Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

ICICT
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde

CURSO DE ESPECIALIZAÇÃO EM INFORMAÇÃO CIENTÍFICA E TECNOLÓGICA EM SAÚDE

MINERAÇÃO DE TEXTO PARA UMA TERMINOLOGIA ESPECIALIZADA EM MODELOS MATEMÁTICOS APLICADOS A DENGUE

por

ALYNE MORAES COSTA

Instituto de Informação Científica e Tecnológica em Saúde (ICICT)

Projeto apresentado ao Instituto de Comunicação e Informação Científica e Tecnológica em Saúde da Fundação Oswaldo Cruz como requisito parcial para obtenção do título de Especialista em Informação Científica e Tecnológica em Saúde.

Orientador (es): Cícera Henrique da Silva,
Doutora em Ciências da Informação e da Comunicação.

Rosane Abdala Lins de Santana,
Mestre em Saúde Pública.

Leonardo de Souza Melo,
Mestrando em Ciências.

Rio de Janeiro, novembro/ 2011

SUMÁRIO

RESUMO

A dengue é uma doença infecciosa que constitui um dos grandes problemas de saúde pública no mundo. No Brasil ocorre a circulação simultânea de três sorotipos virais: DENV-1, DENV-2 e DENV-3, sendo o sorotipo DENV-4 recentemente encontrado em casos isolados nas regiões Norte e Sudeste. O cenário da doença no Brasil não é estimulante, pois as iniciativas de combate ao vetor, ainda se apresentam ineficazes. Com esta preocupação, a Fiocruz criou a Rede Pronex de modelagem, que funciona como um fórum para a discussão de modelos matemáticos para a aplicação no controle da dengue, de forma prevenir futuras situações de epidemias. O volume da literatura biomédica disponível na web está aumentando de tal forma, que se torna difícil localizar e recuperar a informação desejada. A recuperação da informação relevante a partir de textos completos pode ser resolvida através da utilização da mineração de texto, de forma atuar como um instrumento para a construção de vocabulários controlados e apoiar na construção de ontologias. Estudos recentes apontam que a ontologia pode ser a solução para a recuperação da informação de grandes volumes textuais. Nesta perspectiva da utilização da mineração de texto para a recuperação da informação por meio de ontologias, que se apóia o desenvolvimento deste projeto, particularmente trata-se de utilizar o texto científico como objeto capaz de apontar a terminologia da área para a recuperação de informação, uma vez que não foi possível identificar nenhuma linguagem documentária específica para modelos matemáticos aplicados ao controle da dengue.

PALAVRAS-CHAVE: mineração de textos; terminologia; dengue; modelos matemáticos.

1. INTRODUÇÃO

As doenças tropicais negligenciadas constituem um conjunto de enfermidades infecciosas ou parasitárias, prevalentes em regiões de pobreza, principalmente nos países em desenvolvimento (BRASIL, 2010a). O Brasil contribui com a maior parte da carga de doenças negligenciadas na América Latina e no Caribe (HOTEZ, 2008). Isto significa que grande parte do contingente da população mais pobre do Brasil está infectada por uma ou mais destas doenças, como por exemplo, a dengue.

A dengue é uma doença infecciosa considerada um dos grandes problemas de saúde pública no mundo (WORD HEALTH ORGANIZATION-WHO, 2009). No Brasil, ocorre a circulação simultânea de três (DENV-1, DENV-2, DENV-3) dos quatro sorotipos da dengue, sendo o tipo 4 (DENV-4) recentemente encontrado em casos isolados na região Norte (FIGUEIREDO et al., 2008) e Sudeste (FUNDAÇÃO OSWALDO CRUZ- FIOCRUZ, 2011b). Desde a década de 80, a incidência da infecção vem aumentando com uma sucessão de epidemias. E embora, exista financiamento por parte do governo brasileiro para o controle do vetor, as estratégias de combate à dengue ainda se apresentam ineficazes (TEIXEIRA et al., 2002).

Com esta preocupação, a Fiocruz criou a Rede Dengue, ligada a Vice-Presidência de Ambiente, Atenção e Promoção da Saúde. O objetivo da rede é unir todas as atividades realizadas no âmbito da Fundação voltadas para o enfrentamento da doença, visando colaborar com o Programa Nacional de Combate a Dengue e com os estados e municípios brasileiros no controle de epidemias (FIOCRUZ, 2011b). Uma das atividades ligadas à Rede Dengue é a Rede Pronex de modelagem, criada em julho de 2010, que funciona como um fórum para discussão de projetos aplicados a modelos matemáticos para a aplicação no controle da dengue (BRASIL, 2010b).

Atualmente, a construção de modelos matemáticos é uma das ferramentas utilizadas para o estudo de problemas em diversas áreas, entre estas a área de epidemiologia. Estes modelos utilizam o conhecimento com a função de prevenir futuras situações de epidemias, determinar a prevalência e a incidência de

doenças e principalmente, atuar na tomada de decisões para o controle epidemiológico (MONTESINOS-LÓPEZ; HERNÁNDEZ-SUÁREZ, 2007). No controle da dengue, modelos matemáticos são fundamentais para a otimização do uso de estratégias combinadas, minimizações do impacto ambiental e econômico, avaliação de custo-efetividade e a predição da dinâmica evolutiva do vetor em resposta às estratégias antigas e novas (CODEÇO, 2011).

O volume da literatura biomédica disponível na web está aumentando de tal forma, que se torna difícil localizar e recuperar a informação (SPASIC et al., 2005). Embora a web tenha sido projetada para possibilitar o fácil acesso, intercâmbio e a recuperação da informação, esta cresceu de forma exponencial e desordenada, o que se torna um grande problema quando necessitamos recuperar documentos de maneira eficiente (SOUZA; ALVARENGA, 2004).

