

A IDENTIFICAÇÃO DE NOVAS DESCOBERTAS CIENTÍFICAS ATRAVÉS DA ANÁLISE DO CONHECIMENTO CONTIDO EM ARTIGOS CIENTÍFICOS.

Luciana Reis Malheiros, Depto. de Fisiologia e Farmacologia, Universidade Federal Fluminense, Doutoranda do Programa de Pós-graduação em Ciência da Informação UFF/IBICT, malheiro@vm.uff.br

Carlos Henrique Marcondes, Depto. de Ciência da Informação, Universidade Federal Fluminense, Professor do Programa de Pós-graduação em Ciência da Informação UFF/IBICT, marcon@vm.uff.br

Resumo

A publicação de artigos científicos no formato digital é cada vez mais comum, mas eles ainda mantêm o modelo das publicações em papel, não explorando todo o potencial permitido pelo meio eletrônico. A análise da literatura científica através de dados bibliométricos e cienciométricos depende de um longo processo que envolve a leitura, crítica e citação dos artigos publicados pela sociedade científica. Partindo do pressuposto que o conhecimento científico se dá sob a forma de relações entre fenômenos esta pesquisa desenvolveu um modelo para representar o conhecimento contido no texto de artigos científicos, identificando e marcando no texto relações entre fenômenos. A presente pesquisa parte vislumbra a possibilidade da utilização desse modelo para identificar o surgimento de novidades em Ciência, através da comparação do “conhecimento” identificado e marcado no texto de artigos científicos publicados na Web com o conhecimento registrado nas ontologias públicas também disponíveis na Web.

Palavras-chaves: comunicação científica, metodologia científica, publicações eletrônicas, representação do conhecimento, novidade em ciência.

Abstract

Although electronic publishing is a very usual activity to scientists and researches, it is still based in the paper print model and do not take advantages of potentialities of the Web environment. Scientific literature analysis through bibliometric and scietometric techniques largely depends on a slow process of reading, inquiring and citing articles by the scientific community. Based on the assumption that scientific knowledge has the form of relations between phenomena, o knowledge representation model was developed to identify and markup relations found in a scientific article text. This research uses that model with the aim of identifying the raise of new knowledge in Science through the comparison of the knowledge thus identified in the text of scientific articles with those recorded in public Web ontologies

Keyword: scientific communication, scientific methodology, electronic publishing, knowledge representation, new knowledge

1. INTRODUÇÃO

A Ciência da Informação tem como um de seus objetivos a avaliação da ciência identificando as áreas de pesquisa mais relevantes em um determinado período de tempo, tentando fazer previsões das áreas que irão desenvolver-se mais e das que ainda precisariam de mais incentivos e, também, medindo a produção científica, através dos trabalhos publicados/comunicação científica, dados esses que podem ser utilizados como parâmetros indicadores dos rumos e tendências do desenvolvimento da Ciência, para a obtenção de recursos e orientações de políticas na área de Ciência e Tecnologia (C&T).

A análise da literatura científica através de dados bibliométricos e cienciométricos depende, contudo, de um longo processo que envolve a leitura, crítica e citação dos artigos publicados pela sociedade científica. Assim, com o tempo, um artigo que traga um “conhecimento novo” para uma determinada área deverá ser citado mais vezes pela sua comunidade científica. Entretanto, um maior número de citações não significa necessariamente que o artigo seja portador de um “conhecimento novo”, já que citações podem ser feitas contestar um artigo. Uma outra situação em que o grande número de citações pode não ser uma medida precisa da importância de um artigo se dá quando o “conhecimento novo” que um artigo trás foge da interpretação corrente dos fenômenos daquela área. Neste caso a não-citação poderia ser um indício de novidade e a importância do trabalho, provavelmente, passaria despercebida pelos métodos bibliométricos e cienciométricos normalmente utilizados.

A comunicação científica, através de artigos de periódicos permanece, basicamente, a mesma desde sua origem há mais de 300 anos atrás. A revista científica foi criada para atender a necessidade de se trocar resultados de pesquisas de forma muito mais abrangente que a proporcionada pelas cartas trocada entre cientistas. A ciência desenvolvia-se de uma atividade particular para uma atividade social e o principal papel da comunicação científica era, e ainda é, promover o intercâmbio de informações entre os membros da comunidade científica. (LE COADIC, 1996). Para que se tornem contribuições científicas relevantes as novas descobertas advindas da pesquisa científica devem ser comunicadas dentro de uma estrutura formal rígida, mas que permita a sua compreensão, replicação e eventual contestação por parte da comunidade científica. A respeito da revista científica Ziman (1979, p.117) diz: “Contudo, a invenção de um mecanismo por meio do qual os resultados de minuciosas pesquisas podem ser publicados parceladamente talvez tenha sido o passo decisivo para o desenvolvimento do ‘Método Científico’”.

