# A *web* como fonte na construção de Sistemas de Conhecimento

**Rafael Bassegio Caumo[1], Bruna Devens Fraga[2], Roberto Carlos dos Santos Pacheco[3], Denilson Sell[4]**

## *ABSTRACT*

*From the understanding of knowledge as an asset that can be made explicit and modeled, it is revealed the importance of preserving and making it available in the management processes. One way to support the management of intangible assets is to utilize knowledge bases and systems. Thus, it is necessary to adopt a knowledge engineering methodology that supports this process. In this context, knowledge needs to be discovered - if hidden - and/or acquired. Sometimes these tasks are complicated, especially when the knowledge is located externally to the organization. Nowadays, actors that use or produce external knowledge are faced with an opportunities scenario brought by the data revolution: making the web - Internet - an instant low cost source for discovery and acquisition of knowledge. Thus, the purpose here is to bring the state of the art of the bibliographic production on the use of the web as raw material for the structuring of knowledge bases and systems. As results, concepts related to web mining are presented as well as a bibliometric summary of the analyzed publications and perspectives of using web data for acquisition and discovery of knowledge - enabling the construction of knowledge bases and systems.*

***Keywords:*** *web mining; knowledge engineering; knowledge discovery; knowledge acquisition; knowledge based system.*

## RESUMO

A partir da compreensão do conhecimento enquanto ativo que pode ser explicitado e modelado, revela-se a importância de preservá-lo e torná-lo disponível nos processos de gestão. Uma forma de apoiar a gestão de ativos intangíveis consiste na utilização de bases e sistemas de conhecimento. Para tal, é necessário adotar uma metodologia de engenharia do conhecimento que auxilie este trabalho. Nesse contexto, o conhecimento deverá ser descoberto – caso esteja oculto – e/ou adquirido. Por vezes, essas tarefas se apresentam complicadas, especialmente quando o conhecimento se localiza externamente à organização. Atualmente, entretanto, atores que utilizam ou produzem conhecimento externo se deparam com um cenário de oportunidades trazido pela revolução dos dados: fazer da *web* – Internet – uma fonte instantânea e de baixo custo para descoberta e aquisição de conhecimento. Assim, o presente estudo se propõe a trazer o estado da arte da produção bibliográfica acerca da utilização da Internet como matéria prima para a estruturação de bases de conhecimento a serem incorporadas em sistemas inteligentes de apoio à tomada de decisão. Como resultados, são trazidos conceitos associados à mineração na Internet – *web mining* – assim como um resumo bibliométrico das publicações analisadas e perspectivas de utilização de dados da *web* para aquisição e descoberta de conhecimento – viabilizando a construção de bases e sistemas de conhecimento.

**Palavras-chave**: mineração na *web*; engenharia do conhecimento; descoberta de conhecimento; aquisição de conhecimento; sistema de conhecimento.

[1] Doutorando em Engenharia e Gestão do Conhecimento na Universidade Federal de Santa Catarina, Brasil. E-mail: rbcaumo@gmail.com

[2] Doutoranda em Engenharia e Gestão do Conhecimento na Universidade Federal de Santa Catarina, Brasil. E-mail: brunadefraga@gmail.com

[3] Professor do Departamento de Engenharia do Conhecimento da Universidade Federal de Santa Catarina, Brasil. E-mail: pacheco@egc.ufsc.br

[4] Professor do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina, Brasil. E-mail: denilsonsell@gmail.com

# 1 INTRODUÇÃO

Quando o conhecimento é percebido como fator de produção (Schreiber, Akkermans, Anjewierden, Hooge, Shadbolt, Van de Velde & Wielinga, 2000; O´Shea et al., 2007), insumo na cadeia de geração de valor organizacional, passa a ser visto como um ativo. E assim sendo, é importante que seja gerido de forma estratégica (Santos, 2001) para que impulsione da melhor maneira possível o sucesso dos negócios de uma organização (Davenport & Prusak, 1998).

Sob a ótica dos cognitivistas, este conhecimento não existe apenas na forma tácita, mas pode também ser explicitado. Assim, pode ser modelado e inserido em sistemas computacionais. Para Schreiber et al. (2000), por exemplo, conjuntos de dados e informações colocados em prática na realização de tarefas também são considerados conhecimento.

Assim, no contexto em que o conhecimento é ativo, fator de produção, e pode ser explicitado, modelado e inserido em sistemas de conhecimento, abre-se espaço para um processo de gestão que contemple elementos metodológicos da engenharia – Engenharia do Conhecimento – e que tem como principal produto um sistema de conhecimento – sistemas especialistas sócio técnicos que representam conhecimento e apoiam atividades intensivas em conhecimento nas organizações (Schreiber et al., 2000).

Quando da elaboração de um sistema de conhecimento, uma das etapas consiste em adquirir o conhecimento junto a uma fonte (humana ou não), transferindo certas habilidades ou perícias para dentro do sistema (Schwabe & Carvalho, 1987; Cordingley, 1989). Dependendo do negócio da organização, os conhecimentos necessários para seu sucesso podem estar mais ou menos relacionados aos ambientes interno ou externo.