A minimização da dificuldade do usuário em localizar a informação desejada, visto o excesso informacional, é uma das funções desempenhadas pelos sistemas de recuperação da informação (KOWALSKI, 1997).

Os modelos clássicos utilizados no processo da recuperação da informação utilizam estratégias de busca de documentos relevantes para a consulta, entre estes, inclui-se o modelo booleano. Este modelo recupera a informação através da aplicação da expressão booleana formada por conectivos lógicos, que também pode ser combinada com operadores de proximidade. Logo, somente os documentos cujos termos de indexação satisfazem a consulta booleana serão recuperados. Entretanto, este modelo não se aplica aos modernos sistemas de texto completo como os mecanismos de busca na web (CARDOSO, 2000).

O problema de recuperar a informação relevante a partir de textos completos pode ser resolvido, através da mineração de texto. De acordo com Araújo Júnior e Tarapanoff (2006) a mineração de texto pode ser utilizada como um instrumento para a construção de um vocabulário controlado, através de listas de palavras mais frequentes no documento, podendo enriquecer ou apoiar na construção de ontologias.

Diante da problemática apresentada sobre o excesso de informação disponibilizada em meio eletrônico sobre a temática e considerando que os operadores booleanos não dão conta de realizar a recuperação da informação de

forma precisa de textos completos, propõe-se a utilização da mineração de texto como uma ferramenta eficiente para a recuperação da informação de textos completos. Especificamente, a temática da dengue e de modelos matemáticos na literatura disponível nas bases de dados científicas será o objeto deste projeto.

O projeto de pesquisa aqui apresentado é resultante da parceria entre o Instituto de Informação e Comunicação em Saúde (ICICT) e o Centro de Métodos Quantitativos (CEMEQ-PROCC), ambos pertencentes à Fundação Oswaldo Cruz, no estado do Rio de Janeiro, Brasil. Trata-se de uma contribuição para a Rede Dengue, permitindo uma interface entre as áreas de Informação em Saúde e de Modelos Matemáticos Aplicados as Epidemias.

No próximo capítulo, será apresentada a justificativa para o desenvolvimento deste projeto.

2. JUSTIFICATIVA

A dengue representa um problema de saúde pública no mundo. Ocorrendo principalmente em áreas tropicais na Ásia, Oceania, África, Austrália e nas Américas (Figura 1) podendo atingir as áreas subtropicais e temperadas no período de verão. Nos últimos 50 anos com a crescente expansão geográfica, a incidência da dengue aumentou 30 vezes. Estima-se que anualmente ocorre cerca de 50 milhões de infecções pelo vírus da dengue, e que aproximadamente 2,5 milhões de pessoas vivem em áreas endêmicas da doença (WHO, 2009).

Figura 1- Distribuição mundial das áreas de risco da dengue (Azul)
Fonte: HEALTHMAP- CDC dengue(2011)

Atualmente, o Brasil é o país das Américas mais afetado em número de casos de dengue, sendo responsável por, aproximadamente 78% dos casos notificados (TEIXEIRA et al., 2009). O território brasileiro apresenta a circulação simultânea de três sorotipos virais: DENV-1, 2 e 3, sendo o sorotipo 4 (DENV-4) recentemente encontrado em casos isolados na região Norte (FIGUEIREDO et al., 2008) e Sudeste (FIOCRUZ, 2011b).

O monitoramento da circulação viral no território brasileiro demonstra que o sorotipo DENV-3 é responsável pelo maior número de casos da doença e o sorotipo DENV-2 está associado com a maior gravidade dos casos, sendo predominante, principalmente no estado do Rio de Janeiro (BRASIL, 2009).

A dengue é considerada uma arbovirose, devido ser transmitida ao homem através de um vetor artrópode (HENCHAL; PUTNAK, 1990). Os vetores do vírus da dengue são mosquitos domésticos, com atividade hematofágica diurna. As espécies *Aedes aegypti*, *Aedes albopictus, Aedes polynesiensis e Aedes africanus* são responsáveis pela transmissão da doença (GUBLER, 1998).

O mosquito *Aedes aegypti* é, portanto, considerado o mais importante vetor da dengue, além de ser também o transmissor da febre amarela (DOHERTY, 1993). Este mosquito é uma espécie tropical e subtropical amplamente distribuída pelo mundo, especialmente entre as latitudes 35ºN e 35ºS, e alguns fatores extrínsecos, como chuva, temperatura, altitude, topografia e umidade, condicionam sua sobrevivência. Apresentam hábitos domiciliares e peridomiciliares e alta eficiência para transmitir o vírus da dengue (PINHEIRO et al., 1997).

São descritos dois ciclos de transmissão para a doença: o ciclo endêmico/ epidêmico urbano, o qual envolvem o hospedeiro humano e o vetor *Aedes aegypti*, *Aedes albopictus* e outros mosquitos do gênero *Aedes*. E o ciclo selvagem, o qual tem sido descrito nas florestas da África e Malásia, envolvendo primatas não-humanos como hospedeiros e diferentes mosquitos do gênero *Aedes* (WANG et al., 2000).

O ciclo da dengue inicia-se através da picada de fêmeas do mosquito (*Aedes* sp.) nos hospedeiros vertebrados. Após a picada por um mosquito infectado, o vírus inicia no hospedeiro, um período de incubação que leva de 3 a 14 dias, passando a causar sintomas e sinais não-específicos, além de febre (SILER et al., 1926 *apud* BORBA, 2010, p. 31). Durante o período febril, as partículas virais circulam no sangue do hospedeiro podendo assim ser transmitidas, caso um mosquito venha a picar a pessoa infectada (GUBLER *et al.*, 1981). Logo, o mosquito que porventura se alimentou do sangue infectado, após um período de incubação de 8 a 12 dias, denominado período de incubação extrínseca passa a transmitir o

vírus, assim recomeçando o ciclo de transmissão (GUBLER, 1998; THOMAS *et al.*, 2003).