Novas descobertas na Ciência são validades cotejando-as com o conhecimento estabelecido num determinado domínio. Antes do surgimento da Web o que constituía este “conhecimento estabelecido” ou aceito tinha contornos difusos, ressentia-se de uma maior formalização que pudesse, por exemplo, classificá-lo em aceito ou válido cientificamente ou não aceito e inválido cientificamente. Este conhecimento encontrava-se armazenado em coleções de periódicos nas bibliotecas. O principal mecanismo de validação da Ciência é a leitura, interpretação, questionamento e crítica de artigos científicos pela comunidade de pesquisadores, processo que tem como manifestação concreta e medida a citação, os índices de citação, por exemplo, o fator de impacto do ISI. Artigos científicos são citados ou não, até que, num lento processo social, o novo conhecimento científico é pouco a pouco, de uma maneira difusa e pouco formal, incorporado ao que Ziman (1979) chama de “Corpo da Ciência”.

Meadows (1999, p. 127), comentando os impactos da tecnologia da informação sobre o artigo científico, assim se expressa: “[...] talvez seja possível no futuro, incluir software apropriado que permita fazer deduções a partir dos dados e informações, acrescentando assim a estes uma função de conhecimento”. Mas este cenário está rapidamente mudando. No contexto da Web, em especial no contexto da Web Semântica, o fazer Ciência vem apresentando mudanças muito rápidas e profundas sobre práticas científicas como publicar, realizar experimentos, validar novos conhecimentos e trabalhar cooperativamente. Para Berners-Lee et al (2001, p.2) “The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation” .

Hoje, por um lado, cada vez mais o registro e formalização do conhecimento em geral, e do conhecimento científico em especial se dá sob a forma das chamadas ontologias, bases de conhecimento geralmente públicas, disponíveis na Web. Ontologias são especificações formais (possivelmente, em linguagem legível por programas) de uma conceituação comum, compartilhada por uma comunidade, sobre um domínio da realidade, um domínio científico, por exemplo. As ontologias vêm se tornando as bases de conhecimento da Ciência atual: “[...] a knowledge base will acquire a value per se, only to the extent that the knowledge it contains is in fact true, such as to correspond to the world beyond the knowledge base”. (Guarino, 1995). Vários setores da comunidade científica vêm desenvolvendo ontologias específicas como a UMLS e a GeneOntology.

Por outro lado, o mecanismo de publicação da comunidade científica torna-se a publicação digital. Ferramentas de autopublicação como os ambiente Eprints e DSpace tornam-se comuns entre a comunidade científica. No entanto, as publicações eletrônicas, apesar de todos os avanços das Tecnologias de Informação e Comunicação (TICs) são ainda calcadas no modelo impresso. O documento digital tem a possibilidade de se constituir numa nova e poderosa ferramenta cognitiva, em especial no contexto da Web Semântica: “The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users” (BERNERS-LEE et al, 2001, p.2).

Será que o conhecimento contido no texto de artigos científicos publicados na Web não poderia ser extraído e representado também em meio legível por programas? Que possibilidades isso abriria para a Ciência, em termos de comunicação científica, identificação de novidades e validação de novos conhecimentos?

Uma pesquisa vem sendo desenvolvida (MARCONDES, 2006) com o objetivo de criar um programa editor/publicador de artigos científicos que permita produzir simultaneamente textos em formato legível por pessoas e por programas. Partindo do pressuposto que o conhecimento científico se dá sob a forma de relações entre fenômenos - por exemplo: “Papilomavirus humano (HPV) **causa** câncer de colo do útero” - esta pesquisa já desenvolveu um modelo para representar o conhecimento contido no texto de artigos científicos, identificando e marcando no texto relações entre fenômenos.

A presente pesquisa parte do modelo proposto acima e vislumbra a possibilidade de aplicá-lo para identificar o surgimento de novidades em Ciência, através da comparação do “conhecimento” identificado e marcado no texto de artigos científicos publicados na Web com o conhecimento registrado nas ontologias públicas também disponíveis na Web. Isso possibilitaria o processamento por um programa agente de “software” para recuperar o conhecimento aí contido de uma forma semanticamente

muito mais rica, cotejar esse conhecimento contra o conhecimento já estabelecido registrado nas ontologias, verificando sua consistência e validade.

A seguir é apresentado um quadro teórico sobre a questão da identificação de novidades em ciência; na seção seguinte é apresentado o modelo já desenvolvido para a identificação e marcação dos conhecimentos contidos no trecho de artigos científicos e discutidas suas possibilidades para a identificação de novidades em Ciência; por fim, são apresentadas conclusões e futuras direções que esta pesquisa poderá tomar.

