RECIIS – R. Eletr. de Com. Inf. Inov. Saúde. Rio de Janeiro, v.7, n.3, Set., 2013
[www.reciis.icict.fiocruz.br] e-ISSN 1981-6278

* Artigo Original

Otimização de expressão para busca de patentes: estudo de caso sobre diagnóstico de malária

Optimizing patent searches: the case of malaria diagnosis

Leonardo Silva Leite
Graduado em Ciências Biológicas pela Universidade Federal do Rio de Janeiro, possui especialização em Informação Científica e Tecnológica em Saúde e Mestrado em Ciências, ambos pelo Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict/Fiocruz). Trabalha na área de Gestão Tecnológica da Fiocruz (Gestec).
lsleite@fiocruz.br

Cícera Henrique da Silva
Graduada em Biblioteconomia e Documentação pela UniRio, mestre em Ciência da Informação pela Universidade Federal do Rio de Janeiro, convênio com o Instituto Brasileiro de Informação Científica e Tecnológica e doutora em Ciências da Informação e da Comunicação pela Université d'Aix-Marseille III. É professora permanente do Programa de Pós-Graduação em Informação e Comunicação em Saúde – PPGICS e pesquisadora em saúde pública do Laboratório de Informação Científica e Tecnológica em Saúde do Icict/Fiocruz.
chenrique@icict.fiocruz.br

DOI: 10.3395/reciis.v7i3.599pt

Resumo

A questão norteadora do resultado de pesquisa aqui apresentado é a elaboração de uma estratégia de busca de informação em documentos de patente. Mais especificamente, o artigo apresenta uma avaliação do uso das linguagens, da truncagem e localização dos termos selecionados. A estratégia de busca foi elaborada na base de patentes *Derwent Innovations Index*, interface disponível no Portal de Periódicos da Capes, pelo cruzamento de conceitos de malária com o diagnóstico associados à Classificação Internacional de Patentes (CIP) no período entre 2005 e 2009 e o resultado foi analisado com software de mineração de texto. O diagnóstico da malária é justificado pela importância da doença em nível mundial, considerando que um diagnóstico preciso influencia no tratamento mais adequado e pode acarretar reduções nos custos das despesas de saúde. Os resultados revelaram que a busca por termos apenas no título limita o retorno dos resultados, já que a maior parte dos conceitos foi encontrada no resumo. O uso da linguagem natural e controlada na mesma estratégia de busca é fundamental para se obter retorno mais preciso da informação. A truncagem é importante na estratégia de busca; é necessário, porém, conhecimento das combinações dos termos para que não sejam recuperadas informações irrelevantes.

Palavras-chave: Inovação em saúde; Prospecção tecnológica; Recuperação da informação; Estratégia de busca; Patente.

Abstract

This study focuses on the development of a strategy for finding information in patent documents. Specifically, this paper presents an evaluation of the use of language, truncation and the location of selected terms. The search strategy was developed on the Derwent Innovations Index patent database (an interface available in the Capes Journal Portal) by cross-referencing malaria concepts with diagnoses associated with international patent classifications (IPC) from 2005 to 2009. A text-mining software program was used to analyze the results. Diagnosing malaria is important because of the worldwide burden of disease and the fact that an accurate diagnosis helps to determine the most appropriate treatment and can also lead to reductions in healthcare costs. The results showed that searching for the terms only in the title limits the search results because most of the concepts were found in the abstract. The use of natural and controlled language in the same search strategy is crucial for obtaining more accurate information. Truncation is important in the search strategy; however, the combination of terms should be known so that irrelevant information is not retrieved.

Keywords: Innovation in health; Technology prospecting; Information retrieval; Search strategy; Patent.

Introdução

O início do século XXI marca para muitas empresas, organizações e instituições um tempo de grandes incertezas, ameaças e oportunidades geradas por um ambiente cada vez mais competitivo. Como afirmado por Braga (2008), para conviver com essas incertezas, as empresas, organizações e instituições devem se organizar, usando informações que vêm do ambiente para manter ou transformar seus processos e estruturas.

Segundo Kupfer e Tigre (2004), a prospecção tecnológica pode ser definida como:

> ... um meio sistemático de mapear desenvolvimentos científicos e tecnológicos futuros capazes de influenciar de forma significativa uma indústria, a economia ou a sociedade como um todo. Diferentemente das atividades de previsão clássica, que se dedicam a antecipar um futuro suposto como único, os exercícios de prospecção são construídos a partir da premissa de que são vários os futuros possíveis. Esses são tipicamente os casos em que as ações presentes alteram o futuro, como ocorre com a inovação tecnológica. (KUPFER; TIGRE, 2004, p.1)

Dentre as atividades prospectivas divididas em famílias de técnicas, propostas por Porter et al. (2004, *apud* SANTOS et al., 2004, p.197), inclui-se o monitoramento, que provê o pano de fundo necessário no qual a prospecção se baseia e pode ser usado para buscar todas as fontes de informação e produzir um rico e variado conjunto de dados.

Em termos conceituais, Goodrich (1987, p. 6) afirma que o conceito é simples:

> Identificar, acompanhar e analisar sinais de alarme precoce no ambiente. Estes sinais são os precursores de tendências e eventos emergentes que possam ter relevância futura no desenvolvimento dos negócios da organização. Como tal, precisam ser selecionados cuidadosamente dentre a abundância de informação bruta existente e analisada quanto a sua potencial relevância, antes que se executem previsões detalhadas para caracterizar as

> tendências e os eventos emergentes, e para especular sobre suas prováveis consequências para a organização.

Portanto, é fundamental que as empresas, organizações e instituições mantenham um processo de monitoramento contínuo de seu ambiente competitivo, identificando e selecionando fontes de informações úteis e confiáveis para a tomada de decisão estratégica que aumente as possibilidades de sobrevivência e o crescimento organizacional nos mercados em que atuam ao longo do tempo.