Até o momento, não existe uma vacina eficaz que assegure a proteção contra os quatro sorotipos existentes da doença (BARRETO et al., 2011). As medidas de controle da doença dependem do combate ao vetor nos grandes centros urbanos dos trópicos. Entretanto, com a rápida expansão das áreas urbanas, a tarefa tornou-se impossível se não houver a colaboração efetiva da população que vive em áreas endêmicas (GUBLER, 1989).

No Brasil, o cenário para o controle da dengue não é estimulante. As iniciativas de saúde pública brasileira visam o aumento da conscientização acerca dos sintomas, com o propósito de facilitar a chegada mais rápida dos doentes aos serviços de saúde, a fim de permitir o tratamento precoce das formas severas (BARRETO et al., 2011).

A Rede Dengue, também conhecida como Rede de Ações Integradas de Atenção a Saúde no Controle da Dengue, é uma destas iniciativas. A Rede foi criada pela Fiocruz, no ano de 2003 e é ligada a Vice-Presidência de Ambiente, Atenção e Promoção da Saúde. O objetivo da rede é unir todas as atividades realizadas no âmbito da Fundação, como as atividades de: promoção, prevenção, educação, assistência e diagnóstico, visando colaborar com o Programa Nacional de Combate a Dengue e com estados e municípios brasileiros no controle de epidemias (FIOCRUZ, 2011b).

Uma das atividades ligadas à Rede Dengue é a Rede Pronex de modelagem, criada em julho de 2010, que funciona como um fórum para discussão de projetos aplicados a modelos matemáticos para a aplicação no controle da dengue (BRASIL, 2010b). Além da Fiocruz, compõem a Rede Pronex 9 outras instituições: a Universidade Federal de Minas Gerais (UFMG), Universidade Federal Fluminense (UFF), Fundação Getúlio Vargas (FGV), Instituto Nacional de Matemática Pura e Aplicada (IMPA), Universidade Federal de Ouro Preto (UFOP), Universidade Federal de Lavras (UFLA), Universidade Estadual do Oeste do Paraná (UNIOESTE), Universidade de São Paulo (USP) e Universidade Federal do Maranhão (UFMA) (CODEÇO, 2011).

Para o Ministério da Saúde (BRASIL, 2010b) existe uma carência de modelos que possam explicar a evolução da dengue, assim como a fisiopatogenia, com evidências sólidas que possibilitem traçar perspectivas de incidência e re-infecção, de forma prevenir as formas mais graves da doença, como a dengue hemorrágica.

Os modelos matemáticos aplicados ao controle da dengue é então o principal tema de pesquisa do grupo onde se insere a presente autora, onde uma questão tornou-se também primordial para o desenvolvimento da pesquisa da área. Como recuperar informação científica relevante no ambiente eletrônico, particularmente na web, quando o fenômeno “explosão da informação”, identificado por Price em 1965 parece incontornável para o pesquisador?

Estudos recentes da área de informação e da computação apontam que a ontologia pode ser a solução (MARCONDES et al., 2008). Entretanto, na área temática deste projeto não se encontrou qualquer ontologia. Há experimentos que dão conta de ontologias para dengue, como exemplo o estudo de Rajapakse e colaboradores (2008), mas não para modelos matemáticos aplicados ao controle da dengue.

É então na perspectiva teórica de mineração de textos para recuperação de informação por meio de ontologias que se apóia o desenvolvimento deste projeto, particularmente trata-se de utilizar o texto científico como objeto capaz de apontar a terminologia da área para recuperação de informação, uma vez que não foi possível identificar nenhuma linguagem documentária específica para a temática.

3. REFERENCIAL TEÓRICO

Para o embasamento teórico deste projeto, debruçou-se sobre a literatura científica de três pilares da área de Estudos de Informação: recuperação da informação, mineração de textos e linguagens documentárias, mais especificamente a de terminologias e ontologias.

3.1 Recuperação da Informação

A recuperação da informação (RI) é a área da Ciência da Informação que estuda a estrutura, análise, organização, armazenamento, recuperação e a busca de informação (SALTON, 1968 *apud* BEPPLER, 2008).

O sistema de recuperação da informação (SRI) deve responder as demandas do usuário, necessitando que os documentos presentes na base de dados sejam submetidos a um tratamento. Este procedimento permite a extração de descritores, (palavras que identificam um determinado tema ou conceito, para fins de indexação) com o intuito de garantir um rápido acesso à informação (SOUZA; ALVARENGA NETO; MENDES, 2007).

O objetivo de um SRI é minimizar a dificuldade do usuário à informação, permitindo o fácil acesso ao que se deseja (KOWALSKI, 1997). Entretanto, os usuários ainda gastam muito tempo na busca por uma informação a qual necessita (LAI; SOH, 2004).

Os sistemas de recuperação da informação trabalham com diversos modelos, utilizando a palavra como unidade básica de acesso à informação (CORRÊA et al., 2011). Contudo, nos últimos anos estes modelos vêm evoluindo, principalmente após o surgimento da Internet (LIN; DENNER-FUSHMAN, 2006). Vários modelos foram desenvolvidos, oferecendo diferentes estruturas e linguagem de busca, cujo objetivo é facilitar o acesso à informação entre estes o modelo booleano.

