

## A ciência da informação e a ciência da computação desenvolvendo a recuperação da informação

Késsia R. da C. Marchi<sup>1</sup>, Ayslan T. Possebom<sup>1</sup>

<sup>1</sup> Informação e Comunicação/Informática – Instituto Federal do Paraná – IFPR  
Avenida José Felipe Tequinha, nº. 1400 – Jardim das Nações – 87703-536 –  
Paranavaí, PR – Brasil

kessia.marchi@ifpr.edu.br, ayslan.possebom@ifpr.edu.br

**Abstract.** *Information retrieval has been researched in the fields of Information Science and Computer Science. In this context, this work is based on a brief literature review, seeking to contemplate Information Retrieval in both areas, aiming to establish the relationships of this interdisciplinarity that favor the development of Information Retrieval.*

**Resumo.** *A Recuperação da Informação tem sido pesquisada nos âmbitos da Ciência da Informação e da Ciência da Computação. Diante disto, este trabalho fundamenta-se em uma breve revisão de literatura, buscando contemplar a Recuperação da Informação em ambas as áreas, objetivando estabelecer as relações desta interdisciplinaridade que favorecem o desenvolvimento da Recuperação da Informação.*

### 1. Introdução

A recuperação da informação é uma das principais características envolvidas em sistemas informatizados. Sem esta recuperação, os sistemas são incapazes de fornecer relatórios ou efetuar gerenciamentos. Quando falamos sobre informação, outros termos se relacionam indiretamente. São eles: dados, informações e conhecimento. Conforme apresentado por Chen [Chen et al. 2008], podemos ler dados, aprender informações e adquirir conhecimentos, entretanto, devemos diferenciar esses termos em questões de espaço perceptivo e espaço cognitivo.

Liew [Liew 2007] apresenta diversas definições para dados, informações e conhecimentos. Para ele, os dados dizem respeito a informações básicas, não refinadas e geralmente não filtradas. São elementos básicos geralmente representados por valores discretos. Quando estes dados são organizados de alguma forma, estes se tornam informações. Pode-se considerar informação como um dado processado, formalizado, capturado e explicado. As informações são utilizadas em algum contexto e possuem algum significado para as pessoas, se tornando conhecimento. Logo, conhecimento está relacionado à experiência humana e o *insight* quanto ao uso destas informações.

Diante da importância do processamento dos dados e da obtenção de conhecimentos, estes dados devem ser armazenados de alguma forma e recuperados posteriormente para fornecer novos conhecimentos. O processamento de dados, bem como a modelagem do relacionamento entre os dados e a recuperação destes dados, transformando-os em informações, é tarefa de interesse multidisciplinar, envolvendo principalmente a Ciência da Informação e a Ciência da Computação.

A Ciência da Informação (CI) tem como objetivo investigar o fenômeno Informação analisando suas propriedades e comportamento, assim como, a origem, coleta, organização, os meios de processamento, o armazenamento, o uso, a recuperação e a transmissão da informação, tanto manuais, quanto digitais [Borko 1968, Basha and Khapre 2012].

Saracevic [Saracevic 1996] redefiniu a CI como um campo dedicado às questões científicas e à prática profissional, voltadas a problemas de comunicação do conhecimento. O autor ainda complementa que são consideradas as vantagens das tecnologias informacionais. Em seus trabalhos, Borko [Borko 1968] e Saracevic [Saracevic 1996] afirmam que a CI é também uma ciência interdisciplinar derivada e relacionada a campos como a ciência da computação e outros campos do saber.

Freire [Freire 2006] entende que a perspectiva científica da informação foi a inovação no campo da produção e comunicação do conhecimento científico e esta inovação resultou em possibilidades de criação de tecnologias de informação que, em um momento inicial preocupou-se com o armazenamento da informação e a sua disseminação para grupos específicos e hoje preocupa-se com a distribuição e Recuperação da Informação (RI) no contexto da sociedade em geral.

A recuperação da informação tem sido pesquisada tanto na área da Ciência da Informação como na área da Ciência da Computação (CC). Em ambas, a RI refere-se à interação dos usuários com os sistemas de recuperação, preocupando-se com o julgamento sobre resultados relevantes a partir dos documentos obtidos por meio da seleção resultantes da estratégia de busca [Basha and Khapre 2012].

Diante do exposto, é possível observar que a CI estabelece uma relação com a CC no que se refere à Recuperação da Informação. Assim, este trabalho tem por objetivo apresentar uma breve revisão de literatura buscando contemplar os precursores das temáticas de Recuperação da Informação nos âmbitos da Ciência da Informação e da Ciência da Computação. Para isto, é relatado neste documento os fundamentos da Recuperação da Informação, sua relação com a relevância e seus aspectos em cada uma destas áreas de conhecimento, assim como as relações que ocorrem entre as áreas.