2. A QUESTÃO DA NOVIDADE CIENTÍFICA

Se por um lado a estrutura do artigo científico pouco mudou, por outro, o meio social em que o cientista está inserido mudou drasticamente nos últimos anos. Ao final da II Guerra Mundial, o aumento do número de pesquisadores e de sub-áreas científicas fez com que o sistema de avaliação por pares ou *peer review* para a publicação de trabalhos fosse amplamente adotado. Isso significa que todo o trabalho científico apresentado para publicação deve ser submetido à avaliação de um editor e de revisores, que nada mais são do que membros da mesma comunidade científica a qual o cientista pertence. A publicação de um artigo científico numa revista conceituada trás consigo não só a opinião de seus autores, mas o “aval científico” de seus revisores e editores, conferindo-lhe um status científico diferenciado de um artigo publicado numa revista menos conceituada. (ZIMAN, 1979).

As críticas ao método de revisão por pares não são poucas já que os revisores detêm um poder considerável, pois devem avaliar se um trabalho é relevante, se seus métodos e análises de resultados foram adequados, enfim, se o trabalho merece crédito e deve ser publicado. Em áreas onde uma mudança de paradigma está em andamento ou onde uma disciplina nova começa a se estabelecer, o papel de revisores e editores pode favorecer ou abafar possíveis discussões.

Informar-se sobre os artigos científicos de sua área é tarefa fundamental no exercício da prática científica e profissional, entretanto, mesmo com o advento dos computadores e mecanismos de busca rápidos o que conseguimos, normalmente, obter é muito mais “informação” ou “dados” do que “conhecimento”. (ZIMAN, 1979). A recuperação de um volume muito grande de informações nem sempre importantes, sem dúvida, não é desejável. Alper et al (2004) avaliando quanto tempo, hoje, um clínico geral levaria para ler toda a literatura da área publicada no período de um mês, chegaram a conclusão que seriam necessárias 627,5 horas (aproximadamente 26 dias ininterruptos) para que todos os artigos fossem avaliados. Os autores concluem que para fornecer ao médico as melhores evidências do momento são necessárias melhores estratégias de seleção de artigos científicos.

A Ciência da Informação sempre se propôs a mediar a Ciência através de indicadores como a produção científica. Técnicas bibliográficas sofisticadas, baseadas na análise de citações, geram indicadores como o fator de impacto do ISI – Institute for Scientific Information -, largamente usados para avaliar a qualidade de periódicos científicos, a produtividade de autores e outros indicadores da atividade científica (CHRISTOVÃO, 1979). Hoje, mais ainda, estes indicadores bibliométricos, associados às tecnologias de visualização gráfica como no projeto Atlas de la Ciencia, <http://www.atlasofscience.net>, permitem identificar claramente tendências de desenvolvimento da Ciência.

Método Científico e conhecimento científico

O método científico estabelece um conjunto geral de técnicas e métodos sobre os quais a pesquisa científica deve ser realizada e possui como marcos importantes a publicação do *Novum Organum* de Francis Bacon e o Discurso do Método de René Descartes. Essas obras estabeleceram as bases da experimentação científica e do método indutivo onde o resultado de experiências controladas levaria a generalizações de uma lei científica.

A literatura atual sobre metodologia científica, principalmente nas chamadas ciências médicas ou da saúde, é fundamentada numa crítica ao método indutivo, o método hipotético-dedutivo, proposto por Karl Popper. Para ele, as teorias não duram para sempre e conforme um determinado campo científico evolui é natural que uma nova teoria se estabeleça no lugar da antiga. Ele considera racional na ciência aceitar condições para que uma teoria seja testada e refutada.

Marconi e Lakatos (2004) expressam o método hipotético-dedutivo, segundo Popper, da seguinte maneira: a) problemas: decorrentes, normalmente, de conflitos entre a teoria existente e as expectativas; b) conjectura: solução proposta em forma de proposição passível de teste (Se....então) e c) testes de falseamento: consistem em torna falsas as conseqüências deduzidas, por exemplo, se “x” é deduzível de “y”, mas “x” é falso, logicamente, “y” também é falso.

Ainda sobre Popper, para que uma nova teoria surja no lugar da antiga é necessário que haja uma grande atividade de crítica mútua entre os cientistas de determinada área. A nova teoria só emergiria de um confronto entre diversas teorias, como no processo de seleção natural proposto por Darwin, daí a necessidade de um meio conflitante para o bom desenvolvimento da ciência.

Novidade científica e mudança de paradigma

Kuhn, entretanto, apresentou em 1962 no livro “A Estrutura das Revoluções Científicas” uma crítica da “visão evolucionista” das teorias científicas defendida por Popper. Para Kuhn, os testes realizados com o objetivo de falsificar uma teoria não são o motor do desenvolvimento científico, até por que as revoluções científicas são raras. (KUHN, 2005)

Para ele uma disciplina se torna ciência quando entra numa fase onde os problemas com os quais ela lida são unificados e estudados com base em padrões standardizados de abordagem, não quando se atem a procedimentos de falsificação ou verificação de suas teorias. Dessa forma, o estabelecimento da ciência dá-se através do estabelecimento de um consenso sobre as múltiplas formas de se abordar um problema. Nesse momento, ele apresenta o conceito de paradigma: “O paradigma é aquilo que os membros de uma comunidade científica partilham e, inversamente, uma comunidade científica consiste em homens que partilham um paradigma.” (KUHN, 2005, p.221). Ele diz ainda, “No seu uso estabelecido, um paradigma é um modelo ou padrão aceito.” (KUHN, 2005, p.43).