Recentemente, um fenômeno tem trazido uma série de oportunidades para aqueles que buscam dados externos como matéria prima para descoberta – transformação de dados em conhecimento através da identificação de padrões compreensíveis que sejam válidos, novos e potencialmente úteis do ponto de vista prático (Fayyad, 1996) – ou aquisição de conhecimento, subsidiando tomadas de decisão – seguindo comportamentos *data-driven* – e a construção de sistemas de conhecimento: trata-se do fenômeno da revolução dos dados (Kitchin, 2014).

Inserida no contexto da nova era digital (Schmidt & Cohen, 2013), ligada à popularização da Internet, à continuidade do avanço tecnológico e à consolidação da era digital, a revolução dos dados está associada à explosão no volume, na variedade e na velocidade com que dados digitais – as vezes tratados simplesmente por "*big data*", como faz UNECE (2013) ao propor uma taxonomia de classificação – são produzidos, armazenados e disponibilizados. Esta "nova" classe de dados – os digitais – cresce na ordem de 100% ao ano (Helbing et al.,

2016), um fenômeno também chamado de "avalanche de dados" (Miller, 2010), fazendo com que a disponibilidade por dados digitais nos dias atuais supere a de dados analógicos ou mecânicos – também conhecidos por "*small data*", conforme comparação feita por Kitchin (2015). Ou seja, especificamente, a revolução dos dados é caracterizada pelo processo de transformação da predominância de disponibilidade por dados analógicos e *small* – gerados por métodos tradicionais de pesquisa – para a dos dados digitais e *big* (Kitchin, 2014).

Produto da revolução dos dados, os dados digitais derivam de registros armazenados em aplicativos e registros de ligações, mensagens e posicionamento de telefones móveis, redes sociais virtuais, páginas da Internet, mecanismos de buscas na *web*, sensores científicos, de tráfego, de segurança, medidores inteligentes, imagens de satélite, rastreamento por GPS, transações comerciais, financeiras e bancárias, registros administrativos de serviços públicos (hospitais, programas sociais, etc.), entre outros.

O potencial dos dados digitais já tem sido percebido científica e economicamente, sendo tratados como o "novo petróleo" (Cavoukian, 2010; Bossoi, 2014), uma vez que oferecem a perspectiva de análises sobre os mais diversos aspectos da vida dos indivíduos de uma população e possuem interessante relação de custo benefício no processo de construção de conhecimento a respeito de desde aspectos sociais até análises econômicas de mercado – desempenhando papel central na já mencionada quarta revolução industrial, em um período de *data economy*. Para Mayer-Schönberger & Cukier (2013), trata-se de um fenômeno que está pronto para "chacoalhar" tudo à medida que impacta na forma como o conhecimento é produzido, os negócios são conduzidos e a governança é promulgada.

Seu valor econômico já foi percebido no ambiente empresarial (Pulse, 2012) ao passo que sua relevância científica tem sido cada vez melhor compreendida pelo setor público, pela academia e pelo terceiro setor (Asquer, 2013). Os dados digitais têm aberto, portanto, uma oportunidade de inovação que pode alcançar os mais diversos setores da sociedade, gerando, por vezes, soluções radicais e até disruptivas para processos, produtos, serviços e modelos organizacionais tradicionais.

Dentro do contexto da revolução dos dados, uma das grandes fontes de dados correspondem àquelas disponíveis abertamente na Internet – *Internet as a Data Source* (EC, 2012; Askitas & Zimmermann, 2015). Este se apresenta como rica e ainda pouco explorada fonte de dados que pode subsidiar a descoberta e a aquisição de conhecimento no âmbito das mais diversas áreas do conhecimento, conforme exemplos de Beresewicz (2015). Sua mais

atraente característica, além daquelas que unificam os dados digitais, corresponde ao fato de ser gratuita e de livre acesso.

Destacam-se como oportunidades disponíveis abertamente na Internet: os conteúdos de páginas da *web*; as publicações, opiniões e comentários em redes sociais, blogs e fóruns; os logs de utilização da rede – o que foi buscado e acessado em páginas e mecanismos de buscas –; dados secundários, estatísticas públicas e socioeconômicas – produzidos tradicionalmente e disponibilizado por terceiros, tais quais organizações de estado, de pesquisa e outros produtores –; entre outras.

Nesse contexto, o presente estudo se propõe a trazer – em nível exploratório de pesquisa – o estado da arte da produção bibliográfica acerca da utilização da Internet como matéria prima para a estruturação de bases de conhecimento a serem inseridas em sistemas inteligentes de apoio à tomada de decisão. Como resultados, são trazidos conceitos de mineração na Internet – *web mining* –, resumos bibliométricos das publicações analisadas e perspectivas e possibilidades de utilização de dados da *web* para aquisição e descoberta de conhecimento, viabilizando a construção de bases e sistemas de conhecimento.

## 2 PROCEDIMENTOS METODOLÓGICOS

Este estudo se propõe a ser uma pesquisa científica com fins de construção de conhecimento no âmbito das ciências empíricas sociais, conforme definições trazidas por Gil (2008). As bases lógicas para verificação e validação científica do conhecimento construído estão apoiadas no paradigma positivista, com bases lógicas nos trabalhos dos empiristas Bacon, Hobbes, Locke e Hume durante os séculos XVI, XVII e XVIII, posteriormente formalizada por Auguste Comte no século XIX (Triviños, 1992). Dessa forma, entende-se a realidade do mundo como objetiva, composta por coisas e fatos, e o intuito está em descobrir, pelo raciocínio e a observação, as leis que regem esta realidade, testando teorias objetivas e examinando a relação entre as variáveis (Creswell, 2010; Triviños, 1992; Bryman, 2015).