Este artigo está organizado da seguinte maneira. A seção 2 apresenta os fundamentos teóricos envolvidos na recuperação da informação e relevância da informação, em especial quando aplicados às áreas de ciência da informação e ciência da computação. A seção 3 discute os aspectos inerentes da recuperação da informação. Por fim, na seção 4, as conclusões são apresentadas.

## **2. Recuperação da informação e relevância da informação**

O termo Recuperação da Informação foi apresentado por Calvin Mooers no início dos anos 50 e é reconhecido como área de pesquisa interdisciplinar, englobando, entre outras, a Ciência da Informação e a Ciência da Computação [Saracevic 1975].

A explosão informacional vivenciada a partir da II Guerra Mundial tornou a RI necessária em várias áreas de conhecimento, se mostrando como uma solução bem sucedida, entretanto, com seus problemas específicos [Saracevic 1996]. Nesse contexto, Mooers [Mooers 1959] contemplou estes problemas em uma concepção que aborda, entre outros, três pontos que continuam fundamentais até os dias atuais, sendo eles: Como descrever

intelectualmente a informação?; Como especificar intelectualmente a busca? e; Quais sistemas, técnicas ou máquinas devem ser empregados? A partir destas questões, surgiram variedades de conceitos e estudos teóricos, empíricos e pragmáticos, assim como, diversas realizações práticas, no contexto da RI.

A necessidade de recuperar informações tem promovido pesquisas nas áreas da CI e da área da CC. Em ambas, a recuperação da informação refere-se à interação dos usuários com os sistemas de recuperação, preocupando-se com o julgamento sobre resultados relevantes a partir dos documentos obtidos por meio da seleção resultantes da busca [Basha and Khapre 2012].

À vista disto, o termo relevância surge como uma noção fundamental da CI e é atribuído aos problemas da comunicação. A relevância é considerada como uma medida da eficácia de um contato entre uma fonte e um destino em um processo de comunicação. Diferentes pontos de vista surgiram porque a relevância foi considerada em vários enfoques diferentes no processo de comunicação [Saracevic 1975].

De forma geral, relevância tem a ver com eficácia e comunicação. Todos os sistemas de informação, de uma forma ou outra, apresentam alguma interpretação da noção de relevância. Quanto mais abrangente for o entendimento do significado de relevância, melhores serão as chances de evitar falhas e restringir a variedade de erros cometidos em nome da comunicação efetiva, consequentemente, melhorando os sistemas de informação [Saracevic 1975].

Isto posto, é possível constatar que os sistemas de recuperação de informação surgiram como uma correção tecnológica para um problema, ou seja, facilitar que cientistas, e hoje público em geral, localize informações que são relevantes para o seu contexto. Desta forma, a RI é o ponto de convergência de técnicas, métodos e tecnologias para um fim comum.

## 2.1. A recuperação da informação na ciência da informação

A área de CI, quando comparado a outras áreas de produção tecnológicas, possui um enfoque maior no que tange a descoberta de fontes de informações e normatizações para a representação destas informações e do conhecimento. O profissional desta área deve conhecer conceitos básicos sobre hardware e software, além de questões envolvidas aos sistemas operacionais, mecanismos de armazenamento e gerenciamento de dados, segurança das informações e redes de computadores. Questões mais avançadas relacionadas à representação dos dados, web semântica, ontologias e *Resource Description Framework* (RDF), em especial quando voltadas à área de Inteligência Artificial, são também de interesse deste profissional. Mais atualmente, a CI tem focado na pesquisa, recuperação e disseminação de conhecimentos na área de Internet das Coisas, *Big Data* e Ciência de Dados, possuindo forte conexão com a CC.

Para a CI, a Recuperação de Informação é um processo dinâmico, iterativo e variável, que deve ser realizado em canais baseados em tecnologias ou não, escolhidos de acordo com as necessidades de informações e contexto do usuário e concentra-se em fatores cognitivos, emocionais, na avaliação de resultados de pesquisa, no uso final e em fatores ambientais [Basha and Khapre 2012].

Saracevic [Saracevic 1996] considera a recuperação da informação como uma ver-

tente tecnológica da CI, o que permite estabelecer uma relação efetiva com a Ciência da Computação.

## 2.2. A recuperação da informação na ciência da computação

A maioria das pesquisas de recuperação de informações na CC estão explorando maneiras, a fim de fornecer informações relevantes, formas eficazes de armazenamento, de indexação e recuperação de informações [Basha and Khapre 2012].

