Anais do XIII Encontro Anual de Tecnologia da Informação - EATI
| Ano 11, n. 1 | novembro/ 2022

Considerações para o desenvolvimento de um sistema de recuperação de informação em ambientes com características de Big Data, orientados ao contexto de uso da informação

Luís Roberto Momberg Albano[1]

1Escola de Comunicação e Artes – Universidade de São Paulo (ECA-USP)
Av. Prof. Lúcio Martins Rodrigues, 443 - Butantã, São Paulo - SP, 05508-020

albano@usp.br

Resumo. *Entende-se que ambientes com características de Big Data possuem características próprias de seu desenvolvimento tecnológico, como a não disposição para identificação de seu usuário, bem como a necessidade organização de dados produzidos rapidamente, em grande volume e variabilidade. Assim, discute-se o potencial da utilização do contexto de uso da informação, de forma introdutória, a fim de guiar pesquisas nesse sentido. São apresentados conceitos iniciais para nortear uma revisão bibliográfica acerca do tema, aprofundadas no sentido de entender as problemáticas possíveis e potenciais soluções.*

Abstract. *It is understood that environments with Big Data characteristics have their own technological development characteristics, such as the unwillingness to identify their user, as well as the need to organize data produced quickly, in large volume and variability. Thus, the potential of using the context of information use is discussed, in an introductory way, in order to guide research in this direction. Initial concepts are presented to guide a bibliographic review on the subject, in-depth in order to understand the possible problems and potential solutions.*

1. Introdução

Parte das preocupações da Ciência da Informação (CI) envolvem a produção, armazenamento, recuperação e uso da informação (ARAÚJO, 2009; LOGAN, 2012). No entanto, se considerar os impactos dos desenvolvimentos das tecnologias de informação e comunicação (TICs), na produção e uso das informações, observa-se que os avanços tecnológicos têm propiciado um substancial aumento da produção intelectual e, consequentemente, informacional. Nesse sentido, também se observa que, neste cenário, se consolida a necessidade por estratégias próprias com vistas a, majoritariamente, coletar, armazenar, organizar e recuperar a informação, por quem, quando e onde se necessita desta para realizar de tarefas próprias do cotidiano.

Usualmente, os sistemas de informação, particularmente os ditos “tradicionais” , são projetados tendo como referência o potencial usuário. E, para tanto, trabalha-se com a noção de “usuário ideal” interagindo com o sistema em contexto(s) específico(s). Porém, frente à enorme de produção de informações – em termos de volume e variedade,

incluindo a diversidade de contextos e produtores –, o desafio que se coloca é o de qualificar a recuperação da informação, a ponto de se obter a informação adequada (útil) ao usuário no contexto em que este se encontra, no momento de sua interação com o sistema. Tal desafio é potencializado num momento em que cresce a utilização de variados instrumentos na produção, acesso e uso de informações, nos diferentes contextos sociais, econômicos e culturais.

De um modo bem geral, esta crescente produção de informações – em termos de volume, variedade, velocidade, variabilidade, veracidade e valor, dentre outros aspectos – identifica o que se denomina Big Data (BARLOW, 2013; DAVENPORT, 2014; DE MAURO; GRECO; GRIMALDI, 2016; MARQUESONE, 2016). Portanto, o desafio de qualificar a recuperação da informação se assenta no fato de que as perguntas (ou queries) apresentadas aos sistemas podem surgir de “qualquer lugar". Logo, percebe-se que isto representa um “ponto de inflexão” na concepção e implementação de instrumentos utilizados na recuperação de informação. Em parte, isto também se deve ao fato de que, tradicionalmente, os instrumentos de recuperação de informação são baseados em abordagens descritivas e informações estruturadas – com o suporte de metadados, por exemplo. Ocorre, porém, que nem tudo o que é produzido – e na velocidade que é produzido – pode ser descrito e organizado de modo que estas abordagens possam subsidiar a recuperação. A este respeito e frente à ampla utilização das TICs, observa-se que novas abordagens são baseadas em aprendizado de máquina (machine learning) tendem a contribuir para o aprimoramento de instrumentos de recuperação da informação, particularmente, quando se tem ambientes informacionais digitais com características de Big Data, os quais armazenam grandes volumes de informação produzida em curtos períodos de tempo, podendo ser multimídia (textos, sons, imagens, vídeos etc), com pouca ou nenhuma estruturação (DE MAURO; GRECO; GRIMALDI, 2016; MARQUESONE, 2016).