Em relação ao alcance do objetivo proposto, o presente estudo almeja alcançar o nível exploratório – buscando proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito (Gil, 2008). Em termos de sua natureza técnica e operacional, trata-se de uma pesquisa aplicada, que objetiva gerar conhecimentos para o aproveitamento prático e dirigidos à solução de problemas específicos, com delineamento do tipo pesquisa bibliográfica – elaborada a partir de materiais já publicados (Gil, 2008).

O levantamento bibliográfico é realizado de forma sistemática (Forbes, 1998), partindo de uma pergunta claramente formulada e utilizando métodos sistemáticos e explícitos para: identificar, selecionar e avaliar criticamente pesquisas relevantes; e coletar e analisar dados dos estudos (Green & Higgins, 2011), com critérios para a busca sistemática que seguem o método PRISMA (Moher et al., 2009) – apresentados ao início da próxima seção.

## 3 RESULTADOS

No processo de seleção das publicações que fizeram parte desta revisão bibliográfica, o protocolo utilizado considerou a base *Scopus*, por sua interdisciplinaridade e pela reconhecida qualidade dos trabalhos, seguindo as etapas:

i. A **estratégia de busca** selecionou publicações contiveram pelo menos uma expressão de cada um dos três grupos – apresentados na sequência – em seu título, resumo ou palavras chaves. Nos resultados, nenhum filtro adicional foi aplicado. Os grupos foram:
   - Grupo 1: *web mining, mining the web, Internet mining, mining the Internet.*
   - Grupo 2: *knowledge engineering, knowledge management, knowledge acquisition, knowledge elicitation, knowledge discovery, knowledge-discovery.*
   - Grupo 3: *expert system, knowledge system, knowledge based system, knowledge-based system, knowledge base, knowledge-base, intelligent system, decision support system.*

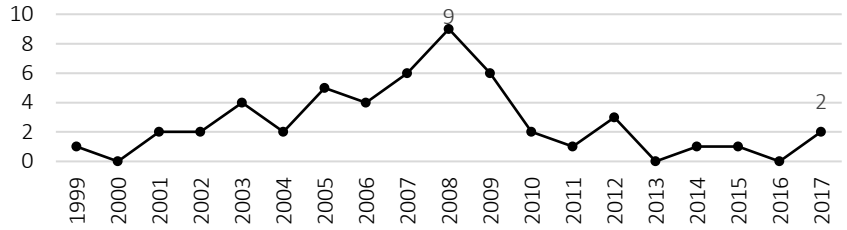
ii. A **seleção definitiva** das publicações se deu após leitura individual de todos os resumos, excluindo-se aquelas que não respeitaram o critério de elegibilidade: tratar da utilização de dados abertamente disponíveis na Internet para descoberta ou aquisição de conhecimento, ou para construção de bases ou sistemas de conhecimento. Nenhuma inclusão manual adicional de publicações foi realizada.

No processo, 65 publicações foram encontradas[5] no item i do protocolo e 14 foram excluídas durante o item ii, de modo que o quantitativo final selecionado de publicações aqui analisadas contemplou 51 itens.

[5] Em busca realizada na data de 02 de janeiro de 2018.

O Gráfico 1 apresenta a evolução do quantitativo selecionado por ano de publicação. Percebe-se que o tema – nos termos das expressões utilizadas na estratégia de busca e considerando a fonte de pesquisa utilizada – já foram de maior interesse de autores, alcançando o pico de 9 publicações no ano de 2008, reduzindo para duas publicações no ano de 2017.

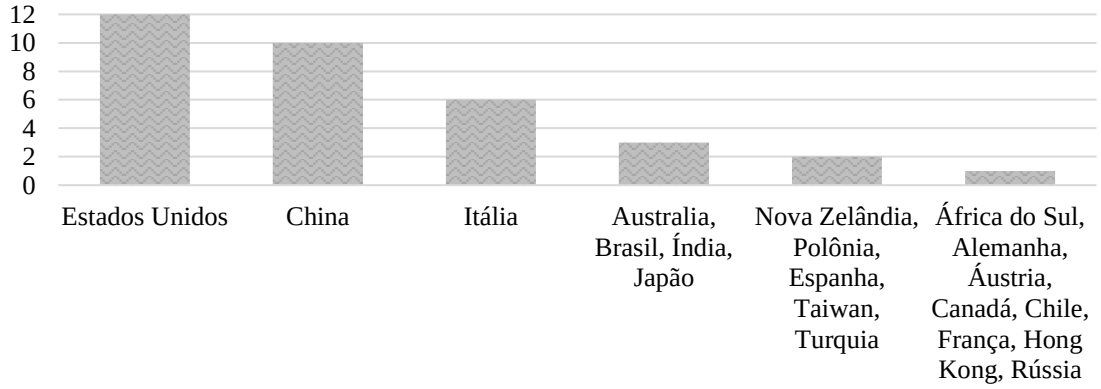
Gráfico 1: Quantitativo de publicações selecionadas, por ano de publicação.