O modelo booleano é baseado na teoria de conjuntos e na álgebra booleana (KORFHAGE, 1997). Neste modelo, um documento é representado por um conjunto de termos de indexação, que podem ser definidos de forma manual ou automática. As buscas são formuladas por meio de expressões booleanas e de conectores lógicos, como: *and, or e not* (e, ou, não). A recuperação de um documento somente acontece, se este responde verdadeiramente a uma expressão booleana. Os operadores booleanos também podem ser combinados com os operadores de proximidade. No entanto, mesmo que os operadores de proximidade possam agregar novos recursos ao sistema de textos completos, estes não alteram as vantagens e limitações do modelo booleano (CARDOSO, 2000).

Atualmente, a área de recuperação da informação volta-se para o desenvolvimento de sistemas inteligentes com base em processamento de linguagem natural, em função da disponibilidade de textos completos e da necessidade de interfaces voltadas para o usuário (BRÄSCHER, 2002).

3.2 Mineração de texto

A mineração de texto também conhecida como *Text Mining* ou Descoberta do conhecimento em textos caracteriza-se como um processo de descoberta, através da coleta e extração do conhecimento a partir de dados não estruturados (HEARST, 1999). Constitui uma ferramenta de ponta, quando se trabalha com grandes volumes de informação textuais com o objetivo de filtrar e resumir o conhecimento (WIVES, 1999).

O texto é o meio predominante de troca de informações entre os especialistas. No entanto, o volume da literatura biomédica está aumentando de tal forma, que se torna difícil localizar, recuperar a informação, sem a utilização da mineração de texto (SPASIC et al., 2005).

Diversas formas de análise ao processo de mineração de texto podem ser empregadas em dados textuais, sendo as principais: a estatística e a semântica. Na análise estatística, a importância dos termos está diretamente ligada a

frequência destes no texto. Já, a análise semântica além de utilizar aspectos estatísticos no tratamento de textos, sua abordagem considera a relevância da linguagem natural nos processamentos de mineração de texto (EBECKEN; LOPES; COSTA, 2005).

O emprego da técnica de processamento de linguagem natural é capaz de avaliar e identificar a funcionalidade correta de um determinado termo e a sua importância no contexto (HAHN; WERMTER, 2006).

De maneira geral, o processo de mineração de texto contém cinco etapas (Figura 2): coleta, pré-processamento, indexação, processamento ou mineração e pós-processamento ou análise da informação.

Figura 2- Etapas do processo de Mineração de texto.
Fonte: Xavier et al. (2011).

Na primeira etapa, realiza-se a coleta ou a recuperação de informação. Os dados podem ser obtidos através das bases de dados científicas ou na web. Estes dados recuperados irão compor a base de textos, também conhecida como *Corpus* (EBECKEN; LOPES; COSTA, 2005).

Na segunda etapa, denominada pré-processamento realiza-se a transformação dos documentos em formato adequado, limpando os dados indesejáveis, a fim de obter uma representação estruturada dos documentos. O pré-processamento é uma etapa onde se aplicam diversas técnicas, como: a *tokenização*, remoção de *stopwods* e *stemming*. A *tokenização* constitui a técnica de fragmentar o conteúdo textual em unidades mínimas, além de realizar a correção ortográfica do conteúdo. A remoção de *stopwords* consiste na retirada de palavras que

aparecem no texto inúmeras vezes e que não apresentam relevância semântica, como: artigos, pronomes, preposições, entre outras. O processo de *stemming* consiste em reduzir a palavra ao seu radical, de forma remover diversas variações de palavras como: plural, sufixos, dentre outras (EBECKEN; LOPES; COSTA, 2005).

A terceira etapa é a indexação, que envolve o processo de organizar todos os termos adquiridos a partir das fontes de dados, com a criação de índices que facilitam o acesso e a recuperação da informação (MONTEIRO; GOMES; OLIVEIRA, 2006).

Após a indexação, inicia-se a fase de processamento, onde se aplicam os algoritmos de mineração de texto. Esta etapa é dividida em: geração de conhecimento, que utiliza técnicas para gerar conhecimento a partir de informação contida em um determinado texto; extração de conhecimento que utiliza técnicas para extrair o conhecimento explícito no texto; e a análise de informação ou pós-processamento. Esta fase é responsável pela avaliação e interpretação dos resultados, com o objetivo de melhorar a compreensão do conhecimento descoberto pelo minerador, validando-o através de medidas de qualidade. A análise dos resultados pode ser realizada com base em técnicas bibliométricas (op. cit).

De acordo com Araújo Júnior e Tarapanoff (2006), a mineração de textos é uma ferramenta capaz de sumarizar um conjunto de documentos em agrupamentos, apresentando-os sob a forma de gráficos indicativos das relações semânticas dos termos que as compõem, o que pode ser utilizado como um instrumento para a construção de um vocabulário controlado, através de listas de palavras mais frequentes no documento, apresentando desta forma potencial para o enriquecimento ou apoio na construção de ontologias.

3.3 Terminologias e Ontologias

A área biomédica, especialmente, é caracterizada por uma vasta gama de terminologias. O estudo de terminologias permitiu um avanço importante para as

diversas áreas, tais como: a inteligência artificial, a mineração de textos, a busca e recuperação da informação (FREITAS, SCHULZ, MORAES, 2009). O termo "terminologia" possui duas definições distintas: a primeira definição refere-se ao conjunto vocabular próprio de uma área de domínio, como por exemplo: a terminologia da biomedicina, da informática, do Direito, entre outras; a segunda definição refere-se ao estudo de fenômenos linguísticos, quando diferentes termos representam o mesmo significado (ALMEIDA, CORREIA, 2008).

As terminologias são relacionadas à organização de termos, enquanto as ontologias permitem uma descrição mais precisa, baseada em lógica (FREITAS, SCHULZ, MORAES, 2009).

