o melhor conjunto de *clusters* de documentos, destacando-se os *clusters* 0, 3, 4, 5, 7 e 8 por manterem-se praticamente com mesmas características desde o seu surgimento.

A análise dos *clusters* em destaque identificou similaridades no conteúdo dos documentos sugerindo os seguintes padrões: *Cluster* 0 (6 documentos) : "tratamento do Schistosoma Haematobium" ,"diagnóstico e controle da esquistossomose urinária"; *Cluster* 3 (21 documentos) : "diagnóstico da esquistossomose","tratamento da esquistossomose"; *Cluster* 4 (8 documentos) : "educação em saúde" , "programas de controle da esquistossomose"; *Cluster* 5 (28 documentos) : "estudos envolvendo roedores", "infecção do Schistosoma Mansoni em Biomphalaria Glabrata", "infecção do Schistosoma Mansoni em Biomphalaria Alexandrina" , "ações em áreas endêmicas"; *Cluster* 7 (7 documentos) : "experimentos envolvendo Biomphalaria Tenagophila"; *Cluster* 8 (8 documentos) : "testes de drogas contra o Schistosoma Mansoni" e "estudos envolvendo vermes adultos".

A classificação pelo método de Agrupamento revelou similaridades nos documentos que podem ser avaliadas como padrões úteis para o entendimento do seu conteúdo, isto é, o conhecimento descoberto. Destacou-se neste tipo de classificação a necessidade de se executar vários experimentos em busca da quantidade ideal de *clusters*, com a participação direta do especialista capaz de avaliar a qualidade dos *clusters* formados.

5 CONSIDERAÇÕES FINAIS

O estudo investigou como descobrir conhecimento sobre a esquistossome em artigos científicos utilizando-se de técnicas de mineração de textos. Levando-se em conta o propósito da mineração de textos que é a busca e extração de informação útil em coleções de textos e o que foi observado neste estudo, considera-se que o objetivo foi atingido: o conhecimento foi descoberto e demonstrado de três formas: lista de termos mais relevantes representando a coleção de documentos, documentos classificados por predição (árvore de decisão) baseado em conhecimento fornecido por especialista e documentos classificados por agrupamento sem suporte prévio de especialista – que resultou em identificação de grupos que sugerem padrões no seu conteúdo. Percebe-se que o conhecimento descoberto no estudo exibe uma nova dimensão para abordagem e entendimento do conteúdo na coleção textual sobre a esquistossomose. O estudo mostrou que há melhorias a serem feitas em trabalhos futuros,

sobretudo no classificador de documentos por árvore de decisão, de forma a elevar o índice de acertos quando comparado à opinião do especialista.

Deve-se também considerar que as técnicas de mineração de textos aplicadas aqui, são passíveis de utilização diferenciada pelo pesquisador, ou seja, as mesmas técnicas escolhidas poderiam produzir resultados diferentes utilizando-se a mesma base de documentos por outro grupo de pesquisas, e não se trata apenas de manipulação de variáveis e parâmetros dos algoritmos, mas da decisão pessoal dos pesquisadores em determinados momentos, como por exemplo: "qual a dimensão ideal de termos para representar a coleção?" ou "qual termo deve ser removido de uma lista já filtrada, mas que ainda apresenta alta dimensionalidade?" ou ainda "os padrões descobertos são significativos?". Conclui-se que os resultados da mineração de textos são dependentes de uma forte "parceria" entre as técnicas/tecnologias e o conhecimento humano.

REFERÊNCIAS

BHUMIKA, S.; SUKHJIT, S.; NAYYAR, A. A Review paper on algorithms used for Text Classification. **International Journal of Application or Innovation in Engineering & Management**, v. 2, n. 3, p. 90-99, 2013.

CORRÊA, G. C.; MARCACINI, R. M.; REZENDE, S. O. **Uso da mineração de textos na análise exploratória de artigos científicos**. São Carlos: Universidade de São Paulo, 2012. Relatórios Técnicos do ICMC.

DAVENPORT, T. H.; PRUSAK, L. **Conhecimento empresarial**: como as organizações gerenciam o seu capital intelectual. Rio de Janeiro: Campus, 1998.

EBECKEN, N; LOPES, M; COSTA, M. **Mineração de textos**. São Paulo: Manole, 2003.

FELDMAN, R; DAGAN, I. **Knowledge discovery in textual databases (KDT)**: Knowledge Discovery and Data Mining, KDD-95: [S.l.]: Association for the Advancement of Artificial Inteligence, 1995.

FELDMAN, R.; SANGER, J. **The text mining handbook**. New York: Cambridge University, 2007.

LOPES, M. C. S. **Mineração de dados textuais utilizando técnicas de clustering para o idioma português**. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2004.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, v. 2, n. 2, p. 159-165, 1958.

MANNING, C. D.; RAGAHAVAN, P.; SCHUTZE, H**. An introduction to information retrieval**. Cambridge: Cambridge University, 2008

BRASIL. Ministério da Saúde. **Portal da Saúde**. Disponível em: <http://portalsaude.saude.gov.br/index.php/o-ministerio/principal/leia-mais-o-ministerio/656-secretaria-svs/vigilancia-de-a-a-z/esquistossomose/11244-situacao-epidemiologica-dados>. Acesso em: 13 fev. 2016.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de Textos**: technical report INF_005/07. goiânia: Universidade Federal de Goiás, 2007.

NEVES, B. J; ANDRADE, C. H.; CRAVO, P. V. L. Natural Products as Leads in Schistosome Drug Discovery. **Molecules**, v. 20, n. 2, p. 1872-1903, 2015.

NONAKA, I.; TAKEUCHI, H. **Criação do conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação**. 2. ed. Rio de Janeiro: Campus, 1997.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Disponível em: <http://www.who.int/mediacentre/factsheets/fs115/en/>. Acesso em: 16 mar. 2016.

SILBERSCHATZ, A.; TUZHILIN, A. **On subjective measures of interestingness in knowledge discovery**. In: Knowldege Discovery and Data Mining, KDD-95 Proceedings, Association for the Advancement of Artificial Inteligence (AAAI), Montreal, Quebec, Canada, 1995.

TEIXEIRA, M. R. F. Gestão do Conhecimento: uma abordagem inicial. *In*: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 2000. Porto Alegre. **Anais...** Porto Alegre: Universidade Federal do Rio Grande do Sul, 2000.

WIVES, L. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2002. Exame de Qualificação EQ-069.