Fonte: Elaborado pelos autores

Analisando o país de origem das publicações, três – EUA, China e Itália – são responsáveis por mais da metade do quantitativo em análise, conforme apresenta o Gráfico 2. Destaca-se aqui que a soma alcança 58 em virtude de algumas publicações serem contabilizadas para mais de um país de origem – são de coautoria de indivíduos de países diferentes.

Gráfico 2: Quantitativo de publicações selecionadas, por país de origem.




Fonte: Elaborado pelos autores

Em relação às publicações mais citadas, os resultados estão apresentados no Quadro 1. Chama atenção o fato de que dos 12 autores que apareceram com mais de uma publicação selecionada, apenas Juan D. Velásquez possui alguma publicação na relação das mais citadas.

E por fim, como uma forma de síntese do conteúdo predominante das publicações, foi construída uma nuvem de palavras a partir das palavras-chave elencadas pelos autores, apresentada na Figura 1.

Quadro 1: Relação das 15 publicações selecionadas mais citadas, identificadas pelo autor, título e ano de publicação, com correspondente quantitativo de citações.

| Autores | Título | Ano | Citações |
|---|---|---|---|
| **Roussinov, D.; Zhao, J. L.** | *Automatic discovery of similarity relationships through Web mining.* | 2003 | 61 |
| **Becerra-Fernandez, I.** | *Searching for experts on the Web: A review of contemporary expertise locator systems.* | 2006 | 44 |
| **Jicheng, W.; Yuan, H.; Gangshan, W.; Fuyan, Z.** | *Web mining: knowledge discovery on the web.* | 1999 | 24 |
| **Abraham, A.** | *i-Miner: A web usage mining framework using hierarchical intelligent systems.* | 2003 | 22 |
| **Zhang, F.; Chang, H. Y.** | *Research and development in web usage mining system-key issues and proposed solutions: A survey.* | 2002 | 21 |
| **Huang, C. C.; Tseng, T. L.; Kusiak, A.** | *XML-based modeling of corporate memory.* | 2005 | 20 |
| **Tuğ, E.; Şakiroğlu, M.; Arslan, A.** | *Automatic discovery of the sequential accesses from web log data files via a genetic algorithm.* | 2006 | 16 |
| **Dujovne, L. E.; Velásquez, J. D.** | *Design and implementation of a methodology for identifying website keyobjects.* | 2009 | 12 |
| **Yu, L.; Huang, W.; Wang, S.; Lai, K. K.** | *Web warehouse - a new web information fusion tool for web mining.* | 2008 | 11 |
| **Sánchez, D.; Moreno, A.** | *Learning medical ontologies from the web.* | 2008 | 10 |
| **Weichbroth, P.; Owoc, M.; Pleszkun, M.** | *Web user navigation patterns discovery from WWW server log files.* | 2012 | 8 |
| **Castellano, G.; Torsello, M. A.** | *Categorization of web users by fuzzy clustering.* | 2008 | 8 |

Fonte: Elaborado pelos autores

Nota: O quantitativo de citações se refere àquelas vinculadas ao portal da base Scopus.

Figura 1: Nuvem de palavras-chave mais frequentes nas publicações selecionadas.


Fonte: Elaborada pelos autores

## 1.1 MINERAÇÃO NA *WEB*: APROVEITANDO-SE DO OFERECIDO PELA INTERNET

Diante de um vasto universo de dados e informações disponíveis na *web*, surgem inúmeras possibilidades de desenvolvimento deste potencial para fins sociais, econômicos e financeiros (Zhang & Chang, 2002). Em paralelo às possibilidades, há a mineração de dados, um processo de identificação de padrões válidos, previamente desconhecidos e potencialmente úteis em dados (Liu, 2007), que vem ganhando cada vez mais espaço pela utilização de seus métodos e técnicas para fins de extração e descoberta de conhecimentos.

Para Liu (2007), a mineração de dados também é chamada de descoberta de conhecimento em bancos de dados (KDD) e pode ser definida como o processo de descoberta de padrões ou conhecimentos úteis a partir de fontes de dados, por exemplo, bancos de dados, textos, imagens, a *web*, etc. Neste sentido, a mineração de dados é usada para identificar padrões válidos, novos, potencialmente úteis e, finalmente, compreensíveis a partir da coleta de dados na comunidade de banco de dados (Jicheng et al., 1999). Trata-se de um campo multidisciplinar que envolve aprendizado de máquinas, estatística, bancos de dados, inteligência artificial, recuperação de informações e visualização (Liu, 2007).

A mineração na *web*, por sua vez, pode ser conceituada como a descoberta e análise inteligente de dados e informações úteis na *web* (Cooley et al., 1997). Assim como a mineração de dados, integra vários campos de pesquisa, como as próprias técnicas de mineração de dados adicionadas de linguística computacional, estatística, informática, entre outras áreas de conhecimento (Jicheng et al., 1999).