A definição de ontologia deriva da palavra aristotélica "categoria", que se utiliza para representar alguma coisa. De acordo com Gruber (1996) uma ontologia é uma especificação explícita de uma conceitualização.

Atualmente, as ontologias são utilizadas em diversas áreas, como a gestão do conhecimento, o processamento da linguagem natural e a recuperação da informação. A formulação de ontologias tem como objetivo suprir a necessidade de um vocabulário compartilhado, permitindo a troca de informações entre os membros de uma comunidade (SOUZA; ALVARENGA, 2004).

> Atualmente, muitas áreas de aplicação beneficiam-se das ontologias, mas o campo das ciências biológicas está ganhando cada vez mais visibilidade neste cenário, já que poucas áreas científicas contêm uma quantidade tão impressionante e rapidamente crescente de termos, conceitos e definições (FREITAS, SCHULZ, MORAES, 2009, p. 9).

Os componentes básicos de uma ontologia são divididos em: classes (organizadas em uma taxonomia), relações (tipo de interações entre os conceitos de um domínio), axiomas (usados para modelar sentenças sempre verdadeiras) e instâncias (utilizadas para representar elementos específicos, ou seja, os próprios dados) (GRUBER, 1996).

As vantagens da utilização da ontologia envolvem melhorias na recuperação da informação, além de permitir formas de representação baseadas em lógica, o que

possibilita o uso de mecanismos de inferência para a criação de novos conhecimentos a partir do existente (ALMEIDA; BAX, 2003).

É importante ressaltar que este projeto será dedicado à utilização de mineração de textos para a construção de ontologias, ou seja, à identificação de terminologia da área de modelos matemáticos aplicados ao controle da dengue. A construção da ontologia propriamente dita poderá se beneficiar dos resultados encontrados no desenvolvimento deste projeto.

4. OBJETIVOS

4.1 Objetivo geral

Identificar a terminologia da produção científica na área de modelos matemáticos aplicados ao controle da dengue.

4.2 Objetivos específicos

- Identificar a literatura científica relevante sobre a dengue;
- Identificar a literatura científica relevante sobre modelos matemáticos;
- Identificar os termos mais frequentes representativos dos conteúdos de cada área.

5. METODOLOGIA

Até meados do século XVII, segundo Ziman (1979), a comunicação científica restringia-se a cartas entre os pesquisadores e publicações esporádicas de livros e panfletos, sendo que estes depois de impresso, facilmente poderiam extraviar-se no trajeto entre a livraria e o leitor, já que não havia um centro que se responsabilizasse pela transmissão dessas publicações. Logo, muitos trabalhos deixavam de ser reconhecidos por outros cientistas.

Em vista disso, surge à iniciativa das Sociedades Reais e Academias Nacionais, para a criação de um importante veículo de disseminação entre os cientistas, a revista científica. Primeiramente, as revistas científicas tinham a função de fornecer um resumo dos problemas científicos que eram discutidos nas reuniões das Sociedades e Academias. E não demorou, para que estas se transformassem em um periódico de publicação regular, como o que conhecemos hoje.

Ainda segundo Ziman (1979), o periódico científico configurou-se como o veículo formal de comunicação tanto para a disseminação do conhecimento como para comunicação entre os pares da comunidade científica. Outros instrumentos formais de comunicação científica têm sido incorporados em seções específicas do próprio periódico ou reunidos e republicados em revistas especiais, tais como os resumos indexados e os artigos de revisão.

O artigo de revisão é um instrumento formal de comunicação científica cujo objetivo é apresentar um quadro explícito do consenso vigente, numa determinada área do conhecimento. Logo, o autor de um artigo de revisão deve ler todos os trabalhos presentes na literatura científica sobre a temática do que será abordado, de forma apresentar um breve resumo, comparando os pontos de vistas convergentes e divergentes presentes na literatura.

Nesta perspectiva, Smith (1999) *apud* Sondergaard, Andersen e Hjorland (2003) corrobora a importância do artigo de revisão apontada por Ziman (1979) e presume que os artigos de revisão chegarão a desempenhar um papel mais proeminente, servindo de guia na literatura para aqueles não suficientemente familiarizados com a área para lidar com a literatura *preprint* “crua” (não avaliada).

É neste contexto que se utilizará o artigo de revisão como objeto de análise para a identificação da terminologia utilizada nas áreas de modelos matemáticos e de dengue. Explicitado o objeto de estudo, segue-se a metodologia para atender os objetivos propostos no âmbito deste projeto, que se encontra dividida em 6 etapas consecutivas:

a) Seleção de bases de dados científicas

Partindo do pressuposto que o indicador da produção científica é o artigo científico e de que os artigos científicos de revisão são capazes de apresentar todo o panorama de pesquisa realizada sobre uma determinada área, inicialmente, será realizado um levantamento no sistema de informação Dialog para a seleção das bases de dados científicas, que mais indexam a temática da dengue e a de modelos matemáticos. A princípio, buscas exploratórias serão realizadas nas bases de dados *Web of Science, Pubmed, Biosis*, mas entende-se que é importante identificar outras bases que indexem a produção científica das duas áreas, além das já conhecidas.

b) Planejamento e construção de uma estratégia de busca

Lopes (2002) define a estratégia de busca como uma técnica ou conjunto de regras que tornam possível o encontro entre uma pergunta formulada e a informação armazenada em uma base de dados. Logo, somente será recuperado um conjunto de itens que constituem a resposta de uma determinada pergunta.