Com isto e considerando que as abordagens baseadas em aprendizado de máquina dependem, fundamentalmente, do projeto e implementação de algoritmos computacionais – atualmente, desenvolvidos com base em princípios da Matemática, Estatística e Ciência da Computação –, parte-se do pressuposto que, conhecendo o usuário e o contexto em que este se encontra, os processos de recuperação da informação nestes ambientes informacionais digitais podem ser aprimorados. Dessa forma, neste trabalho, apresenta-se a proposta de um modelo conceitual para recuperação da informação em ambientes digitais com características de Big Data.

2. Big Data e Informação

Pode-se, de um modo geral, caracterizar Big Data como um ativo informacional, heterogêneo, de grande volume, produzido em curto período de tempo, com variabilidade em termos de codificação dos conteúdos, por vezes não estruturado, dependente de ferramentas computacionais para processamento (BARLOW, 2013; DE MAURO; GRECO; GRIMALDI, 2016; RAUTENBERG; CARMO, 2019; RIBEIRO, 2014). Este ativo é constituído por dados. Considerando dado como fato sobre alguma condição

específica – e que, sem um contexto e ou relação com outros dados, tem pouca serventia – e o mesmo é produzido em grandes volumes e em curtos espaços temporais, espera-se que, com estruturação e organização, seja possível "minerá-lo" e produzir informações, com valor para o usuário demandante. Ocorre que, apesar das abordagens – por exemplo, "mineração de dados" (AMBRÓSIO; MORAIS, 2007) – utilizadas para organizar e obter informações para os potenciais usuários de um sistema de informação com características de Big Data, ao menos, duas problemáticas se destacam: (1) a necessidade de processar, em tempo real ,grandes volumes de dados; e (2) a ausência de um usuário específico para recuperação da informação (BARLOW, 2013). Em parte, tais problemáticas evidenciam a forte dependência de algoritmos computacionais. Neste caso, estes algoritmos são utilizados na tentativa de oferecer uma resposta "mais adequada", a qual possa contemplar a tríade "informação disponível" X "demanda informacional" X "apresentação do conteúdo de forma que este possa ser apropriado pelo usuário demandante". Em geral, isto se dá pela identificação de possíveis relações entre usuário, palavras-chaves e estudos do comportamento informacional. Contudo, a construção destas relações não se apresenta como algo trivial em ambientes informacionais digitais em que não sabe de onde as perguntas surgem e, menos ainda, como as respostas serão utilizadas.

Por isto, observa-se que a modelagem e implementação de algoritmos computacionais para recuperação de informação em ambientes informacionais com características de Big Data tende a incorporar determinadas estratégias tidas como "ideais" na perspectiva do desenvolvedor (O'NEIL, 2021), o qual usualmente se pauta em estudos de casos de uso e lógica de negócios da aplicação, quando este conhece as condições de uso. Contudo, apesar das diversas iniciativas para se desenvolver sistemas e ferramentas computacionais centrados no usuário e sensíveis ao contexto (context-aware), ainda se observa dificuldades na implementação e oferta de métodos computacionais capazes de oferecer a informação plenamente adequada (relevante) ao usuário demandante. Assim, neste trabalho, parte-se de conjectura que a incorporação de elementos associados aos contextos de uso pode contribuir para o aprimoramento de algoritmos computacionais de recuperação da informação, uma vez que são esperadas respostas relevantes (no mínimo, com pertinência) para o usuário demandante, ao mesmo tempo em que tais uso e usuários não podem não ter sido previamente pensados na fase de projeto do sistema.