Para realizar o monitoramento de informação, é necessário identificar qual a melhor tipologia de informação que representa o que se deseja monitorar. A tipologia da informação que mais apresenta caráter tecnológico é a patente, que é conceituada como:

> Um título de propriedade temporária sobre uma invenção ou modelo de utilidade, outorgados pelo Estado aos inventores ou autores ou outras pessoas físicas ou jurídicas detentoras de direitos sobre a criação. Em contrapartida, o inventor se obriga a revelar detalhadamente todo o conteúdo técnico da matéria protegida pela patente. (INSTITUTO..., 2011)

O documento de patente é considerado a mais importante fonte primária de informação tecnológica (FRANÇA, 2007). Estudos revelam que 70% das informações tecnológicas contidas nesses documentos não estão disponíveis em qualquer outro tipo de fonte de informação (INSTITUTO..., 2011).

Para Longa (2007), a maioria das instituições de pesquisa pública e universidades não utilizam a informação em patentes como instrumento capaz de subsidiar o desenvolvimento de pesquisas/projetos, buscar parcerias, licenciamento e transferência de tecnologia. Na verdade, o que essas instituições executam na área da busca em bases de patente – e quando assim procedem - é somente a utilização da informação para aferição da patenteabilidade de suas pesquisas.

França (2007) afirma que as patentes recém-publicadas podem atuar como indicadores do estado da arte, apresentando a informação mais recente em um dado setor da técnica, pois o pedido de patente deve necessariamente demonstrar o que preexistia e o que está sendo reivindicado como novidade. Ainda segundo o autor, a patente pode ser útil em uma negociação de transferência de tecnologia, já que permite tanto a identificação de alternativas técnicas para o atendimento das necessidades da indústria, como busca de empresas que atuam em determinado segmento tecnológico.

Outra possibilidade de utilização é a análise de um setor industrial por meio de um conjunto de patentes depositadas ao longo de um determinado período de tempo, que pode mostrar a evolução do setor e apresentar indícios para novos caminhos de desenvolvimento.

A identificação dos atores de uma dada tecnologia é outro ponto importante informado em documentos de patente. Os dados bibliográficos provêm informação, tanto sobre quem inventou, quanto sobre quem possui uma dada tecnologia. A identificação dos atores pode ser utilizada, por exemplo, para a identificação de clientes ou potenciais competidores por parte de uma empresa; para agências governamentais de controle, como elementos de interesse para verificar se há uma concentração de empresas em um dado ramo industrial; para a geração de um cadastro de inventores independentes em uma dada tecnologia.

Para Araújo (1981), o uso da patente como fonte de informação tecnológica pode ser destacado na identificação de tecnologias emergentes em um campo específico da técnica e dentro de certo período de tempo, não somente refletindo a atividade inventiva e a "produção"

de novo conhecimento técnico em um país, mas também possibilitando a identificação de atividades industriais vindouras, indicando, assim, novas tendências tecnológicas e novos desenvolvimentos, muito antes que seus efeitos sejam sentidos no mercado.

Macedo e Barbosa (2000) afirmam que uma das principais vantagens do sistema de informação patentária é a sua padronização internacional. A padronização é feita por um código numérico denominado como *International Agreed Numbers for the Identification of Data* (INID). O código identifica, na primeira página do documento, denominada "folha de rosto do documento", os campos com dados bibliográficos específicos, presentes em todos os documentos de patente, independente do idioma. Isso significa que, independente de onde a patente seja depositada, independente da língua, pelo código pré-estabelecido qualquer pessoa consegue identificar o conteúdo de cada campo na folha de rosto.

Outra fonte de informação no documento de patente é a Classificação Internacional de Patentes (CIP). A CIP, criada em 1971 pelo Acordo de Estrasburgo, teve por objetivo a necessidade de recuperação da informação em patentes. Já o guia da CIP, fornecido pela *World International Patent Office* (WORLD...) e na sua oitava edição (2006), considera a CIP como um meio internacionalmente usado para se obter uma classificação uniforme de documentos de patentes, tendo, por finalidade principal:

> Criar uma ferramenta de busca eficaz para a recuperação de documentos de patentes pelos escritórios de propriedade intelectual e demais usuários, a fim de instituir a novidade e avaliar a etapa inventiva ou não obviedade (avaliando, inclusive, o avanço técnico e os resultados úteis ou sua utilidade) das características técnicas dos pedidos de patentes. (WORLD..., 2006, p. 7)

Além disso, de acordo com o mesmo guia, a Classificação tem outras finalidades importantes, como, por exemplo, servir de:

- instrumento para disposições organizadas dos documentos de patente, a fim de facilitar o acesso às informações tecnológicas e legais contidas nos mesmos;
- base de disseminação seletiva de informações a todos os usuários das informações de patentes;
- base para investigar o estado da técnica em determinados campos da tecnologia;
- base para preparar estatísticas sobre propriedade industrial que permitam a avaliação do desenvolvimento tecnológico em diversas áreas.

Em suma, a patente, considerada a mais rica fonte de informação tecnológica, revela, de acordo com sua estrutura, a informação mais recente em um dado setor da técnica, o estado da arte. Permite tanto a identificação de técnicas alternativas para o atendimento às necessidades da indústria, como identificação de empresas que atuam em determinado segmento tecnológico, entre outras utilidades.

Coelho e Silva (2003) apontam alguns requisitos essenciais para a execução de busca de patentes: bom conhecimento do tema a ser pesquisado, das bases de dados a serem utilizadas e de propriedade industrial, além de boa definição da estratégia de busca, uso de ferramentas adequadas para tratamento dos dados e boa análise dos resultados obtidos.

Segundo Kremer (1985), é fundamental a qualidade dos recursos oferecidos pelas bases de dados e a qualidade das estratégias de busca utilizadas. É interessante enfatizar que, quanto maior for a sofisticação de recursos existentes em um sistema de recuperação de informação, maior será a importância das estratégias utilizadas (KREMER, 1985).