A construção de uma boa estratégia será fundamental para a recuperação da informação relevante sobre as temáticas pesquisadas.

c) Execução da busca, seleção de textos para mineração, extração e limpeza dos dados

A seleção dos itens nas bases de dados científicas será realizada, seguindo o critério definido para este projeto; serão selecionados para a extração 10 artigos completos de revisão na língua inglesa com maior índice de relevância de acordo

com a indexação em cada base científica. Os artigos científicos extraídos estarão em formato PDF (*Portable Document Format*), e deverão ser submetidos à limpeza dos dados, que consiste na retirada de imagens, gráficos dentre outros elementos não-textuais.

d) Mineração de texto

A mineração de texto será utilizada para a determinação da contagem de palavras de maior frequência e relevância na temática da dengue e de modelos matemáticos.

Após a limpeza, os textos serão submetidos a tratamento de dados, preferencialmente por software livre, que realizará a mineração de texto. Na mineração de texto serão realizados os seguintes passos: pré-processamento (remoção de *stopword, stemming*, *tokenização*, correções ortográficas), a indexação (organização de termos), e o processamento (aplicação de algoritmos de mineração).

e) Análise dos resultados

Nesta etapa serão realizadas a avaliação e interpretação dos resultados com o objetivo de melhorar a compreensão do conhecimento descoberto pelo minerador, validando-o através de medidas de qualidade.

A análise dos resultados pode ser realizada com base em técnicas bibliométricas (MONTEIRO; GOMES; OLIVEIRA, 2006), como a aplicação da Lei bibliométrica de Zipf, que relaciona a frequência da ocorrência de palavras em textos.

f) Avaliação da terminologia

Nesta etapa, se contará com os especialistas da equipe do PRONEX para a avaliação do protótipo de terminologia obtido e sua potencialidade para a construção da ontologia.

6. RESULTADOS ESPERADOS

Espera-se com a identificação das palavras de maior frequência e relevância extraída de artigos de revisão sobre a área de dengue e modelos matemáticos a obtenção de um protótipo de terminologia que possibilite a construção de ontologia na área de modelos matemáticos para o controle da dengue.

Ressalte-se que a ontologia é uma forma de representar o conhecimento de grandes volumes textuais, envolvendo melhorias na recuperação da informação, além de possibilitar a criação de novos conhecimentos a partir do existente.

8. REFERÊNCIAS CONSULTADAS

ALMEIDA, G.M.B.; CORREIA, M. Terminologia e corpus: relações, métodos e recursos. In: TAGNIN, S.E.O.; VALE, O.A. (Org.). **Avanços da Linguística de *Corpus* no Brasil**. São Paulo: Humanitas, cap. 3, p. 67-94, 2008.

ALMEIDA, M.B.; BAX, M.P. Uma visão sobre ontologias: pesquisa sobre definições,tipos, aplicações, métodos de avaliação e de construção. **Ciência da Informação**, v. 32, n. 3, p. 07-20, 2003.

ARAÚJO JUNIOR, R.H.; TARAPANOFF, K. **Precisão no processo de busca e recuperação da informação**: uso da mineração de textos. Ciência da Informação, v. 35, n. 3, p. 236-247, 2006. Disponível em: <http://repositorio.bce.unb.br/bitstream/10482/943/1/ARTIGO_PrecisaoProcessoBuscaRecuperacao.pdf>. Acesso em: 10 de setembro de 2011.

BARRETO, M.L.; TEIXEIRA, M.G.; BASTOS, F.I.; XIMENES, R.A.A.; BARATA, R. B.; RODRIGUES, L.C. Sucessos e fracassos no controle de doenças infecciosas no Brasil: o contexto social e ambiental, políticas, intervenções e necessidades de pesquisa. **The Lancet -Series Saúde no Brasil**, v. 3, 2011.

BEPPLER, F.D. **Um modelo de recuperação e busca da informação baseado em ontologia e no círculo hermenêutico.** 2008. Tese (Doutorado em Engenharia do Conhecimento). Universidade Federal de Santa Catarina, Florianópolis.

BORBA, L. Dengue: **Caracterização de marcadores moleculares de virulência utilizando a tecnologia de genomas infecciosos.** 2010. Tese (Doutorado em Biologia celular e molecular), Universidade Federal do Paraná, Curitiba.

BRASCHER, M. A ambiguidade na recuperação da informação. **DataGramaZero-Ciência da informação**, v.3, n.1, 2002.

BRASIL-Ministério da Saúde. **Dengue no Brasil**. Secretaria de Vigilância em Saúde- Informe epidemiológico 17/2009. Disponível em: <www.combatadengue.com.br/downloads/boletimEpidemiologico_n026.pdf>. Acesso em: 20 de setembro de 2011.

_____. **Doenças negligenciadas**: estratégias do Ministério da Saúde. Rev. Saúde Pública, v.44, n.1, p. 200-202, 2010a. Disponível em: <http:// www.scielo.br/pdf/rsp/v44n1/23.pdf>. Acesso em: 12 de outubro de 2011.

_____. **Rede Dengue**: inovação da abordagem e da gestão em pesquisa à saúde. Rev. Saúde Pública, v.44, n.6, p. 1159-1163, 2010b. Disponível em: <http://www.scielo.br/pdf/rsp/v44n6/IT-decit.pdf>. Acesso em: 14de outubro de 2011.

CARDOSO, O.N.P. **Recuperação de Informação**. In: SEMANA DE CIÊNCIA DA COMPUTAÇÃO DA UNIVERSIDADE FEDERAL DE LAVRAS, 2000, UFLA, (Minas Gerais).

CODEÇO, C.T. **Rede Pronex de desenvolvimento de modelos matemáticos para a aplicação no controle da dengue**. In: II OFICINA TÉCNICA DA REDE PRONEX DE MODELAGEM EM DENGUE, 2011, Instituto de Matemática Pura e Aplicada -IMPA (Rio de Janeiro).