Cooley et al. (1997) apontam as três categorias existentes de mineração da *web*, que dependem da parte da *web* a ser explorada. A primeira categoria trata da informação contida dentro dos documentos da *web*, denominada mineração de conteúdo. A segunda está relacionada às informações entre os documentos da *web*, chamada de mineração de estrutura. E por fim, a terceira trata das informações utilizadas ou mesmo na interação com a *Web*, no qual classificam como mineração de uso.

A mineração da *web* é considerada um campo de pesquisa que atrai interesse de muitas comunidades de conhecimento (Li, Wu & Ji, 2008). E com a evolução da *World Wide Web*, a *Web* Semântica surge para oferecer uma gama de infraestruturas conectadas e interativas de informações e documentos, subsidiando a interpretação e a análise das informações disponíveis.

Na *Web* Semântica, a informação é dada com um significado bem definido, permitindo melhor interação entre os computadores e as pessoas (Berners-Lee, 2001). Desta forma, são desenvolvidos padrões tecnológicos para estabelecer uma linguagem para o compartilhamento

mais significativo de dados entre dispositivos e sistemas de informação (Souza & Alvarenga, 2004). Para Berners-Lee, Hendler & Lassila (2001) a questão não é somente facilitar as trocas de informações entre agentes pessoais de forma estática, mas trabalhar de forma cooperativa.

## 1.2 PERSPECTIVAS E USO DE DADOS DA *WEB*

Utilizando-se das publicações selecionadas como referência, parece haver indícios de que tanto a mineração de conteúdo quanto a realizada em dados de uso são bastante exploradas, sem uma grande distinção na preferência por uma ou outra, conforme sugere a Tabela 1.

Tabela 1: Distribuição relativa das publicações analisadas por tipo de dado utilizado

| **Tipo dos dados** | **Frequência relativa (%)** |
|---|---|
| Conteúdo | 50,0 |
| Uso | 33,3 |
| Conteúdo e uso | 10,4 |
| Conteúdo, uso e estrutura | 6,3 |
| Total | 100,0 |

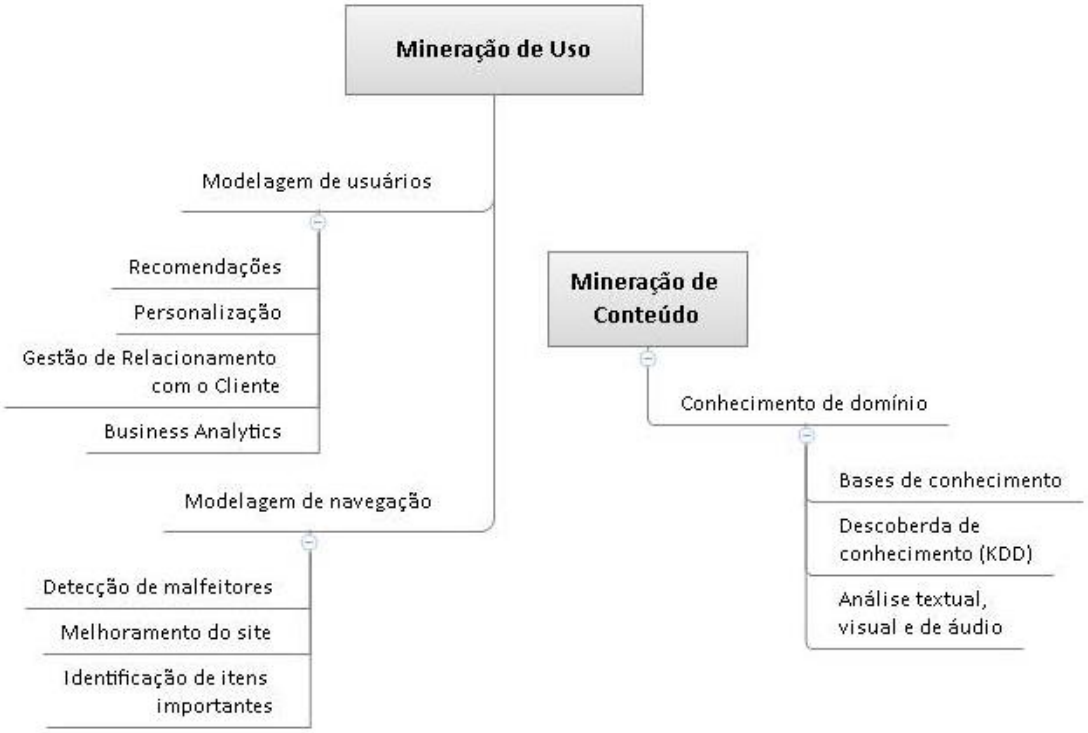
Fonte: Elaborada pelos autores

Aproximadamente 50% das publicações analisadas utilizaram-se de dados de conteúdo, enquanto aproximadamente 33% exploraram dados de uso. Estes dois tipos, trabalhados individualmente ou em conjunto, foram o alvo das aplicações em 93,7% dos trabalhos.

Em termos de aplicações práticas, os autores analisados desenvolveram soluções para as mais diversas finalidades. Algumas estão apresentadas na Figura 2, que relaciona a finalidade com os dois principais tipos de dados extraídos da Internet.