O estabelecimento de um paradigma dá, na visão de Kuhn, uma moldura para que os cientistas trabalhem com o que ele chama de ciência normal. A ciência normal não busca a novidade, mas, sim corroborar o paradigma já estabelecido e é exercida por uma comunidade que volta seus esforços para “encaixar” um número cada vez maior de fenômenos dentro da moldura apresentada pelo paradigma.

Contudo, por mais que a ciência normal represente a prática comum da ciência, chega um momento em que mudanças ocorrem. Em princípio, os fenômenos não contemplados pelo paradigma passam despercebidos ou são tidos como falhas, ou

anomalias de menor importância pelos pesquisadores. Porém, chega um momento em que as anomalias atingem a estrutura básica do paradigma e não cedem as tentativas de eliminá-las. Começa, aí, uma revolução científica ou uma revolução paradigmática e interrompe-se, o modo cumulativo de produção de conhecimento. (OLIVA, 1998). Ao final do período de transição e estabelecido um novo paradigma, os cientistas terão modificado radicalmente a visão que tinham de sua área de atuação e teremos uma nova fase de ciência normal regida por outro paradigma.

Uma pergunta que nos fazemos: Como a questão da mudança de paradigma se reflete na literatura? Charles Pierce (apud Santaella, 2004, p.102) definiu o conceito de abdução da seguinte maneira: “Abdução é o processo de formar uma hipótese explicativa. É a única operação lógica que introduz qualquer idéia nova”. Santanella (2004, p103) explica “[...] os argumentos abduativos formulam sinteticamente explicações tentativas para todas as situações na ciência ou fora dela, sem eles, permaneceriam como fatos inexplicáveis”. Talvez a mudança de paradigma se reflita na literatura pela publicação de artigos “abduativos”.

Conhecimento como relação

Em que consiste o conhecimento contido no texto de artigos científico? Essa é uma discussão que vem assumindo uma centralidade cada vez maior na Ciência da Informação, tanto quanto em outras áreas do conhecimento. Quais as possibilidades e quais as formas de termos acesso a ele? Existe um conhecimento público, intersubjetivo, ou, ao contrário, o conhecimento é um processo individual? Estas são perguntas que há muito tempo a Filosofia vem se fazendo. A trajetória histórica da Filosofia, desde os gregos, segundo Michel Dummett, citado por Marcondes (2004, p. 9), abrange o enfoque a três questões: a questão ontológica – o ser, seus componentes fundamentais, seus estados, com os pré-socráticos; a questão do conhecimento – suas condições e possibilidades – desde o estabelecimento do método científico com Descartes, Bacon e Galileu, passando pelos empiristas com Locke, Berkley e Hume, até a síntese formulada por Kant; evoluindo mais recentemente, a partir de fins do século XIX e início do século XX, quando a humanidade passa pelo impacto crescente da Ciência, para a questão da linguagem como viabilizadora do conhecimento – a Filosofia da Linguagem.

As ciências utilizam da razão (ou pensamento) para estudar seus objetos. O pensamento pode ser definido como a capacidade de estabelecer **relações** por meio do processo de unir ou separar conceitos e objetos: a realidade é composta de complexos, e esses complexos têm ligações com algumas coisas e não com outras. Pensar é justamente o processo de descobrir ou realizar associações e disjunções, ou seja, de traçar relacionamentos. Como o significado de um complexo encontra-se nos relacionamentos que ele estabelece com outros complexos, cada relacionamento descoberto ou inventado é um significado, e grandes ordens de sistemas de relacionamentos constituem grandes corpos de significados. (MATTAR NETO, 2002, p. 33).

Textos científicos, em especial, os artigos científicos, têm como objetivo serem portadores de conhecimento novo. Como é desde há muito discutido na Filosofia, desde Aristóteles, passando por Kant – para quem que a Relação é um dos seus quatro conceitos puros do entendimento ou *categorias*, (KANT, 2001, p. 70) -, até a moderna Lógica Formal conhecimento consiste em estabelecer relações entre fenômenos da

realidade. Entre os elementos de metodologia científica, em especial é a hipótese o componente que encerra uma relação. A hipótese poderá ser simbolizada de duas formas: “Se x, então y, sob as condições r e s”, ou “Se x1, x2 e x3, então y” (MARCONI & LAKATOS, 2004, p.141), ou seja, dada certa condição X segue-se como consequência Y.). Ou então: “As leis científicas são enunciados gerais que indicam relações entre dois ou mais fatores”... “em todos os casos em que se realizam as condições A, serão realizadas as condições B” (ALVES-MAZOTTI, 2002, p. 11).