3. Usuário e contexto

Entende-se por usuário da informação o indivíduo que demanda informações para o desenvolvimento de suas atividades, sejam estas profissionais ou cotidianas (GUINCHAT; MENOU, 1994; SANZ CASADO, 1994). Nesse sentido, cabe observar que quando o usuário busca uma informação, este o faz sob a influência de diversos aspectos, geralmente relacionados às condições específicas de uso da informação, ou seja, seu uso (CHOO, 2003; LE COADIC, 2004). Tal uso pode estar relacionado com a resolução de um problema (FIGUEIREDO, 1979, 1994) ou preenchimento de uma lacuna informacional. Independente, porém, da razão pela qual um usuário realiza uma busca, observa-se um esforço cognitivo para explicitar a informação demandada (CHOO, 2003). Tal esforço está relacionado com a perspectiva que o indivíduo tem acerca do problema

com o qual se depara (CAPURRO; HJØRLAND, 2003; CASTELLS, 1996; PEIRCE, [s.d.]).

Logo, sistemas de informação criados com o intuito de recuperar a informação "mais adequada" para um determinado usuário deveriam ser projetados com direcionamento às suas necessidades e demandas de uso, levando em conta questões socioculturais deste usuário (GUINCHAT; MENOU, 1994; SARACEVIC et al., 1988). Pois, as condições específicas de uso da informação, normalmente, são dadas pelo contexto em que este usuário se encontra.

Cabe ponderar que a definição de contexto não é tarefa simples, dado que há certa variação de entendimento a depender da abordagem adotada. Por exemplo, Pete Steggles (1999), em painel organizado por Abowd e Dey (1999), menciona que "qualquer descrição do mundo que pode ser relevante para um aplicativo conta como um 'tipo de contexto'" (ABOWD; DEY, 1999, p. 5, tradução nossa). Na mesma linha, Sezer, Dogdu e Ozbayoglu (2018, p. 10), explicitaram que contexto pode ser

> [...] qualquer informação que possa ser utilizada para caracterizar a situação de uma entidade. Uma entidade é uma pessoa, lugar ou objeto considerado relevante para a interação entre um usuário e um aplicativo, incluindo o usuário e os próprios aplicativos.

E, anteriormente, trazendo elementos da Comunicação, Foresti, Varvakis e Godoy Viera (2016, p. 4) afirmam "que os sentidos do termo contexto abarcam os dados, a informação, o conhecimento, o texto, o ambiente, o emissor e o receptor." E, em perspectiva normativa sobre usabilidade, a norma NBR 9241-11 (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2021), com base na norma ISO/IEC 25063 (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2014) apresenta alguns pilares sobre contexto de uso, indicando a necessidade de se considerar o usuário, a tarefa a ser realizada, o equipamento utilizado e seu ambiente de operação. Estes pilares remetem ao contexto de uso. Neste caso, observa-se que o contexto de uso representa um balizador importante para a oferta da informação "mais adequada" ao usuário, ou seja, aquela que privilegia elementos como forma, conteúdo e acessibilidades (física e cognitiva) na oferta de resposta ao usuário. Assim, de modo complementar, em termos de recuperação da informação, além do contexto de uso, deve-se observar o contexto de produção. Este está relacionado com as relações espaço-tempo (ABOWD; DEY, 1999). Logo, o ambiente digital contemporâneo deve balizar contexto de uso, necessidade informacional e contexto de produção da informação com potencial de ser utilizada em resposta à uma demanda de usuário. A este respeito, Dey (2000) propôs que informações contextuais poderiam ser utilizadas para o projeto de sistemas de informação mais consistentes (*context-aware* ou *context-awareness*). Esta abordagem auxilia na seleção de estratégias que possam aprimorar a oferta de informação com relevância (pertinência), a um dado usuário, com base no contexto em que este se encontra.