Lopes (2002) define estratégia de busca no âmbito da recuperação da informação como uma técnica ou conjunto de regras para tornar possível o encontro entre uma pergunta formulada e a informação armazenada em uma base de dados. Uma das etapas decisivas para o êxito de uma estratégia de busca é a de planejamento da estratégia. Para Oldroyd e Citroen (1977), o processo de planejamento estratégico da busca é realizado em três etapas decisórias: decisão da melhor base para o contexto da busca; decisão das palavras-chave; decisão da fórmula lógica da estratégia.

Kremer (1985) afirma que buscas de informação bem-sucedidas dependem, em primeiro lugar, de uma eficiente determinação das necessidades do usuário. Ainda segundo a autora, nem sempre as solicitações refletem as suas reais necessidades; muitas vezes o próprio usuário não é capaz de identificá-las com precisão.

> O problema da diferença entre demanda e necessidade também costuma ocorrer por causa de ruídos na comunicação entre usuários e os serviços de informação, causados por deficiência ou incorreções no uso da terminologia técnico-científica, ou porque as pessoas tendem a pedir somente aquilo que pensam poder obter, sem conhecer realmente todos os recursos informacionais que se encontram à disposição. (KREMER, 1985, p. 193)

Kremer (1985) defende a ideia de que o ideal, nem sempre alcançado, seria que os usuários fossem capazes de realizar suas próprias buscas de informação, pois eles, melhor do que ninguém, podem conhecer suas verdadeiras necessidades. Outra vantagem é que eles conhecem, ou deveriam conhecer, a terminologia em suas áreas de especialização melhor do que os especialistas. Para tanto, é importante que os usuários sejam treinados para o uso das fontes de informação disponíveis e dos recursos de busca existentes nos sistemas de recuperação da informação. Mesmo que os usuários não realizem as buscas, o treinamento continua pertinente para conscientização da necessidade de uma explicação mais detalhada nas solicitações de busca. Se falhar a comunicação entre os usuários e o serviço de informação, será impossível alcançar a eficiência no seu atendimento (KREMER, 1985).

Para a execução de uma estratégia de busca é necessário, pois, conhecimento e perfeito domínio de todos os recursos de busca disponíveis. Além de planejamento, é preciso fazer escolhas, como, por exemplo, a base/banco de dados mais adequado; a abrangência do assunto; tipos de documentos indexados; campos de busca disponibilizados e a linguagem empregada. Para Lopes (2002) a estratégia de busca requer um constante julgamento na sua utilização, pois o intermediário tem que tomar a decisão sobre o melhor momento de implementá-la durante o processo de busca. Essas escolhas podem determinar a eficácia da busca de informação independente do tipo de solicitação. Para atingir a eficácia, faz-se necessário ora restringir os resultados alcançados, ora ampliá-los para obter informações mais relevantes, conforme o pedido de busca demandado.

De acordo com Van der Drift (1991) e Ziman (1969), citados por Dirnberger (2011), para avaliar a qualidade da estratégia de busca na recuperação da informação, são utilizadas as medidas de precisão e a de revocação. Enquanto a revocação é a porcentagem de documentos recuperados relevantes para o número total de documentos, a precisão é definida como a proporção de documentos recuperados relevantes para o número de documentos recuperados. Ainda segundo Dirnberger (2011), de acordo com tais definições, a meta de qualquer estratégia de busca seria uma revocação elevada e uma alta precisão, isto é, a recuperação de praticamente todos os documentos relevantes. No entanto, revocação e precisão tendem a se correlacionar inversamente, ou seja, uma maior precisão leva a uma menor revocação e vice-versa, o que significa que deve haver um equilíbrio adequado entre estes polos de acordo com o tipo de pesquisa escolhida.

Outro ponto importante para a qualidade da estratégia é o vocabulário. Para Lopes (2002), a linguagem controlada caracteriza-se como a utilizada apenas nos campos de descritor, termos de indexação e identificadores. Já a linguagem natural é a encontrada no título e no resumo dos documentos referenciados. Em termos gerais, a linguagem natural é o texto livre e a linguagem controlada é um conjunto limitado de termos autorizados para uso na indexação.

Para busca de informação em bases de dados estruturadas, ao se trabalhar com linguagem natural ou linguagem controlada, os sistemas dispõem de recursos que visam otimizar o resultado da busca, como os operadores booleanos (AND, OR, NOT), que são usados para relacionar termos ou palavras em uma expressão de pesquisa. Lopes (2002) considera outras duas técnicas importantes: a truncagem de termos e a busca com operadores de proximidade.

> A truncagem de termos permite ao intermediário que operacionaliza a busca usar a raiz do termo sem especificar todas as possíveis variações desse termo (prefixos e/ou sufixos). Já a técnica de busca com operadores de proximidade ou de adjacência permite especificar, na estratégia, a posição relativa de dois ou mais termos entre eles próprios. (LOPES, 2002, p. 50)

Ainda segundo a autora, nas buscas relativas ao levantamento das últimas tecnologias em uma determinada área, ou a um novo assunto, ou a um novo produto e, ainda, na busca em documentos de patentes, o uso da estratégia em linguagem natural pode ser o melhor caminho para o encontro da informação desejada. A terminologia pode ser muito atual, as aplicações ainda não são tão significativas para serem indexadas; portanto, os novos termos não foram incorporados a nenhuma lista autorizada de linguagem controlada.

Com base nas informações apresentadas, este trabalho possui como questão norteadora identificar como se deve elaborar uma estratégia de busca em documentos de patente para monitoramento. Mais especificamente, visa verificar o papel do título, do resumo e da classificação na recuperação de informação de documentos de patente.

Como estudo de caso para a realização da estratégia de busca foi utilizado como base o diagnóstico da malária. Também conhecida como paludismo, a malária é uma doença infecciosa aguda ou crônica causada por protozoários parasitas do gênero *Plasmodium*, que são transmitidos para o ser humano através da picada da fêmea do mosquito *Anopheles.*

A malária é reconhecida como grave problema de saúde pública no mundo. Ocorre em quase 50% da população, em mais de 109 países e territórios. Sua estimativa é de 300 milhões de novos casos e 1 milhão de mortes por ano, principalmente em crianças menores de 5 anos e mulheres grávidas do continente africano (BRASIL, 2010).