No que concerne aos dados de uso, estes serviram basicamente para observar e gerar conhecimento a partir do comportamento e das ações dos indivíduos que navegam pela Internet. Baseado nos conteúdos que acessam, seja em páginas ou em mecanismos de busca, usuários podem ter seus perfis modelados, enquanto a forma e os caminhos que percorrem em um site podem ser utilizados para modelagem da navegação (Abraham, 2003; Catellano, & Torsello, 2008; Tarakci & Cicekli, 2012). Desta forma, podem ser gerados perfis dinâmicos de usuários, gerando sistemas inteligentes de personalização e recomendação de informações na *web* (Weichbroth, Owoc & Pleszkun, 2012).

Figura 2: Relação entre tipos de dados e possíveis finalidades, conforme publicações analisadas




Fonte: Elaborada pelos autores

Ao modelar o perfil de um usuário, é possível realizar tarefas de agrupamento, segregação, categorização, classificação, entre outras, junto aos indivíduos. Assim, torna-se possível a construção de sistemas de recomendação e de estruturas que se personalizam de forma automática para cada usuário, além de permitir uma gestão de relacionamento com o cliente mais precisa e a identificação de tendências de mercado e demandas potenciais através de *Business Analytics* (Tug, Sakiroglu & Arslan, 2006; Zhou, Huang & Chen, 2008).

Quando da modelagem de navegação, esta permite, ao analisar o comportamento dos visitantes no que diz respeito à forma e aos caminhos que percorrem dentro dos sites, que sejam identificados comportamentos anômalos, servindo de indícios para a previsão e detecção de malfeitores – como aponta o trabalho de Castellano, Mastronardi, Aprile, Minardi, Catalano, Dicensi & Tarricone (2007). Além disso, permite que os objetos e links mais atraentes do site sejam identificados e viabilizam o melhoramento da eficiência das páginas e mecanismos de buscas no que concerne a layouts que tragam desempenho de tempo e precisão.

Dados de conteúdo, por sua vez, permitem predominantemente a exploração de conhecimento de domínio, com a identificação, indexação, o relacionamento e a criação de bases de conhecimento, a descoberta e a aquisição de conhecimento a partir de dados e informações e análises de documentos, páginas, instâncias, itens e conteúdo de natureza textual,

visual e de áudio. Permitindo a construção de ontologias, o resumo e a captura de contexto ou frases chaves, o estabelecimento de relacionamentos entre diferentes itens, a tradução de conteúdo, lexicalização, filtragem de e-mails, indexação e recuperação, classificação, reconhecimento de voz e imagem, entre outras tarefas (Roussinov & Zhao, 2003; Becerra-Fernandez, 2006; Choi & Huang, 2010; Da Silva & Omar, 2014).

A partir desta contextualização da finalidade dos dados utilizados na mineração da *web* (com foco em uso e conteúdo), foram analisadas as publicações que constróem um sistema de apoio a decisão, investigando-se de que forma contribuem para o contexto em que estão inseridos. Nos 51 trabalhos, 26 desenvolveram sistemas de apoio à decisão.

No que tange ao contexto organizacional, Becerra-Fernandez (2006) apresenta os sistemas SAGE e *Expert Seeker*, ambos para localização de especialistas em domínios específicos, o primeiro para universidades da Flórida e o segundo para funcionários da NASA. Outro trabalho apresenta o sistema Armazém *Web*, onde são utilizados uma série de serviços da *web*, incluindo serviço de *wrapper*, serviço de mediação, serviço de ontologia e serviço de mapeamento para auxílio na descoberta de conhecimento (Yu, Huang, Wang & Lai, 2008).

No trabalho de De Rezende, Pereira, Xexéo & De Souza (2007), os autores descrevem o sistema *Olympus*, desenvolvido para auxiliar os alunos a encontrar e compartilhar conhecimento na sua área de interesse. Este mesmo sistema classifica o indivíduo pelo conteúdo que busca e partir desta informação, recomenda cadeias/redes de conhecimento semelhantes. É importante destacar que dentre os sistemas de apoio à decisão desenvolvidos, são apontados dados de conteúdo, uso e estrutura para mineração e descoberta de conhecimento na *web*.

## 4 CONSIDERAÇÕES FINAIS

Este artigo buscou alimentar o debate sobre a utilização de dados digitais, em especial de conteúdos de páginas da Internet, como fontes para descoberta de conhecimento a ser adquirido e inserido com bases e sistemas de conhecimento. O potencial e o contexto de oportunidades que se coloca com a revolução dos dados tem sito percebidos e imagina-se que cada vez mais os dados digitais vão substituir formas mais caras e operacionalmente complicadas de gerar e capturar dados e informações.

A revolução dos dados está inserida no contexto da nova era digital, relacionada à popularização da Internet e à continuidade do avanço tecnológico, culminando em um grande volume, variedade e velocidade de dados digitais disponíveis. Desta forma, amplia-se esta

lacuna de estudo para áreas como a mineração na *web*, gerando inúmeras possibilidades de desenvolvimento deste potencial para fins sociais, econômicos e comerciais.

Entretanto, os achados deste artigo sugerem uma redução no interesse sobre o tema no âmbito acadêmico, diagnosticada pela redução do número de trabalhos na área de mineração *web* para fins de construção de bases e sistemas de conhecimento nos últimos anos.