4. Recuperação da Informação

Conforme já mencionado, usualmente, o usuário, frente a um problema (ou necessidade informacional), empreende ações de busca por informações que possam auxiliar na solução deste problema (ou suprir a necessidade). Neste cenário, identifica-se o que se denomina comportamento informacional e os sistemas de recuperação de informação, nesse sentido, podem ser entendidos como capazes de, por meio do acesso a um "estoque", recuperar e "entregar" informações que possam auxiliar o usuário na solução do referido problema. E, de uma forma geral, quando o usuário percebe a pertinência da informação recuperada e se apropria da mesma, ocorre uma alteração em seu estado de conhecimento. Cabe lembrar que a plena utilização da Informação recuperada por parte do usuário depende de sua capacidade de se apropriar da mesma, de forma que possa operacionalizar com ela. Com isto, o projeto de sistemas de recuperação de informação, portanto, deve pressupor a oferta de interfaces capazes de desempenhar a mediação (ou aproximação) entre os usuários e os estoques de informação, favorecendo a apropriação. Normalmente o usuário realiza sua busca em linguagem natural, ou linguagem com a qual foi aprendeu a se comunicar. A recuperação da informação em ambientes digitais, por outro lado, demanda de mediação entre a linguagem humana e a linguagem de máquina (código), realizando, através de algoritmos, traduções entre busca e potencial recuperação (MITLETON-KELLY; LAND, 2008). A recuperação de informação digital, ao não ter seu projeto orientado a um usuário específico, possui dificuldade para lidar com qualquer tipo de incerteza, pois demanda de compreensão contextual específica do usuário. Assim, compreender o contexto de uso de uma informação para um usuário se torna de fato interessante para o projeto deste tipo de sistema, tendo em vista que insere o usuário como protagonista do processo de busca e recuperação da informação.

5. Um modelo para recuperação da informação baseado em contexto

Considera-se que um modelo conceitual pode prover bases para o desenvolvimento de um sistema de recuperação de informação baseado em contexto, que mitigue os problemas supracitados. Para fins de modelagem, os atributos contextuais de uso considerados para coleta são da ordem de perfil do usuário (estrutura linguística, simbólica, cognitiva e emocional), palavras-chave de busca (intenção e conhecimento prévio da busca), razão da busca (conteúdo) e utilização da informação a ser recuperada (forma, conteúdo e acessibilidades), notados a partir de revisões feitas em Talja et al. (1999), Foresti, Varvakis e Godoy Viera (2016), Choo (2003), Adams, Schilit e Want (1994), Henrique, Nassif e Venâncio (2007), Johnson (2003). Entende-se que o modelo, em um primeiro momento, deve se expandir no sentido de coletar variações de atributos nessas ordens, vinculados aos eventos de busca e recuperação, com o intuito de criar relações. Estes atributos do contexto de uso devem ser coletados a partir de processo interativo e, também, - dentro das permissões possíveis legais e por parte do usuário - a partir de cookies ou outros algoritmos sensíveis de percepção do sistema.

O treino interno do sistema deve se dar pela relação entre atributos do contexto de uso e atributos do contexto de produção da informação, tendo os pesos de cada atributo

calibrados a cada treino feito em cada processo de recuperação de informação, a depender de feedback dado pelo usuário. De forma sucinta, o sistema opera conforme segue: o usuário apresenta para o sistema mediador seu problema ou demanda, em linguagem natural. O sistema mediador, então, realiza processo interativo a respeito de condições contextuais que podem auxiliar na identificação do usuário e suas demandas, relacionadas a seu uso (quem é usuário, o que, por que e para que). O usuário fornece esses dados, fazendo com que o sistema mediador solicite ao estoque informacional a seleção das informações a serem recuperadas a partir do explicitado, comparando contexto de uso com contexto de produção – o estoque informacional já pode conter condições possíveis previamente treinadas entre de busca e recuperação da informação. As informações do estoque informacional são organizadas em nível de forma e conteúdo e enviadas para o sistema mediador, para que este as apresente opções para o usuário.