Neste sentido, as hipóteses, ao proporem relações entre fenômenos, têm importância decisiva enquanto manifestação do conhecimento novo em Ciência. Marconi e Lakatos (2004, p. 141) expressam assim este fato: “[...] podemos considerar a hipótese como um enunciado geral de relações entre variáveis (fatos, fenômenos)”. A manifestação concreta deste novo conhecimento, nos marcos institucionais da Ciência, é o artigo científico. É em torno de hipóteses que se orienta e se organiza toda a pesquisa científica, e, conforme sugerido aqui, em torno dela que se organiza toda a argumentação no texto de um artigo científico.

Todo artigo ou propõe enquanto um problema, ou desenvolve e quantifica de forma original, ou re-contextualiza uma hipótese já estabelecida por outro autor. Sob alguma destas formas, uma hipótese (ou relação entre fenômenos) deve aparecer implicitamente no texto de artigos científicos.

Também na área de Inteligência Artificial e Sistemas Especialistas (SOWA, 2000) uma base de conhecimento é formada por “regras de produção”. Estas não são mais que relações entre fatos ou fenômenos, na forma de cláusulas em lógica de primeira ordem *Se (fato ou fenômeno) então (fato ou fenômeno)*. Esta é a noção utilizada aqui, diferente de como conhecimento é visto pela Ciência da Informação (BARRETO, 1999), um processo ocorrendo na mente de pessoas.

Ontologias na Web, o “Conhecimento Público” e “e-Science”

Sendo a compreensão, e não a mera acumulação de dados, o objetivo principal da Ciência, a publicação dos resultados de pesquisa, sua discussão e análise é parte importante da incorporação da Ciência no que Ziman (1979, p.155) chama de “Conhecimento Público”. Bibliotecas e centros de pesquisa sempre tiveram um papel central no que tange ao acesso, armazenamento e disseminação de informações científicas, mas esse quadro foi ampliado e redefinido com o surgimento da Web e das TICs.

Apesar de a Web ser um grande repositório e distribuidor de informações os mecanismos de busca ainda são precários. Por exemplo, se quero saber se o uso de aspirina pode causar câncer de estômago, encontrarei artigos relacionando “aspirina” com “câncer de estômago”, mas não necessariamente artigos relacionando aspirina como **causa** de câncer de estômago. O desafio, então, é fazer com que todas as informações possam ser processadas, “interpretadas”, “entendidas” e interoperacionalizadas por computadores, diminuindo a intervenção humana e aumentando a precisão das informações obtidas, deixando para os computadores a comparação de milhões de informações para a identificação de padrões ou similaridades.

A Web Semântica proposta por Tim Berners-Lee seria uma forma de vencer esse desafio. Para que a Web Semântica se torne uma realidade é preciso que várias áreas do conhecimento trabalhem cooperativamente. Uma importante iniciativa nesse sentido é o W3C, *World Wide Web Consortium*, criado em 1994 com o objetivo de fazer com que a

Web atinja todo o seu potencial através do desenvolvimento de protocolos e diretrizes que garantam seu crescimento a longo prazo.

Um dos pontos centrais para a concretização da Web Semântica é a construção de ontologias cujo objetivo é permitir que múltiplos sistemas e agentes “entendam” o conteúdo de um recurso da Web e que possam “integrar esse conhecimento com o conteúdo de outros recursos; o sistema ou agente deve ser capaz de interpretar a semântica de cada recurso [...]” (JACOB, 2003 p.19).

Existem varias projetos que visam à construção de ontologias. Na área biológica uma das ontologias mais conhecidas é a *Gene Ontology* (GO) que é um esforço colaborativo para atender a necessidade de descrições consistentes de produtos genéticos de diferentes bases de dados.

Para “atender as necessidades de cientistas que precisam de novas tecnologias e conceitos para lidar com o número crescente de dados tanto de experimentos e simulações como de trabalhos em grupo” foi criado, em 1999, o programa Nacional de *e-Science* no Reino Unido. Um dos objetivos do programa é fazer com que os cientistas troquem e discutam os resultados de sua pesquisa de forma mais rápida e efetiva do que o modelo tradicional (leitura, discussão e crítica de artigos publicados em revistas científicas).

É cada vez mais comum a publicação de artigos científicos em formato digital. Os periódicos científicos publicados na Web são uma ferramenta cognitiva cujas potencialidades estão longe de serem avaliadas. Embora publicados na Web, os periódicos eletrônicos ainda estão calcados no modelo de publicações em papel e não utilizam todo o potencial do meio eletrônico. Eles sofrem o mesmo processo de leitura, avaliação e citação pelos pares para que os novos conhecimentos possam ser incorporados ao acervo de conhecimento público aceito num determinado campo. A Web Semântica e o futuro ambiente de E-Science trazem a perspectiva de mudanças:

Examples are the integration of authoring and reviewing processes in on-line documents. Such environments allow structured discussions of the evolution and development of an idea, paper or concept. The structured discussion is another annotation that can be held in perpetuity. This means that the reason for a position in a paper or design choice is linked to the object of discussion itself. (DE ROURE, 2001, p. 59.)