Devido às necessidades de controle da malária, o diagnóstico parasitológico é de grande importância na avaliação da eficiência e eficácia dos programas ou medidas de controle, bem como no estabelecimento da terapêutica e evolução das manifestações clínicas, principalmente em áreas rurais e remotas, de alta endemicidade. Um diagnóstico preciso acarreta o tratamento mais adequado e, segundo Vitzthum et al. (2005), pode levar a reduções nos custos das despesas de saúde em todo o mundo.

Para identificar o movimento de patenteamento e consequente entrada no mercado de diagnósticos para detecção da malária, necessita-se, portanto, de um monitoramento contínuo de informação que, além de estratégico para o reposicionamento competitivo das empresas, cria barreiras para a entrada de novos concorrentes no mercado.

O presente artigo objetiva apresentar, em uma base de documentos de patente, uma avaliação comparativa das linguagens natural e controlada, o uso da truncagem e a localização

dos termos selecionados como forma de auxiliar no processo de elaboração da estratégia de busca.

Metodologia

A metodologia utilizada ao longo da pesquisa baseou-se em princípios bibliométricos para elaborar um procedimento de busca sobre reagentes e *kits* de diagnóstico para malária em documentos de patente. Trata-se, então, de estudo exploratório, que visou identificar a melhor forma possível de elaborar uma estratégia de busca.

A base escolhida para coleta dos dados foi a *Derwent Innovations Index,* disponível através do Portal de Periódicos da Capes, não só pela praticidade e facilidade de acesso mas também por ser esta considerada uma das melhores bases comerciais de patentes disponíveis no mercado.

A estratégia de busca definitiva foi desenvolvida a partir dos termos encontrados em busca preliminar, procedimento similar à técnica *snowball*, utilizada em pesquisa social para captação de indivíduos em pesquisa por amostragem (BLAIKIE, 2010; YIN, 2001). A lógica da estratégia foi definida em três partes: na primeira, utilizaram-se termos que representassem os conceitos de malária; na segunda, os termos representativos do conceito diagnóstico; na terceira e última, o código da CIP mais representativo quanto à qualidade e quantidade dos documentos recuperados na estratégia preliminar. O código G01N-33/*, que se refere à investigação ou análise dos materiais pela determinação de suas propriedades químicas ou físicas, incluindo análise física de material biológico, de material biológico líquido, do sangue, imuno-ensaio etc., foi selecionado para esta etapa.

Assim, a estratégia de busca definitiva, executada em 28 de maio de 2011, configurou-se da seguinte forma: (TS=(*malaria** OR *plasmodium** OR *falciparum* OR *vivax* OR *ovale* OR *knowlesi*)) AND (TS[1]=(*diagnostic** OR *diagnosing* OR *diagnosis*) OR IP[2]=(G01N-033/*)).

A escala temporal para recuperação da informação foi definida com base na obsolescência e vida média de uma tecnologia patenteada, que, segundo os estudos de Chen *et al.* (2009), é de cinco anos. As referências recuperadas foram baixadas da base no formato *plain text* (txt) e importadas para a ferramenta de mineração de texto denominada *VantagePoint™, versão 7.0,* para tratamento e análise. Por meio desta ferramenta, os dados foram uniformizados e classificados pelo ano de prioridade da patente, ou seja, foi criado um subgrupo com os documentos de patente no intervalo entre 2005 e 2009, que foram analisados pela leitura do título, foco tecnológico e resumo (que inclui campos como novidade, uso, vantagens e descrição detalhada) dos documentos de patente.

Os critérios para análise das referências dos documentos de patente foram categorizados inicialmente sob três aspectos. O primeiro deles é a relevância, se o documento é de utilidade para malária ou não. Os documentos categorizados como não relevantes foram descartados e as análises seguiram com base nos documentos considerados relevantes. O segundo aspecto é a especificidade, se o documento é específico ou não específico (utilidade para malária e para outras doenças). O terceiro aspecto é o escopo de proteção, se o documento se refere ao tratamento, prevenção, diagnóstico ou prognóstico da malária.

As demais análises foram realizadas com base nos documentos de patente que apresentaram alguma relação com o diagnóstico da malária, podendo estar associados a tratamento e/ou prevenção e/ou prognóstico. São elas: análise das palavras-chave utilizadas, uso da CIP, uso

[1] Código da base *Derwent Innovation Index* que representa "Topic", busca no título e resumo do documento de patente.

[2] Código da base *Derwent Innovation Index* que representa "Int. Classificação da patente", busca no título e resumo do documento de patente.

da truncagem e análise da localização dos termos utilizados na estratégia de busca, ou seja, a verificação se tais termos estão no título, no resumo e/ou no foco tecnológico[3].