Após apresentação, o usuário seleciona qual a melhor combinação de forma e conteúdo parece lhe atender, e é questionado sobre a qualidade da recuperação, indicando resposta positiva ou negativa. Em caso positivo, o processo é encerrado e os dados relacionando atributos contextuais de uso de produção são enviados para o estoque informacional, criando um mapa relacional que será utilizado para aprimoramento do sistema. Em caso negativo, a partir de relações anteriormente feitas com condições similares, novo processo de inquirição é feito, dando continuidade ao ciclo que se encerra apenas quando o usuário entende que a recuperação feita lhe atende. O sistema, dessa forma, é treinado no sentido de relacionar atributos contextuais de uso e produção, a fim de recuperar por perfilamento individual a melhor informação possível para um usuário específico.

7. Considerações finais

O modelo proposto contribui no sentido de dar instrumentos teóricos para a modelagem de sistemas de informação com características de Big Data, a fim de mitigar a problemática da identificação de seu usuário, bem como do contexto de sua demanda (BARLOW, 2013). O processo de recuperação de informação é um processo de mediação, onde o usuário explicita seu problema e o sistema de informação identifica, em seu estoque, a melhor informação possível a ser recuperada (NETO; ALMEIDA JÚNIOR, 2017). Ao se utilizar processos de inquirição e negociação (CHOO, 2003) para a identificação dos atributos contextuais de uso, fica possível uma melhor qualificação do perfil do usuário e de sua necessidade, fazendo com que a recuperação de informação seja mais personalizada e individualizada. Isto mitiga o problema dos algoritmos computacionais de seleção de perfis para agrupamento, utilizados em ambientes com características de Big Data que, considerando aspectos de proximidade, qualificam indivíduos como possuindo as mesmas características (BOYD; CRAWFORD, 2012; MARQUESONE, 2016). Ao realizar esse perfilamento mais individualizado, o modelo é capaz de, comparando atributos contextuais de uso, produção, e recuperações anteriores, apresentar uma informação mais delimitada para seu problema. Assim, são consideradas as lacunas cognitivas de conhecimento do usuário sobre sua busca (CHOO, 2003), onde o conteúdo a ser recuperado é alinhado com seu estado anômalo de conhecimento (BELKIN, 1980). Atributos linguísticos (FORESTI; VARVAKIS; GODOY VIERA, 2016), nesse modelo, são alinhados à atributos locais espaciais (ADAMS; SCHILIT;

WANT, 1994), atuando como delimitadores idiomáticos, históricos, de estilo e também de estado anômalo de conhecimento, através do recorte instrumental de conhecimento de termos e jargões. Isto, alinhado à atributos simbólicos, capturados em processo de inquirição e sensoriamento, podem ser utilizados para a seleção das melhores formas de apresentação da informação para um usuário. Nesse sentido, as estruturas de mediação desse tipo de sistema podem ser aprimoradas, aproximando um usuário distinto de um ambiente informacional com características de Big Data. Nota-se, assim, que a noção de contexto de uso representa um elemento central tanto para a concepção de algoritmos computacionais de recuperação da informação, quanto de interfaces disponibilizadas aos usuários destes sistemas.

Entende-se que o presente modelo deverá fazer uso de ferramentas de *machine learning,* porém, com uma lógica diferente da utilizada nesse tipo de sistema. O modelo proposto deve, a nível de algoritmo, criar uma trilha de relações restritas entre contexto, usuário e demanda. O refinamento do treino, ao se utilizar dessa tecnologia, deve ser dado em ciclos de interação entre busca, atributos do contexto de uso e seleção das melhores recuperações de informação, e não pela comparação entre perfis coletivos e buscas por esses perfis. Como sugestão para continuidade da pesquisa, cabe o desenvolvimento de seu algoritmo computacional e suas relações matemáticas, bem como a operacionalização dos sistemas de inteligência artificial que dão cabo do modelo apresentado.

6. Referências

ABOWD, G.; DEY, A. (EDS.). PANEL: Towards a Better Understanding of Context and Context-Awareness. Lecture Notes in Computer Science. Anais...Atlanta, Geórgia, USA: Graphics, Visualization and Usability Center and College of Computing, Georgia Tech, jan. 1999.