Manning [Manning et al. 2008] classificou os sistemas de RI em três escalas, onde, na escala inicial encontra-se a RI pessoal, a qual apresenta como objetivo realizar a recuperação de informação em computadores pessoais. Na escala intermediária, encontra-se os sistemas empresariais, institucionais e de busca em um domínio específico. E na escala mais avançada, tem-se a busca na Web, na qual o sistema realiza a pesquisa na Web e deve ser capaz de trabalhar de forma eficiente, tornando possível que o usuário empregue o menor esforço produzindo o melhor resultado.

Independentemente da escala, consultar refere-se à utilização de símbolos em combinação com outras sintaxes como comandos para localizar o índice no conteúdo de um sistema de RI. Estes símbolos podem ser palavras, conjuntos de palavras, documentos, entre outros, e pode haver mais de uma combinação possível. Basicamente, são três técnicas aplicadas em sistemas de RI citadas na literatura como modelos considerados clássicos, sendo o Booleano, o modelo vetorial e o de recuperação probabilística [Cardoso 2004, Basha and Khapre 2012].

O modelo booleano é baseado na teoria dos conjuntos e na álgebra booleana. Os documentos são representados por um conjunto de palavras-chave que podem ser extraídas de forma automática ou manual. Os documentos recuperados contêm os termos que satisfazem a expressão lógica da consulta, sendo essa consulta uma expressão booleana formada com os conectivos lógicos AND, OR e NOT [Manning et al. 2008].

O Modelo Vetorial é utilizado por conseguir representar documentos e consultas na forma de vetores de termos. Ele apresenta vantagens por sua facilidade de computar a similaridade, fato esse que permite que o modelo se comporte bem com documentos genéricos. Suas desvantagens estão nas limitações que impedem a utilização da sinonímia, da polissemia e do contexto local e global e na ausência de ortogonalidade entre os termos, o que poderia acarretar relações entre termos que não apresentam similaridade e também a não seleção de documentos relevantes que não apresentem os termos das consultas [Micarelli et al. 2007].

Por fim, o modelo probabilístico é representado por meio de um vetor binário de pesos semelhante ao modelo booleano, com a diferença de que o peso de cada termo é gerado por meio do cálculo da probabilidade de que o documento realmente seja relevante para a consulta, baseando-se no princípio da ordenação probabilística. A principal vantagem deste modelo está relacionada à possibilidade de classificar os documentos em ordem decrescente, considerando a probabilidade da sua relevância. Suas limitações estão no fato de assumir a independência entre os termos, não explorar a frequência dos termos, ignorar os problemas de filtragem de informações e depender da precisão das estimativas de probabilidade [Manning et al. 2008, Micarelli et al. 2007].

A ordenação dos resultados oferecidos por meio das tarefas de RI é estabelecida

de acordo com o modelo de RI que está sendo aplicado. No entanto, nem sempre as informações oferecidas pelos usuários são suficientes para apresentar resultados relevantes. É neste cenário que se faz necessário encontrar recursos alternativos. Uma alternativa interessante é o uso de modelos de análise semântica mais sofisticados. Neste contexto, técnicas de Inteligência Artificial têm sido propostas na busca de melhorias no tratamento textual.

O campo de pesquisa de RI no âmbito da computação tem se concentrado em melhorar o desempenho técnico, entretanto, diversos aspectos tem sido considerados em pesquisas interdisciplinares referentes a RI, o que tem levado a melhorias nos processos e sistemas de recuperação de informação.

### **3. Aspectos de RI inerentes a CI e a CC**

Por sua característica interdisciplinar, as pesquisas pertinentes a RI apresentam aspectos inerentes e complementares entre a CI e a CC. Dentre estes aspectos, Basha e Khapre [Basha and Khapre 2012] relatam a relevância como um campo transversal com base na perspectiva da informação. No âmbito da CC, os algoritmos desenvolvidos consideram taxas de precisão e recall como medidas de avaliação de desempenho dos sistemas de RI. No campo da CI, as pesquisas concentram-se no estado de conhecimento, intenções, objetivos e motivações do usuário.

Considerando a perspectiva do usuário, Basha e Khapre [Basha and Khapre 2012] relatam os princípios da incerteza e da interação. A incerteza normalmente parte da perspectiva do usuário, refere-se a um estado cognitivo, formado pela falta de entendimento ou conhecimento limitado e apresenta-se nos estágios iniciais da RI. Objetivando auxiliar os usuários na definição de consultas mais claras, a CC tem realizado pesquisas de RI que se concentram no uso de algoritmos para reduzir as incertezas, refinar as consultas, amenizar problemas de vocabulários e apresentar feedback de relevância.