Por outro lado, a queda na produção bibliográfica relacionada pode não corresponder à uma redução no interesse sobre o assunto, mas sim a uma questão de modificação na taxonomia utilizada para construir e referenciar os trabalhos, ou até simplesmente um mais recente aprofundamento nos métodos e técnicas que culmina em artigos que já não mencionam ou abordam o aspecto mais macro das aplicações – utilização da *web* como fonte para a descoberta e aquisição de conhecimento, subsidiando a construção de bases e sistemas de conhecimento. Isto é, pode ser que os termos aqui utilizados na busca não estejam mais funcionando de maneira adequada para encontrar artigos que estejam trabalhando dentro do tema maior de interesse.

Como exemplo, nenhum artigo que explora informações que derivam da utilização e da interação entre indivíduos em redes sociais virtuais foi encontrado a partir da busca realizada. Sabe-se, entretanto, que muitos já foram publicados e exploram tais dados para fins de construção de conhecimento. Ou seja, a busca não foi suficientemente adequada para capturar todo o escopo de utilização da *web* para os fins aqui propostos.

Quanto aos trabalhos selecionados pelas buscas, foi possível realizar a revisão sistemática destes e apontar alguns resultados a respeito à exploração de dados e informações da *web* para fins de descoberta e aquisição de conhecimento. Os autores com mais publicações na área, total de 3, foram Chen, H. e Zhou, Y. Seus trabalhos envolveram questões como o desenvolvimento de artefato para facilitar a transliteração de idiomas para descoberta de conhecimento em diferentes idiomas como chinês, árabe, inglês.

Dentre os países que mais desenvolvem trabalhos tem-se Estados Unidos, China e Itália, que somados, correspondem a mais da metade das publicações. Dentre os sistemas de apoio a tomada de decisão identificados nos trabalhos analisados, a grande maioria, 93,7%, utilizam-se de dados de uso, conteúdo, ou ambos – sendo baixíssimo o uso de dados de estrutura.

Espera-se que o estudo tenha conseguido reforçar a importância da exploração da mineração da *web* no que tange à descoberta e à aquisição de conhecimento para as mais diversas finalidades práticas, subsidiando a estruturação de bases e sistemas de conhecimento. Da mesma forma, espera-se também que possa ter contribuído para que pesquisadores em gestão e engenharia do conhecimento percebam de que forma o tema vem sendo trabalhado, permitindo que vislumbrem possibilidades de aplicações em seus domínios de interesse.

## REFERÊNCIAS

Abraham, A. (2003, May). i-miner: A *web* usage mining framework using hierarchical intelligent systems. In *Fuzzy Systems, 2003. FUZZ'03. The 12th IEEE International Conference on* (Vol. 2, pp. 1129-1134). IEEE.

Asquer, A. (2013). The governance of big data: Perspectives and issues.

Askitas, N., & Zimmermann, K. F. (2015). The Internet as a data source for advancement in social sciences. *International Journal of Manpower*, *36*(1), 2-12.

Becerra-Fernandez, I. (2006). Searching for experts on the *Web*: A review of contemporary expertise locator systems. *ACM Transactions on Internet Technology (TOIT)*, *6*(4), 333-355.

Beręsewicz, M. E. (2015). On representativeness of Internet data sources for real estate market in Poland. *Austrian Journal of Statistics*, *44*(2), 45-57.

Berners-Lee, T. et. al (2001). The semantic toolbox: building semantics on top of XML-RDF. *W3C Note*, *24*.

Berners-Lee, T.; Hendler, J.; Lassila, O (2001). The Semantic *Web*: A new form of *Web* content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, *284*(5), 34-43.

Bossoi, R. A. C.. (2014). A proteção dos dados pessoais face às novas tecnologias. Direito e novas tecnologias, Florianópolis: *CONPEDI*.

Bryman, A. (2015). *Social research methods*. Oxford university press.

Castellano, G., & Torsello, M. A. (2008). Categorization of *web* users by fuzzy clustering. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 222-229). Springer, Berlin, Heidelberg.

Castellano, M., Mastronardi, G., Aprile, A., Minardi, M., Catalano, P., Dicensi, V., & Tarricone, G. (2007). A Decision Support System base line Flexible Architecture to Intrusion Detection. *Journal of Software*, *2*(6), 30-41.

Choi, B., & Huang, X. (2010). Creating New Sentences to Summarize Documents. In *Proceedings of the 10th IASTED International Conference* (Vol. 674, No. 143, p. 458).

Cooley, R., Mobasher, B., & Srivastava, J. (1997, November). *Web mining*: Information and pattern discovery on the world wide *web*. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on* (pp. 558-567). IEEE.

Cordingley, E. S. (1989). Knowledge elicitation techniques for knowledge-based systems. In *Knowledge elicitation: principle, techniques and applications* (pp. 87-175). Springer-Verlag New York, Inc..

Creswell, J. W. (2010). Projeto de pesquisa métodos qualitativo, quantitativo e misto. In *Projeto de pesquisa métodos qualitativo, quantitativo e misto*.

da Silva, L. R., & Omar, N. (2014). Knowledge Sharing Using *web mining* for Categorization and Disambiguation of Structured and Unstructured Data. In *European Conference on*

*Knowledge Management* (Vol. 3, p. 1265). Academic Conferences International Limited.