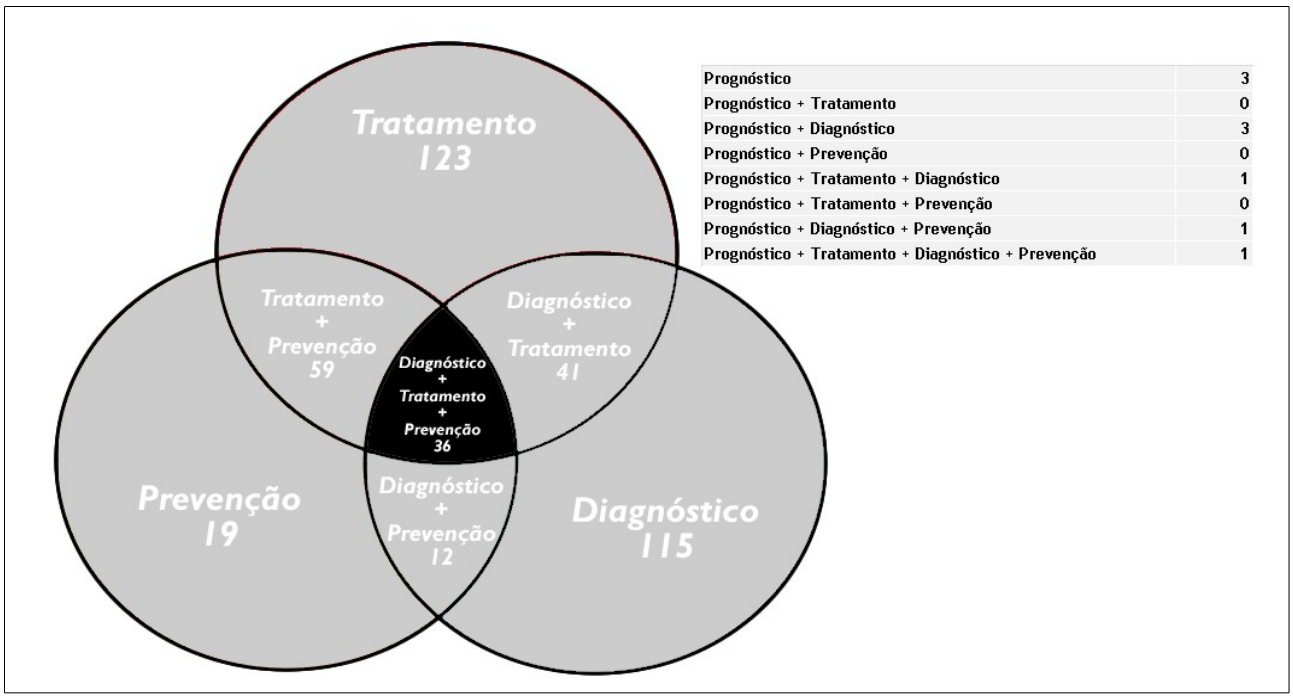
Resultados e Discussão

De acordo com a estratégia utilizada, recuperaram-se 516 referências de documentos de patente. Deste total, verificou-se que 89 (17%) documentos são exclusivos para malária; 325 (63%) documentos referem-se tanto à malária quanto a outras doenças e 102 (20%) documentos não apresentaram relevância/aplicabilidade para malária, ou seja, apesar de terem sido recuperados pela presença de uma palavra representativa do conceito malária, o escopo de proteção não se refere à doença. Estes documentos podem ser considerados ruídos da recuperação e algumas ocorrências mereceriam análise mais aprofundada, a depender de um especialista no tema em estudos futuros.

Quanto ao escopo de proteção, verificou-se a ocorrência de casos em que o documento de patente protege mais de um tipo de uso (tratamento, prevenção, diagnóstico e prognóstico) conforme pode ser observado na Figura 1. Na área de patentes é usual que o escopo de proteção seja solicitado de forma ampla e existem várias razões para tal. Dentre elas, pode-se citar a ampliação da abrangência da barreira comercial, de forma a impedir a entrada de produtos no mercado de concorrentes; a maior possibilidade de licenciamento e, assim, de retorno através do recebimento de *royalties* associados às atividades de Pesquisa & Desenvolvimento; os gastos com o registro e a colocação do produto no mercado. A ampliação do escopo de proteção é uma estratégia adotada pelos depositantes, já que, na avaliação realizada pelos examinadores dos escritórios oficiais de propriedade intelectual, estes podem restringir o escopo de acordo com a legislação local de cada país.

Figura 1 - Avaliação dos documentos de patente quanto ao escopo de proteção com base na malária.

[3] A base *Derwent* apresenta em alguns documentos, além do resumo tradicional da base, um resumo equivalente denominado foco tecnológico. O *software* de mineração de texto diferencia o campo do resumo (AB) e o campo foco tecnológico (TF). Por este motivo, foram levados em consideração os campos título (TI), resumo (AB) e foco tecnológico (TF).