3. NOSSA HIPÓTESE

Em Marcondes (2005) é proposto uma ambiente de auto-publicação na Web – programa um editor/publicador de artigos científicos- capaz de extrair e representar em formato legível por programas agentes de “software”o conhecimento contido no texto dos artigos, simultaneamente e como subproduto do processo de editar um artigo científico pelo autor/pesquisador. Para viabilizar a construção desta ferramenta de software é proposto um modelo do raciocínio científico contido no textos dos artigos, através do qual são identificados os diferentes tipos de raciocínio e marcadas as partes componentes de cada tipo de raciocínio.

Para se chegar a este modelo foi escolhida a área de Ciências da Saúde devido ao alto grau de formalização e padronização da sua documentação. A pesquisa analisou 20 artigos da versão eletrônica do periódico **Memórias do Instituto Oswaldo Cruz**, <http://www.scielo.br/revistas/mioc>, e outros 20 do periódico **Brazilian Journal of Medical and Biological Research**, <http://www.scielo.br/bjmbr>. Artigos

científicos, em especial nessa área, têm uma estrutura altamente formalizada, a assim chamada IMRD, “*Introduction, Methods, Results, and Discussion*”, cujo objetivo é, literalmente, refletir o método científico.

Ao apresentar os resultados de sua pesquisa, autores desenvolvem linhas de argumentação ou de raciocínio. A análise de um artigo inicia-se por classificá-lo com base no tipo de raciocínio empregado. Esta classificação é baseada em Hutchins (1997) e Gross (1990) e em textos de e sobre Pierce (HOFFMAN,1997), um dos maiores estudiosos de Lógica.

Artigos podem ser *teóricos* ou *experimentais*; artigos teóricos seriam os que propõem novas hipóteses; artigos experimentais testam experimentalmente hipóteses já formuladas ou formulam e testam experimentalmente uma nova hipótese; estes podem usar os métodos de raciocínio *dedutivo* (no primeiro caso) ou *indutivo* (no segundo).

Artigos teóricos se caracterizam por discutirem questões de maior abrangência. Analisam criticamente diversas hipóteses anteriores, mostrando suas fragilidades. Estes artigos são os que têm mais potencial de apresentarem contribuições para a Ciência, já que discutem ou questionam o paradigma vigente (OLIVA, 1994). Sua contribuição é uma nova hipótese, indicando um novo caminho de pesquisa. O tipo de raciocínio empregado é o **abdução**, ou seja, o “insight” sobre a solução de questões não explicadas na Ciência e a formulação de novas hipóteses de solucioná-las.

Artigos experimentais se dividem em dedutivos e indutivos. Ambos se caracterizam por discutirem questões num escopo de abrangência limitado. Não discutem os rumos de uma teoria científica, mas se limitam a confirmá-la ou aperfeiçoá-la. Sempre trazem resultados experimentais.

A característica dos artigos que utilizam o raciocínio **dedutivo** é trabalharem a partir de hipóteses já formuladas anteriormente, cujas referências vêm citadas, aplicando-as a um contexto específico. Os artigos que utilizam o raciocínio **indutivo** se caracterizam por formularem e testarem uma proposta com certo grau de originalidade, dentro do paradigma científico vigente.

Os elementos identificados, que formam a argumentação e o raciocínio científico no texto de um artigo, são os seguintes; um PROBLEMA expressa uma carência, insatisfação ou deficiência conceitual com o atual estado de coisas num determinado domínio de conhecimento.

A partir do PROBLEMA, este é inserido numa relação que pode resolver a carência ou deficiência; esta relação é a HIPÓTESE. Uma hipótese enuncia relações entre fenômenos. Uma HIPÓTESE se desdobra em ANTECEDENTE, TIPO DE RELAÇÃO e CONSEQUENTE. Um autor num artigo pode formular uma hipótese original – HIPÓTESE(o)- ou tomar a hipótese anterior – HIPÓTESE(a) - de outros autores; neste caso uma ou mais citações referentes à HIPÓTESE(a) – CITAÇÕES(h) - são feitas.

Um autor também pode analisar várias HIPÓTESE(a) para mostrar que elas são insatisfatórias como soluções para o PROBLEMA e formular sua HIPÓTESE(o).

Um artigo teórico se justifica simplesmente por propor uma nova HIPÓTESE(o).