Davenport, T. H., & Prusak, L. (1998). Conhecimento empresarial: como as empresas gerenciam seu capital intelectual. *Rio de Janeiro: Campus*.

De Rezende, J. L., Pereira, V. B., Xexéo, G., & De Souza, J. M. (2007). Olympus: personal knowledge recommendation using agents, ontologies and *web mining*. In *International Conference on Computer Supported Cooperative Work in Design* (pp. 53-62). Springer, Berlin, Heidelberg.

EC (European Commission). (2012). Internet as data source - Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering. Disponível em: https://www.dialogic.nl/file/2016/12/2010.080-1226.pdf. Acesso em: 15 dez 2017.

Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert: Intelligent Systems and Their Applications*, *11*(5), 20-25.

Forbes, D. A. (1998). Strategies for Managing the Behavioural Symptomatology Associated with Dementia of the Alzheimer Type: A Systematic Overview. *Canadian Journal of Nursing Research Archive*, *30*(2).

Gil, A. C. (2008). *Métodos e técnicas de pesquisa social*. 6. ed. Ediitora Atlas SA.

Higgins, J. P., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4). John Wiley & Sons.

Helbing, D.; Frey, B. S.; Gigerenzer, G.; Hafen, E.; Hagner, M.; Hofstetter, Y.; Van Den Hoven, J.; Zicari, R.; Zwitter, A. (2016). Behavioural Control or Digital Democracy? - *A Digital Manifesto*.

Jicheng, W., Yuan, H., Gangshan, W., & Fuyan, Z. (1999). *Web mining*: knowledge discovery on the *Web*. In *Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on* (Vol. 2, pp. 137-141). IEEE.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS*, *31*(3), 471-481.

Li, H., Wu, Z., & Ji, X. (2008). Research on the techniques for Effectively Searching and Retrieving Information from Internet. In *Electronic Commerce and Security, 2008 International Symposium on* (pp. 99-102). IEEE.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Mayer-Schonberger, V.; Cukier, K.(2013*). Big data:* A Revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging?. *Journal of Regional Science*, *50*(1), 181-201.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Reprint—preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Physical therapy*, *89*(9), 873-880.

O'Shea, R. P., Allen, T. J., Morse, K. P., O'Gorman, C., & Roche, F. (2007). Delineating the anatomy of an entrepreneurial university: the Massachusetts Institute of Technology experience. *R&d Management*, *37*(1), 1-16.

Roussinov, D., & Zhao, J. L. (2003). Automatic discovery of similarity relationships through *Web mining*. *Decision Support Systems*, *35*(1), 149-166.

Santos, A. R. dos. (2001). *Gestão do conhecimento: uma experiência para o sucesso empresarial.* Universitária Champagnat.

Schmidt, E.; Cohen, J. (2013). *A nova era digital: como será o futuro das pessoas, das nações e dos negócios*. Intrínseca.

Schreiber, A. Th.; Akkermans, J. M.; Anjewierden, A.A.; Hoog, R. de; Shadbolt, N. R.; Van de Velde, W.; Wielinga, B. J. (2000). *Knowledge engineering and management: the CommonKADS methodology*. MIT press, 471 p.

Schwabe, D., & Carvalho, R. D. (1987). Engenharia de conhecimento e sistemas especialistas. *Buenos Aires: Kapelusz*.

Souza, R. R., & Alvarenga, L. (2004). A *Web* Semântica e suas contribuições para a ciência da informação. *Ciência da Informação*, *33*(1).

Tarakçi, H., & Cicekli, N. K. (2012). UCASFUM: A Ubiquitous Context-aware Semantic Fuzzy User Modeling System. In *KEOD* (pp. 278-283).

Triviños, A. N. S. (1992). *Introdução à pesquisa em ciências sociais: a pesquisa qualitativa em educação*. Atlas.

Tuğ, E., Şakiroğlu, M., & Arslan, A. (2006). Automatic discovery of the sequential accesses from *web* log data files via a genetic algorithm. *Knowledge-Based Systems*, *19*(3), 180-186.

UNECE. (2013). Classification of Types of Big Data. Disponível em: https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data.

Pulse, U. G. (2012). Big data for development: Challenges & opportunities. *Naciones Unidas, Nueva York, mayo*.

Cavoukian, A. (2013). Personal Data Ecosystem (PDE)–A Privacy by Design Approach to an Individual's Pursuit of Radical Control. *Digital Enlightenment Yearbook 2013: The Value of Personal Data*, 89-101.

Weichbroth, P., Owoc, M., & Pleszkun, M. (2012). *Web* user navigation patterns discovery from WWW server log files. In *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on* (pp. 1171-1176). IEEE.

Yu, L., Huang, W., Wang, S., & Lai, K. K. (2008). *Web* warehouse–a new *web* information fusion tool for *web mining*. *Information Fusion*, *9*(4), 501-511.

Zhang, F., & Chang, H. Y. (2002). Research and development in *web* usage mining system-key issues and proposed solutions: a survey. In *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on* (Vol. 2, pp. 986-990). IEEE.

Zhou, Y., Huang, F., & Chen, H. (2008). Combining probability models and *web mining* models: a framework for proper name transliteration. *Information Technology and Management*, *9*(2), 91-103.