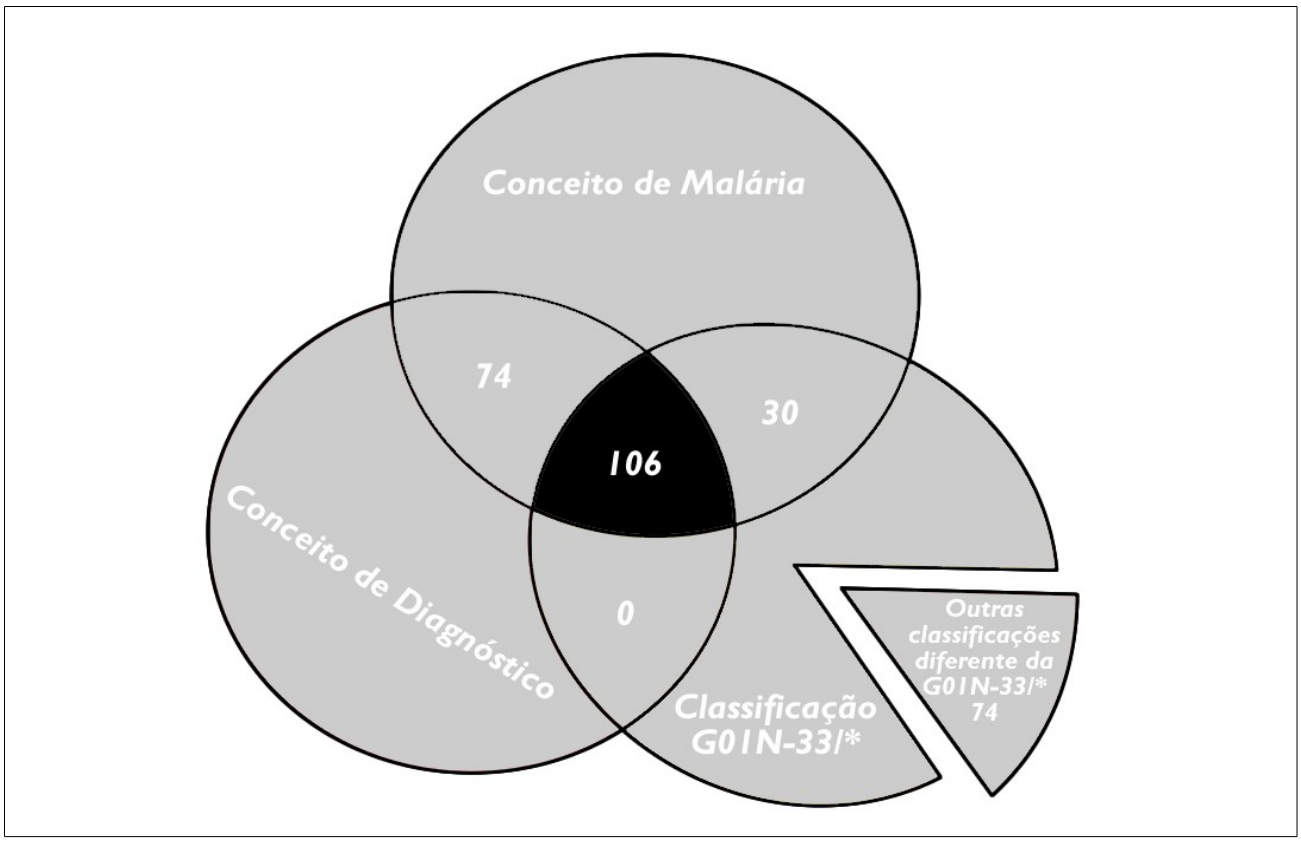
Fonte: Elaboração própria

Dos 516 documentos de patente recuperados, apenas 210 (40,7%) apresentaram em seu escopo proteção para diagnóstico da malária. Os documentos em questão foram recuperados pelos conceitos de malária e diagnóstico; porém, em muitos documentos não houve a associação diagnóstico para malária – o que foi visto foi a malária como uma possibilidade que não afetaria o diagnóstico de determinada doença.

O resultado obtido vai ao encontro do apontado por Dirnberger (2011) sobre os conceitos de revocação e precisão: uma maior precisão leva a uma menor revocação e vice-versa. A busca não foi considerada imprecisa pelo princípio observado anteriormente e pela técnica denominada *snowball* utilizada para construção da lógica estrutural da busca. Talvez o uso de outros vocabulários controlados, como o *Derwent Manual Code*, pudesse favorecer o retorno da busca.

A análise dos termos representativos do conceito malária, em associação com o conceito de diagnóstico ou com a CIP G01N-33/*, revelou que 50,48% dos documentos recuperados apresentaram a associação dos termos de diagnóstico e da CIP; 35,24% foram recuperados exclusivamente pelo uso dos termos de diagnóstico e 14,28% foram recuperados exclusivamente pelo uso da CIP (Figura 2).

Figura 2 - Linguagem Natural X Linguagem Controlada.



Fonte: Elaboração própria

Os resultados apontam que a recuperação por linguagem natural (LN) foi 75% maior em comparação com a linguagem controlada (LC), corroborando o estudo de Markey, Aherton e Newton (1980), que afirmaram que a busca com termos da LN pode, frequentemente, ser a melhor opção quando se deseja alto índice de retorno. Ainda sobre a LN, Lopes (2002) também afirma que o uso desta pode ser o melhor caminho para o encontro da informação desejada nas buscas relativas ao levantamento das últimas tecnologias em uma determinada área, a um novo assunto, ou a um novo produto e, ainda, na busca em documentos de patentes.

Em contrapartida, 14,28% dos documentos foram recuperados exclusivamente pelo uso da LC, no caso a CIP. Estes resultados também corroboram os estudos de Markey, Aherton e Newton (1980), que apontam que o melhor desempenho da estratégia de busca é aquele que utiliza os dois métodos concomitantemente.

Na análise da ocorrência das palavras utilizadas na estratégia de busca, seja pela palavra utilizada em si, seja pela truncagem dos termos, identificaram-se onze palavras associadas ao conceito malária, a saber: *malaria*; *malariae*; *malarial*; *malaria-infected*; *malaria.separating*; *malaria-causing; plasmodium; falciparum; ovale; vivax; knowlesi*, sendo as seis primeiras selecionadas em virtude do truncamento. Já com relação ao conceito diagnóstico, foram identificadas quatro palavras-chave: *diagnostic; diagnostics; diagnosing e diagnosis*. A Tabela 1, a seguir, apresenta a distribuição do número de documentos contendo as palavras recuperadas na estratégia representativas dos conceitos malária e diagnóstico.

Tabela 1 – Número de documentos recuperados x localização das palavras recuperadas.

Palavras	Título	Resumo	Foco

selecionadas	(TI)	(AB)	Tecnológico (TF)
Diagnosing	50	89	15
Diagnosis	14	65	11
Diagnostic	18	71	17
Diagnostics	0	9	0
Malaria	73	163	29
Plasmodium	40	80	33
Malariae	3	9	7
Falciparum	24	49	30
Ovale	2	10	5
Vivax	9	26	15
Knowlesi	0	0	1
Malarial	3	20	2
Malaria-infected	0	2	10
Malaria.Separating	0	0	1
Malaria-causing	0	0	2

Fonte: Elaboração própria

Cabe afirmar que, nos números encontrados, foram identificadas correlações entre as palavras pesquisadas, ou seja, a presença de duas ou mais palavras no mesmo documento (título e/ou resumo e/ou foco tecnológico), o que impulsionou a análise mais detalhada do retorno das palavras-chave pesquisadas nos 210 documentos selecionados quanto a sua localização específica.

Pode-se verificar na Tabela 2 que, das 15 palavras encontradas, 12 estão concentradas no resumo e as outras 03 no foco tecnológico, ou seja, nenhuma das palavras-chave apresentou relevância no título. É importante apontar que, neste ponto, há uma grande diferença entre artigos científicos e patentes. Nos artigos científicos, a principal representação do conteúdo do artigo está no título, que é de suma importância para recuperar a informação desejada (PESSOA JÚNIOR, 2007). Já no caso dos documentos de patente, o título apresenta-se com menor relevância em comparação com os artigos científicos, mesmo na base selecionada para este estudo, a *Derwent*, que reindexa os títulos de tais documentos. O título da patente geralmente expressa a natureza da invenção sem, no entanto, conter expressões definidoras de limitações como se verifica nos artigos.