AMBRÓSIO, A. P.; MORAIS, E. A. M. Mineração de Textos. Chácaras Califórnia, Goiânia, BRA: Instituto de Informática Universidade Federal de Goiás, dez. 2007.

ARAÚJO, C. A. Á. Correntes teóricas da ciência da informação. Ciência da Informação, v. 38, n. 3, p. 192–204, 2009.

BARBOSA, S. D. J.; SILVA, B. S. DA. Interação Humano-Computador. Rio de Janeiro, Rio de Janeiro, BRA: Elsevier Science, 2010.

BARLOW, M. The Culture of Big Data. Sebastopol, Califórnia, USA: O'Reilly Media, 2013.

BOYD, D.; CRAWFORD, K. Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, v. 15, n. 5, p. 662–679, maio 2012.

CHOO, C. W. Como ficamos sabendo – um modelo de uso da informação. Em: A organização do Conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. São Paulo, São Paulo, BRA: Editora SENAC, 2003. p. 61–120.

DAVENPORT, T. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Boston, Massachusetts, USA: Harvard Business Review Press, 2014.

DE MAURO, A.; GRECO, M.; GRIMALDI, M. A formal definition of Big Data based on its essential features. Library Review, v. 65, n. 3, p. 122–135, jan. 2016.

DEY, A. Providing Architectural Support for Building Context-Aware Applications. Tese—Atlanta, Geórgia, USA: Georgia Institute of Technology, nov. 2000.

GUINCHAT, C.; MENOU, M. Introdução geral às ciências e técnicas da informação e documentação. Brasília, Distrito Federal, BRA: Instituto Brasileiro de Informação em Ciência e Tecnologia, 1994.

HUVILA, I. Making and taking information. Journal of the Association for Information Science and Technology, v. 73, n. 4, p. 528–541, 2022.

INGWERSEN, P.; JÄVERLIN, K. The Turn: Integration of Information Seeking and Retrieval in Context. Berlim, Brandemburgo, GER: Springer, 2005.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 25063: Systems and software engineering — Systems and software product Quality Requirements and Evaluation (SQuaRE) — Common Industry Format (CIF) for usability: Context of use description. Genebra, CH: [s.n.].

JOHNSON, J. D. On contexts of information seeking. Information Processing and Management, v. 39, n. 1, p. 735–760, 2003.

KITCHIN, R. Big Data, new epistemologies and paradigm shifts. Big Data and Society, v. 1, n. 1, p. 1–12, 2014.

LE COADIC, Y.-F. A ciência da informação. Brasília, Distrito Federal, BRA: Briquet de Lemos Livros, 2004.

MARQUESONE, R. Big Data: Técnicas e tecnologias para extração de valor dos dados. São Paulo, São Paulo, BRA: Casa do Código, 2016.

PEIRCE, C. S. The Collected Papers of Charles Sanders Peirce. Cambridge, East of England, GBR: Harvard Business Review Press, [s.d.]. v. Vol. I-VI. C. Hartshorne et P. Weiss (eds.), Vol. VII-VIII Arthur Burks (ed.), 1931-1958, Referenciado como CP, seguido do número do volume, ponto, e número do parágrafo.

SANZ CASADO, E. Manual de estudios de usuarios. Madrid, Madrid, ESP: Fundación Germán Sánchez Ruipérez, 1994.

SARACEVIC, T. et al. A study of information seeking and retrieving. Journal of the American Society for Information Science, v. 39, n. 3, p. 161–176, maio 1988.

SEZER, O. B.; DOGDU, E.; OZBAYOGLU, A. Context-Aware Computing, Learning, and Big Data in Internet of Things: A Survey. IEEE Internet of Things Journal, v. 5, n. 1, p. 1–27, fev. 2018.

TALJA, S.; HEIDI, K.; TARJA, P. The production of ‘context’ in information seeking research: a metatheoretical view. Information Processing and Management, v. 35, n. 1, p. 751–763, 1999.