O princípio da interação, por sua vez, é considerado como fator chave em pesquisas de RI. A interação é uma parte importante da RI, incluída em quase todos os modelos e modos de recuperação de informações. Saracevic [Saracevic 1996] enfatizou a importância da interação no processo de recuperação de informações, desenvolvendo um modelo que se concentra nos diferentes níveis de interação entre usuário e sistema, fornecendo uma série de filtros interativos.

O acesso à informação é outro aspecto analisado por Basha e Khapre [Basha and Khapre 2012] nas áreas da CI e da CC e o autor afirma que quanto mais fácil o acesso à informação, maior a probabilidade de as pessoas usarem. Para tornar as informações mais acessíveis o pesquisador deve estudar profundamente, a expressão de consulta, a ajuda de contexto e a visualização de informações. É especialmente adequado para a CC porque esta área está mais preocupada com o design e desenvolvimento de sistema. Já no campo da CI, muito trabalho visa melhorar a facilidade de acesso. Portanto, a arquitetura é o núcleo das duas disciplinas.

### **4. Conclusões**

Com base no exposto a partir da breve revisão de literatura que considerou aspectos inerentes a Recuperação da Informação nos âmbitos da Ciência da Informação e da Ciência

da Computação pode-se verificar, não somente uma estreita relação entre estas áreas, mas que as preocupações, métodos e tecnologias utilizadas são comuns entre elas, embora em contextos diferentes.

A Recuperação da Informação na CI engloba os aspectos intelectuais da descrição de informações e suas especificidades para a busca, objetivando desenvolver novas técnicas de manipulação da informação. Já a aplicação das teorias e técnicas da CI permitem criar, modificar e melhorar os sistemas de manipulação e gestão da informação [Basha and Khapre 2012, Saracevic 1975].

As pesquisas relacionadas à RI no campo da Ciência da Computação são construídas objetivando fornecer informações úteis aos usuários, logo, estudam o armazenamento e a recuperação automática de documentos, considerados objetos de dados. Portanto, essas pesquisas seguem o princípio racional e tradicional de que a informação é a existência factual do mundo externo das coisas, pois, a informação é considerada uma unidade de presença física, como texto, imagens e vídeos [Cardoso 2004, Basha and Khapre 2012].

Em se tratando das contribuições entre as áreas da CI e da CC nas pesquisas relacionadas à RI, é possível observar a aplicação das tecnologias e da organização informacional. Neste contexto, a CC fornece meios tecnológicos para o desenvolvimento de ambientes de RI e a CI contribui com métodos e técnicas de tratamento informacional, fazendo uso das aplicações tecnológicas.

A atuação conjunta destas áreas tem favorecido o desenvolvimento de estruturas tecnológicas e informacionais que contemplam metodologias de tratamento sedimentadas na CI e aplicações tecnológicas implementadas na CC, desta forma, estas áreas se fazem complementares possibilitando resultar em métodos e ferramentas que fornecem melhores resultados aos usuários de sistemas de recuperação de informação.

## Referências

- Basha, S. and Khapre, S. (2012). A theoretical paradigm of information retrieval in information science and computer science. *International Journal of Computer Science Issues*, 9(5):232–240.
- Borko, H. (1968). What is information science? *Journal of Soviet Education*, 28(10):51–54.
- Cardoso, O. N. P. (2004). Recuperação de informação. *INFOCOMP Journal of Computer Science*, 2:33–38.
- Chen, M., Ebert, D., Hagen, H., Laramée, R. S., Liere, R. V., Ma, K.-L., Ribarsky, W., Scheuermann, G., and Silver, D. (2008). Data, information, and knowledge in visualization. *IEEE computer graphics and applications*, 29(1):12–19.
- Freire, G. H. (2006). Ciência da informação: temática, histórias e fundamentos. *Perspectivas em Ciência da Informação*, 11(1):6 – 19.
- Liew, A. (2007). Understanding data, information, knowledge and their inter-relationships. *Journal of knowledge management practice*, 8(2):1–16.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

- Micarelli, A., Sciarrone, F., and Marinilli, M. (2007). *Web Document Modeling*, volume 4321, chapter 5, pages 155–192. Springer, Berlin, Heidelberg. The Adaptive Web: Methods and Strategies of Web Personalization.
- Mooers, C. N. (1959). The next twenty years in information retrieval: some goals and predictions. *American Documentation*, 11(638):81–86.
- Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321 – 343.
- Saracevic, T. (1996). Ciência da informação: origem, evolução e relações. *Perspectivas em Ciência da Informação*, 1(1):41 – 62.