Tabela 2 – Detalhamento da localização das palavras selecionadas.

Palavras selecionadas	TI	AB	TF	TI+AB	AB+TF	TI+TF	TI+AB+TF	TOTAL
Diagnosing	3	34	1	41	8	0	6	93
Diagnosis	3	45	2	11	9	0	0	70
Diagnostic	1	44	1	11	10	0	6	73
Diagnostics	0	9	0	0	0	0	0	9
Malaria	1	84	5	55	7	0	17	169
Plasmodium	2	36	14	25	6	0	13	96

Malariae	0	5	3	0	1	0	3	12
Falciparum	1	17	13	16	10	0	7	64
Ovale	0	6	1	0	2	0	2	11
Vivax	0	11	5	5	6	0	4	31
Knowlesi	0	0	1	0	0	0	0	1
Malarial	0	16	1	3	1	0	0	21
Malaria-infected	0	2	1	0	0	0	0	3
Malaria.Separating	0	0	1	0	0	0	0	1
Malaria-causing	0	0	1	0	0	0	0	1
	11	309	50	167	60	0	58	

Fonte: Elaboração própria

A recuperação da informação nas bases de dados de patentes pode ser realizada por vários campos, dentre eles o(s) inventor(es); depositante(s); e CIP, geralmente associada ao título e/ou resumo. A informação contida no resumo do documento de patente revela a natureza e o(s) objetivo(s) reivindicado(s). Geralmente o resumo é harmonioso com a invenção reivindicada e com todo e qualquer objeto mencionado na reivindicação.

Ao analisar a relevância do truncamento das palavras-chave (ver Quadro 1), foram identificadas 06 palavras referentes ao conceito de malária. Destas, 04 não foram relevantes para a busca por apresentar em sua totalidade (100%) a presença de outras palavras no mesmo documento. Assim, conclui-se que o documento seria recuperado utilizando a mesma estratégia com a exclusão destas palavras, a saber: *malaria-causing*, *malaria.separating*, *malaria-infected* e *malariae*. Acredita-se que três das palavras em questão somente estavam presentes pela estratégia de busca por um erro de indexação da base ao utilizar sinais de pontuação de forma indevida - no caso, usados para ligar os elementos de palavras compostas - e sinais usados para indicar o final de um período. Já as palavras *diagnostics* e *malarial* foram consideradas importantes para a estratégia de busca: 44,44% e 28,57% dos documentos recuperados apresentaram as palavras independentemente da localização - no título, no resumo e/ou no foco tecnológico.

Quadro 2 - Ocorrência de palavras recuperadas a partir da truncagem

Palavras truncadas	Outras palavras encontradas nos mesmos documentos
Diagnostics	*Diagnosing* (TI, AB, TF); *diagnosis* (AB) e *diagnostic* (AB).
Malaria-causing	*Malaria* (TI, AB, TF) e *plasmodium* (TF).
Malaria.Separating	*Malaria* (TF).
Malaria-infected	*Malaria* (AB, TF); *plasmodium* (TI, AB, TF); *malariae* (AB, TF); *falciparum* (AB, TF); *ovale* (AB, TF); *vivax* (AB, TF) e *malarial* (AB).
Malarial	*Malaria* (TI, AB, TF); *plasmodium* (TI, AB, TF); *malariae* (AB, TF), *falciparum* (TI, AB, TF); *ovale* (AB); e *vivax* (AB, TF).
Malariae	*Malaria* (TI, AB, TF); *plasmodium* (TI, AB, TF); *knowlesi* (TF); *ovale* (TI, AB, TF); *vivax* (TI, AB, TF); *malarial* (TI, AB) e *malaria-infected* (TF).

Fonte: Elaboração própria

A truncagem permite a utilização da raiz da palavra para recuperar todas as possibilidades de expansão da mesma. Para a formulação da estratégia de busca, fez-se necessário utilizar este recurso para avaliar a relevância de cada palavra truncada; porém, caso não seja esta a finalidade, é fundamental o conhecimento das combinações dos termos para que não sejam recuperadas informações irrelevantes, denominadas ruídos de informação.

Conclusão

Com base nas informações apresentadas, verificou-se que a qualidade da estratégia e a linguagem (vocabulário) são fatores importantes para a recuperação da informação. Este estudo, realizado em base de patentes, revelou que o melhor desempenho da estratégia de busca é aquele que utiliza as duas linguagens concomitantemente, ou seja, a linguagem natural associada à linguagem controlada, esta representada pela CIP.

Ainda em termos de uso da linguagem, verificou-se que, para busca em linguagem natural, a truncagem pode ser útil para reduzir o número absoluto de palavras em uma estratégia de busca e para verificar quais são as variações quanto aos prefixos truncados. Porém, caso não seja esta a finalidade, é necessário conhecimento das combinações dos termos para que não sejam recuperadas informações irrelevantes (ruídos de informação).

Quanto à localização dos termos utilizados na estratégia de busca, verificou-se que, para busca em documentos de patente, o resumo é mais relevante do que o título. Esse fato não coincide com o que ocorre com os artigos científicos, nos quais a principal chave de busca é normalmente o título.

Não há estratégia de busca que apresente 100% de relevância quanto aos documentos recuperados. A meta de uma estratégia de busca eficaz seria uma revocação elevada e uma alta precisão, isto é, a recuperação de um alto percentual de documentos relevantes. No entanto, revocação e precisão tendem a se correlacionar inversamente, ou seja, uma maior precisão leva a uma menor revocação e vice-versa, o que significa que o buscador de patentes tem de encontrar o equilíbrio adequado entre estes polos de acordo com o tipo de busca escolhida.