A hipótese, num artigo experimental, deve ter uma MANIFESTAÇÃO concreta observável empiricamente. Em um artigo científico, significa ter RESULTADOS observados segundo determinada MEDIDA, em determinado CONTEXTO segundo determinada METODOLOGIA. Este CONTEXTO onde os fenômenos relacionados na HIPÓTESE são observados pode ser desdobrado em AMBIENTE – comunidade ou instituição onde o fenômeno ocorre -, ESPAÇO - o lugar onde o fenômeno ocorre -, TEMPO ou época em que o fenômeno ocorre e GRUPO de indivíduos onde o fenômeno ocorre.

Depois de classificados quanto ao tipo de raciocínio científico, os artigos foram analisados, identificado-se os componentes descritos anteriormente. Numa segunda fase de análise, os componentes da HIPÓTESE foram relacionados ao “conhecimento público” de uma determinada área, identificando-os com termos contidos numa base de conhecimento disponível na Web, no caso a UMLS (para o ANTECEDENTE e o CONSEQUENTE) e UMLS Semantic Network (para TIPO DE RELAÇÃO).

Por exemplo, no artigo: Camara, Geni NL, Cerqueira, Daniela M, Oliveira, Ana PG *et al.* Prevalence of human papillomavirus types in women with pre-neoplastic and neoplastic cervical lesions in the Federal District of Brazil. **Mem. Inst. Oswaldo Cruz.** [online], v.98, n.7, p.879-883, 2003.

As marcações seriam as seguintes:

```
<ANTECEDENTE> papillomavirus humano</ANTECEDENTE>,  
<tipo_de_relacao>causa("causes", T147/UMLS SN)</tipo_de_relacao>  
<consequente>lesões cervicais pre-neoplásicas e neoplásicas </consequente>  
<contexto>  
    <grupo>Mulheres</grupo>  
    <espaco>Distrito Federal</espaco>  
    <espaco>Brasil</espaco>  
</contexto>
```

O conhecimento assim representado, como uma relação, permite que programas agentes de “software” realizam “inferências” sobre uma base de dados de artigos publicados de forma simultânea, como texto e segundo a proposta descrita aqui.

Artigos analisados segundo este formalismo permitiriam, por exemplo, as seguintes consultas:

- “*Que artigos (também) têm hipóteses relacionando HPV como causa de lesões pré-neoplásicas e neoplásicas em mulheres?*”
- “*Que artigos têm hipóteses relacionando outros fatores que não HPV como causa de lesões pré-neoplásicas e neoplásicas em mulheres?*”
- “*Que artigos têm hipóteses relacionando HPV como causa de lesões pré-neoplásicas e neoplásicas em outros grupos?*”
- “*Que artigos têm hipóteses relacionando HPV como causa de outras patologias em mulheres?*”
- “*Em que diferentes condições contextuais existem artigos com hipóteses relacionando HPV como causa de lesões pré-neoplásicas e neoplásicas em mulheres?*”

Um conjunto de artigos publicados na Web e representados da forma proposta, com os elementos do raciocínio empregado em cada artigo identificados e marcados, constituindo uma ontologia, pode ser facilmente comparado com ontologias públicas e com ontologias derivados de outros artigos.

Num esquema de marcação como o proposto, uma falha em estabelecer estas relações entre o “conhecimento” contido no artigo – as marcações dos elementos que compõe o raciocínio desenvolvido no artigo - e as bases de conhecimento públicas – ontologias ou outros artigos - pode ser um indício de uma nova descoberta científica.

Por exemplo, um outro artigo com a seguinte marcação:

```
<ANTECEDENTE>deficiência de vitamina E </ANTECEDENTE>,  
<tipo_de_relacao>causa("causes", T147/UMLS SN)</tipo_de_relacao>  
<consequente> lesões cervicais pre-neoplásicas e neoplásicas </consequente>
```

Se em nenhuma ontologia pública ou se em nenhum outro artigo representado segundo o formalismo proposto a “deficiência de vitamina E” é identificada como causa de “lesões cervicais neoplásicas e pré-neoplásicas”, isso pode ser um indício de novidade científica.

Esta pesquisa pretende explorar e aprofundar esta hipótese.

4. CONCLUSÕES

Todos os artigos examinados até o momento apresentaram algum grau de novidade. Mas o que seria realmente uma “novidade” em Ciência? Somente 2 artigos, no total de 40 analisados, se enquadraram no tipo de raciocínio “abduutivo”. Seriam estes tipos de artigo que encerrariam as maiores novidades científicas? Este conceito precisaria ser mais trabalhado e desenvolvido.

Pretende-se agora ampliar a validação do modelo através da análise de artigos de periódicos Internacionais em Ciências da Saúde que se caracterizem por divulgarem pesquisas inovadoras em suas áreas científicas. Recentemente, uma área de pesquisa dentro das Ciências da Saúde que tem se caracterizado por trazer avanços significativos para a Ciência tem sido a pesquisa na área de células-tronco.

Várias outras questões correlatas a nossa pesquisa ainda precisam ser melhor formuladas. Por exemplo: Como a questão da mudança de paradigma se reflete na literatura? Em que momento um fenômeno científico novo se reflete nas linguagens documentárias/ontologias?