CORRÊA, R.F.; MIRANDA, D.G.; LIMA, C.O.A.; SILVA, T.J. **Indexação e recuperação de teses e dissertações por meio de sintagmas nominais**. A.to.Z. Novas práticas em informação e conhecimento, v.1, n.1, p.11-22, 2011. Disponível em: <http://www.atoz.ufpr.br/index.php/atoz/article/view/2/21>. Acesso em: 03 de novembro de 2011.

DOHERTY, R. Australia´s contribution to tropical health: past and present.The **Med. Journal of Australia**, v.158, n.8, p.552-557, 1993.

EBECKEN, N.F.F.; LOPES, M.C.S.; COSTA, M.C.A. Mineração de textos. In: REZENDE, S.O. (coord.) **Sistemas inteligentes**: fundamentos e aplicações. São Paulo: Manole, 2005, cap. 13 , p. 337-370.

FIGUEIREDO, R. M. P.; NAVECA, F.G.; BASTOS, M. S.; MELO, M.N.; VIANA, S.S.; MOURÃO, M.P.G.; COSTA, C. A.; FARIAS, I. P. **Dengue vírus type 4, Manaus, Brasil**. Emerging Infectious Diseases, v. 14, n. 4, p. 667-669, 2008. Disponível em:< http://wwwnc.cdc.gov/eid/article/14/4/pdfs/07-1185.pdf>. Acesso em: 12 de outubro de 2011.

FREITAS, F.; SCHULZ, S. MORAES, E. **Pesquisa de terminologias e ontologias atuais em biologia e medicina**. RECIIS – Rev. Eletr. de Com. Inf. Inov. Saúde, v.3, n.1, p. 08-20, 2009. Disponível em: <http://www.reciis.cict.fio cruz.br/index.php/reciis/article/view/239/248>. Acesso em: 30 de outubro de 2011.

FUNDAÇÃO OSWALDO CRUZ. **Apresentação Rede Dengue**, 2011a. Disponível em: <http://www.fiocruz.br/rededengue/cgi/cgilua.exe/sys/start.htm?sid=52011b>. Acesso em: 17 de outubro de 2011.

_____. **Fiocruz detecta Denv-4 no RJ**. Notícias Rede Dengue, 2011b. Disponível em: <http://www.fiocruz.br/rededengue/cgi/cgilua.exe/sys/start.htm?infoid=62&sid=3>. Acesso em: 20 de outubro de 2011.

GUBLER, D.J. *Aedes aegypti* and *Aedes aegypti*-borne disease control in the 1990s: top down or bottom up. **Am. J. Trop. Med. Hyg**. , v.40, p. 571-578, 1989.

_____. Dengue and dengue hemorrhagic fever. **Clin. Microb. Rev**., v.11, n.3, p.480-496, 1998.

_____; SUHARYONO, W.; TAN, R.; ABIDIN, M.; SIE, A. Viremia in patients with naturally acquired dengue infection. Bull World Health Organ., n. 59, p.623-30, 1981.

GRUBER,T. **What is ontology?** 1996. Disponível em:<http://www.ksl.stanford.edu/kst/ what-is-an-ontology.html>. Acesso em: 20 de outubro de 2011.

HAN, U.; WERMTER, J. Levels of natural language processing for text mining. In: ANANIADOU, S.; MCNAUGHT, J. (editors). **Text Mining for Biology and Biomedicine**. Norwood: Artech House, Inc., 2006, cap. 2, p.13-41.

HEALTHMAP. **Center Disease Control- CDC dengue**. Disponível em: <http://www.healthmap.org/dengue/index.php>. Acesso em: 03 de novembro de 2011.

HEARST, M.A. **Untangling Text Data Mining**. In: 37th Annual Meeting of the Association for Computational Linguistics, 1999, University of Maryland (College Park).

HENCHAL, E.A.; PUTNAK, R. **The dengue virus**. Clin. Microb. Rev., v.67, p.376-96, 1990. Disponível em: < http://cmr.asm.org/content/3/4/376.full.pdf+html>. Acesso em: 30 de setembro de 2011.

HOTEZ, P. **The giant anteater in the room**: Brazil's neglected tropical diseases problem. Plos Neglected Tropical Diseases, v. 2, n.1 p. 01-03, 2008. Disponível em:<http://www.plosntds.org/article/info%3Adoi%2F10.1371%2Fjournal.pntd.0000177>. Acesso em: 05 de outubro de 2011.

KORFHAGE, R. R. **Information Storage and Retrieval**. New York: Wiley Computer Publishing, 1997.

KOWALSKI, G. Introduction to information retrieval systems. In: **Information Retrieval Systems**: theory and Implementation. USA: Kluwer Academic Publishers, 1997.

LAI, J.; SOH, B. **Similarity Score for Information filtering thresholds**. In: IEEE International Symposium on Communications and Information Technology (ISCIT), p.216-221, 2004.

LIN, J.; DEMMER-FUSHMAN, D. **The Role of Knowledge in Conceptual Retrieval: a study in the domain of clinical medicine**. In: 29Th International Conference on Research and Developed in Information Retrieval-SIGIR, p. 99-106, 2006.

LOPES, I.L. **Estratégia de busca na recuperação da informação**: revisão da literatura. Disponível em: <http://www.scielo.br/pdf/ci/v31n2/12909.pdf>. Acesso em: 23 de setembro.

MARCONDES, C.H.; MENDONÇA, M.A.R.; MALHEIROS, L.R.; COSTA, L.C.; SANTOS, T.C.P. Ontologias como novas bases de conhecimento científico. **Perspectivas em Ciência da Informação**, v.13, n. 3, p.20-39, 2008.