De certa forma, esta conclusão vem ao encontro da questão colocada por alguns autores citados de que a lógica da busca para monitoramento tecnológico é de assumir uma busca "imperfeita", pois se deve trabalhar com o risco de se recuperar documentos na frequência baixa, que pode concentrar tanto documentos relacionados à inovação como aqueles que são considerados ruídos estatísticos, por não apresentarem qualquer relação com o que se procura (QUONIAM, 1995; ROSTAING,1996 *apud* SILVA, 2002).

Acredita-se que o presente trabalho seja de grande contribuição para a recuperação da informação na área de patentes, uma vez que são raros os estudos que teorizem a área no aspecto abordado, particularmente na área da saúde. Além disso, considera-se que a metodologia, os passos para a formulação e a execução da estratégia de busca, que, no caso deste trabalho, foi aplicada à malária, podem ser utilizados para outras doenças, configurando-se como ferramenta importante no trabalho de um buscador que trabalhe na área da saúde.

Referências

BLAIKIE, Norman. **Designing social research**. Cambridge: Polity Press, 2010.

BRAGA, Fabiane dos Reis. **Um modelo de monitoramento ambiental (*environmental scanning*) orientado para o planejamento estratégico da CNEN**. 2008. Dissertação (Mestrado) -IBICT/UFRJ/ECO.

BRASIL. Ministério da Saúde. Portal da Saúde. **Malária**. Disponível em: <http://portal.saude.gov.br/portal/saude/profissional/visualizar_texto.cfm?idtxt=31086&janela=2>. Acesso em 01 setembro 2010

CHEN, D. Z. et al. Constructing a new patent bibliometric performance measure by using modified citation rate analyses with dynamic backward citation Windows. **Scientometrics** v 82, p.149–163, 2010.

COELHO, G. M.; SILVA, C. H. da. Prospecção tecnológica em patentes no setor de óleo e gás. In: Workshop Brasileiro de Inteligência Competitiva e Gestão do Conhecimento, 3., Salvador, 2003. Disponível em: < http://biblio.int.gov.br/phl82/INT_DOCELE/Memoria/1521.pdf>. Acesso em 20 jul. 2011>.

DIRNBERGER, D. A guide to efficient keyword, sequence and classification search strategies for biopharmaceutical drug-centric patent landscape searches - A human recombinant insulin patent landscape case study. **World Patent Information**, v. 33, p. 128-143, 2011.

FRANÇA, R. O. A patente. In: CAMPELLO, B. S.; CENDÓN, B. V.; KREMER, J. M. **Fontes de Informação para Pesquisadores e Profissionais**. 2. ed. Belo Horizonte: Editora UFMG, 2007. cap. 12, p. 153-182.

GONZALEZ, M. O. A.; TOLEDO, J. C.. A integração do cliente no processo de desenvolvimento de produto: revisão bibliográfica sistemática e temas para pesquisa. **Prod.**, v.22, n.1, p. 14-26, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-65132012000100002&lng=en&nrm=iso>. Acesso em 10 dez. 2012.

GOODRICH, R. S. Monitoração do ambiente externo: uma necessidade para as organizações tecnológicas. **Revista de Administração Empresarial**, Rio de Janeiro, v. 27, n. 1, 5-10, janeiro/março 1987.

Instituto Nacional da Propriedade Industrial - INPI. Apresentação. Disponível em: <http://www.inpi.gov.br/menu-esquerdo/informacao>. Acesso em: 09 jul. 2011.

KIRKBRIDE, P. Full text, free text and controlled vocabulary strategic search planning. In: ONLINE/CD-ROM 91. Chicago: Online, 1992. p. 73-78.

KREMER, J. M. Estratégia de Busca. **R. Esc. Bibliotecon. UFMG**, Belo Horizonte, v. 14, n. 2, p. 187-220, set. 1985.

KUPFER, D.; TIGRE P. B. **Modelo SENAI de prospecção**: documento metodológico. Montevideo: OIT/Cinterfor, 2004. Capítulo 2: Prospecção Tecnológica. (Papeles de la oficina técnica, 14).

LONGA, L. C. D. **O Gerenciamento da informação tecnológica contida na literatura patentária**: uma proposta para a FIOCRUZ. 2007. Dissertação (Mestrado) – Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz, Rio de Janeiro, 2007.

LOPES, I. L. Estratégia de busca na recuperação da informação: revisão da literatura. **Ci. Inf.**, Brasília, v. 31, n. 2, p. 60-71, maio/ago. 2002

LOPES, I. L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ci. Inf.**, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002.

MACEDO, M. F. G.; Barbosa, A. L. F. **Patentes, Pesquisa & Desenvolvimento**: um manual de propriedade intelectual. Rio de Janeiro: Fiocruz, 2000.

MARKEY, K.; ATHERTON, P. **Online training and practice manual for ERIC database searchers**. Syracuse: Syracure University, 1978.

OLDROYD, B. K; CITROEN, C. L. Study of strategies used in online searching. **Online Review**, v. 1, n. 4, p. 295-310, 1977.

PESSOA JÚNIOR, A. **Preparo de artigos científicos**. São Paulo: USP, 2007. 92 slides, color.

PORTER, A. L. et al. Technology Futures analysis: toward integration of the field & new method. **Technol Forecast Soc Change**. v. 71, n. 3, p. 287-303, 2004.

SANTOS et al. Prospecção de tecnologias de futuro: métodos, técnicas e abordagens. **Parcerias estratégicas**, n.19, Dez. 2004.

SILVA, C. H. **Services d´information dans un monde globalisé**: tendances et stratégies**.** 2002. Tese (Doutorado) - Universite d'Aix-Marseille III, Marseille, 2002.

VITZTHUM et al. Proteomics: from basic research to diagnostic application. A Review of requirements & needs**. Journal of Proteome Research,** v.4, p. 1086-97, 2005.

World Intellectual Property Organization- WIPO. **Preface to the International Patent Classification**. 8t. ed. Geneva, 2006. v. 5.

YIN, R. K. **Estudo de caso**: planejamento e métodos. 2. ed. Porto Alegre: Bookman, 2001.

Recebido 19-06-2012

Aceito 06-08-2013