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALPER, Brian S.; HAND, Jason A.; ELLIOTT, Susan G.; KINKADE, Scott; HAUAHN, Michael J.; ONION, Daniel K.; SKLAR, Bernard M. How much effort is needed to keep up with the literature relevant for primary care? **Journal of the Medical Librarian Association**, v. 92, n.4, p. 429-437, 2004.

ALVES -MAZZOTTI, Alda, GEWANDSZNAJDER, Fernando. **O Método nas Ciências naturais e sociais: pesquisa quantitativa e qualitativa**. São Paulo : Pioneira Thomson Learning, 2002. BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. **Scientific American**, n.5, 2001. BARRETO, Aldo de Albuquerque. A oferta e a demanda da informação: condições técnicas, econômicas e políticas.

Ciência da Informação, Brasília, v..28, n.2, maio/ago. 1999. p.168-142. Disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651998000200003&lng=pt&nrm=iso>. Acesso em 18 junho 2005.

CHRISTOVÃO, Heloisa Tardin. Da comunicação informal à comunicação formal: identificação da frente de pesquisa através de filtros de qualidade. **Ciência da Informação**, Brasília, v.8, n.1, p.3-36, 1979.

DE ROURE, David; JENNINGS, Nicholas; SHADBOLT, Nigel. **Research agenda for the Semantic Grid: a future s-Science infrastructure**. (Report commissioned for EPSRC/DTI Core e-Science Programme, 2001.

GROSS, Alan. **The Rhetoric of Science**. Cambridge, Massachusetts; London: Harvard University Press, 1990.

GUARINO, Nicola. **Formal ontology, conceptual analysis and knowledge representation**. Disponível em <http://agentes.puntodev.com.ar/download/Guarino1995_Formal_Ontology.pdf>.. Acesso em 5 jan. 2006.

HOFFMANN, Michael. Is there a “Logic” of Abduction? In: CONGRESS OF THE IASS– AIS International Association for Semiotics Studies, 6., Guadalajara, Mexico, **Anais...** Disponível em <<http://www.unibielefeld.de/idm/personen/mhoffman/papers/abduction-logic.html>>. Acesso em 14 dez. 2005.

HUTCHINS, John. 1977. **On the structure of scientific texts**. Proc. 5th. UEA Papers in Linguistics, Norwich..Norwich, UK: University of East Anglia, 1977. p. 18-39.(Conference Proceedings) Disponível em: <<http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf>>. Acesso em: 30 mar. 2006.

JACOB, Elin K. Ontologies and the Semantic Web. **Bulletin of the American Society for Information Science and Technology**, v.29, n.4, p. 19-22, 2003.

KANT, Immanuel. **Crítica da Razão Pura**. São Paulo : Nova Cultural, 1991. (Coleção os Pensadores, 7).

KUHN, Thomas Samuel. **A estrutura das revoluções científicas**. São Paulo: Perspectiva, 2005.

LE COADIC, Yves-François. **A Ciência da informação**. Brasília: Briquet de Lemos, 1996.

MARCONDES, Carlos H. From scientific communication to public knowledge: the scientific article Web published as a knowledge base. In: Egelen, Jan, Dobrev, Milena, ed. ICCC EIPub - INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, Leuven, Bélgica, 2005, 9, **Anais...** Leuven, Bélgica, 2005. p.119-27. Disponível em <<http://elpub.scix.net>>

MARCONDES, C. H; MENDONÇA, M. A. R; MALHEIROS, L.. Uma proposta de representação do conhecimento contido no texto de artigos científicos publicados na Web em formato legível por programas.**DatagramaZero** (no prelo), 2006.

MARCONDES, Danilo. **Filosofia analítica**. Rio de Janeiro : Jorge Zahar, 2004. (Coleção Passo a passo).

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Metodologia científica**. São Paulo : Editora Atlas, 2004.

MATTAR NETO, José Augusto. **Metodologia científica na era da informática**. São Paulo : Saraiva, 2002.

MEADOWS, Arthur Jack. **A comunicação científica**. Brasília : Briquet de Lemos, 1999.

OLIVA, Alberto. Kuhn: o normal e o revolucionário na reprodução da racionalidade científica. IN: PORTOCARRERO, Vera. (Org.). **Filosofia, história e sociologia das ciências**: abordagens contemporâneas. Rio de Janeiro: Fiocruz, 1998. p. 67-102.

SANTAELLA, Lucia. O método anticartesiano de C.S.Peirce. São Paulo: Editora UNESP, 2004.

SOWA, John. Knowledge representation: logical, philosophical and computational foundations. Pacific Grove : Brooks/Cole, 2000.

ZIMAN, John. **Conhecimento público**. Belo Horizonte: Ed. Itatiaia; São Paulo : EdUSP, 1979.