MONTEIRO, L.O.; GOMES, I.R.; OLIVEIRA, T. **Etapas do processo de mineração de textos**: uma abordagem aplicada a textos em português do Brasil. In: Anais do XXVI Congresso da SBC, Mato Grosso, Campo Grande, 2006.

MONTESINOS-LÓPEZ, O.A.; HÉRNADEZ SUÁREZ, C.M. **Modelos matemáticos para enfermedades infecciosas**. Salud Publica Mex., v. 49, p. 218-226, 2007. Disponível em: <http://www.scielosp.org/pdf/spm/v49n3/07.pdf>. Acesso em: 23 de setembro de 2011.

PINHEIRO, F.P.; CORBER, S.J. Global situation of dengue and dengue hemorrhagic fever, and its emergence in the America. **World Health Statistics Quarterly**, v. 50: 161-169,1997.

PRICE, D. J. S. Networks of scientific papers: the pattern of bibliographic references indicates the nature of the scientific research front. **Science**, v. 149, n.3683, p. 510-515, 1965.

RAJAPAKSE, M.; KANAGASABAI, R.; ANG, W.T.; VEERAMANI, A.; SCHREIBER, M.J.; BAKER, C.J.O. Ontology-centric integration and navigation of the dengue literature. Journal of Biomedical Informatics, v.41, p.806-815, 2008.

SONDERGAARD, T. F.; ANDERSEN, J.; HJORLAND, B. Documents and the communication of scientific and scholarly information: Revising and updating the UNISIST model. **Journal of Documentation**, v.59, n.3, p. 278-320, 2003.

SOUZA, R. R.; ALVARENGA, L. **A web semântica e suas contribuições para a ciência da informação**. Ciência da Informação, v. 33, n.1, p. 132-141, 2004. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/view/50/50>. Acesso em: 10 de setembro de 2011.

_____; ALVARENGA NETO, R.D.C.; MENDES, K. C.I. Mapeamento semântico através da análise de ocorrência de descritores sobre a gestão do conhecimento. **Transinformação**, v.19. n.1, p. 19-30, 2007.

SPASIC, I.; ANANIADOU, S.; MCNAUGTH, J.; KUMAR, A. Text mining and ontologies in biomedicine: making sense of raw text. **Briefings in Bioinformatics**, v. 6, n. 3, p. 239-251, 2005.

TEIXEIRA, M.G.; BARRETO, M.L.; FERREIRA, I.D.A.; VASCONCELOS, P.F.C.; CAIRNCROSS, S. **Dynamics of dengue virus circulation**: a silent epidemic in a complex urban area. Trop. Med. Int. Health, v.7, p. 757-762, 2002.

_____;COSTA, M. DA C.; BARRETO, F.; BARRETO, M.L. **Dengue**: twenty-five years since reemergence in Brazil. Cad. Saúde Pública, v.25, supl.1, 2009. Disponível em:< http://www.scielosp.org/scielo.php?script=sci_arttext&pid= S0102-311X2009001300002>. Acesso em: 16 de outubro de 2011.

THOMAS, S.J.; STRICKMAN, D.; VAUGHN, D.W. Dengue Epidemiology: virus epidemiology, ecology, and emergence. In: MARGNIOROSCH, K.; MURPHY, F.A.; SHATKIN, J. (Org.). **Advances in Virus Research**. California: Elsevier, v.61, p. 235-289, 2003.

WANG, E.; NI, H.; XU, R.; BARRETT, A.D.T.; WATOWICH, S. J.; GUBLER, D. J.; WEAVER, S. C. Evolutionary Relationships of Endemic/Epidemic and Sylvatic Dengue Viruses. **Journal of Virology**, 74: 3227-3234, 2000.

WIVES, L.K. **Estudo sobre agrupamentos de documentos textuais em processamento de informação não-estruturadas usando a técnica de**

clustering. Dissertação (Mestrado em Ciência da Computação). Universidade Rio Grande do Sul, Porto Alegre, 1999.

WORLD HEALTH ORGANIZATION-WHO. **Dengue guidelines for diagnosis, treatment, prevent and control**. 2009. Disponível em: <http://whqlibdoc.who. int/publications/2009/9789241547871 _eng.pdf>. Acesso em: 05 de outubro de 2011.

XAVIER, E.C.; BARRETO, J.R.M.; NEVES, R.S.; SANTOS, V. R.P. **Mineração de textos, suas técnicas e uma aplicação envolvendo sumarização através da ferramenta gistsumm para a produção de um site informativo.** In: Departamento de Ciência da computação, Universidade Federal da Bahia, 2011.

ZIMAN, J.M. Conhecimento Público. In: **Coleção o homem e a ciência**. Belo Horizonte: Itatiaia, v.8, 1979.

8. CRONOGRAMA

	MESES											
	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**
Seleção das bases												
Execução das buscas												
Seleção e preparação dos textos para mineração												
Pré-processamento												
Indexação e processamento												
Análise dos resultados												
Construção da terminologia												
Avaliação na equipe												
Elaboração do texto final												

9. ORÇAMENTO

Os custos previstos para o desenvolvimento da pesquisa são:

- Aquisição de notebook para o assistente de pesquisa -R$ 2.000,00
- Remuneração do assistente de pesquisa pelo período de um ano-R$ 18.000,00
- Total: R$ 20.000,00

Não estão previstos demais custos para a realização da pesquisa, uma vez que se poderá contar com os recursos já disponíveis nas instituições envolvidas como o acesso à Internet, instalações e acesso ao sistema Dialog.

10. COMITÊ DE ÉTICA

O projeto deverá ser apresentado ao Comitê de Ética para cumprir formalidades, mas não há necessidade de (Termo de consentimento livre e esclarecido- TCLE) ou qualquer outro instrumento.