A TEORIA DO BIG DATA

Patrícia de ALMEIDA1

1Universidadade de Coimbra, Coimbra (Portugal)

Resumo

O desafio da contemporaneidade de encontrar a individualidade num mundo massificado é equiparado ao desafio para os profissionais da informação de (fazer) encontrar a singularidade num mundo de Big Data. O crescimento informacional tem estado na origem de grande volume e variedade de dados em diferentes domínios sociais, pelo que novas perspetivas estão a ser necessárias para a representação, a organização e a recuperação da informação. A teoria que suporta o momento Big Data é variada, no entanto uma tem merecido destaque na literatura, enquanto modelo para a estruturação e localização da informação: a teoria facetada. Neste contexto, tem-se como objetivo verificar se a teoria facetada é eficaz na recuperação veloz de informação verdadeira e com valor. Para tal, propõe-se um caso prático, concretamente a consulta de duas páginas de comércio eletrónico, uma nacional e outra internacional, em busca de um determinado item. Os resultados mostram que os sistemas facetados trazem bastante exatidão e validade, em tempo quase imediato, mesmo em grandes e variados volumes de dados. Conclui-se que a teoria facetada é (a) adequada ao momento Big Data.

Palavras-chave: Big Data; Teoria facetada; Ranganathan; Recuperação da informação.

Abstract

The contemporary challenge of finding individuality in a mass world is compared with the challenge for information professionals to find the uniqueness in a Big Data world. Information growth has been the source of a large volume and variety of data in different social domains, therefore new perspectives are needed to represent, organize and retrieve the information. The theory that supports the Big Data moment is diverse, however, one has deserved special importance in the literature, as a model for structuring and locating information: the faceted theory. In this context, the objective is to verify if the faceted theory is effective in the rapid recovery of valuable and true information. For this, a practical case is proposed, namely the check of two sites of electronic commerce, one national and another international, in search of a certain item. The results show that faceted systems bring a lot of accuracy and validity, with almost immediate speed, even in a large and varied data volume. It is concluded that the faceted theory is (the one) appropriate to the Big Data moment.

Keywords: Big Data; Faceted theory; Ranganathan; Information retrieval.

1 INTRODUÇÃO

Se um dos grandes desafios da contemporaneidade é encontrar a individualidade num mundo massificado, então, poder-se-á dizer que o desafio para os profissionais da informação é (fazer) encontrar a singularidade num mundo de Big Data. Esta expressão é já banal para a descrição de quaisquer conjuntos de dados que, excedendo as capacidades dos sistemas informáticos tradicionais, não possam ser processados sem recurso a estruturas de computação específicas (Souza, Almeida & Baracho, 2015). É real o sempre crescente e infinito universo de dados, pois, se até 2003, se gerou cinco exabytes (10 bytes elevados a 18ª potência) de dados, esse mesmo volume é criado a cada dois dias e prevê-se que o armazenamento de dados na internet duplique a cada dois anos (Victorino, Shiessl, Oliveira, Ishikawa, Holanda & Hokama,

2017). Neste contexto de enorme volume e variedade, obter rapidamente aquela informação exata e singular é, realmente, um privilégio nos tempos modernos. Seja qual for a área de ação humana, procura-se valor, veracidade e velocidade na recuperação da informação.

O crescimento informacional originou Big Data em diferentes domínios sociais e novas perspetivas estão a ser necessárias para a representação, a organização e a recuperação de dados, de que trata

a Ciência da Informação. Esta é, verdadeiramente, uma ciência dinâmica e desafiante, com objetos de estudo e trabalho fluídos e ubíquos (Souza, Almeida & Baracho, 2015). Os Big Data apresentam-se como mais um desafio, uma permanente provocação e um verdadeiro estímulo para os profissionais da informação. A expansão das tecnologias nos séculos XX e XXI veio acicatar esta questão em diferentes domínios de intervenção – criação, representação, armazenamento, organização e consumo da informação –, pelo que o trabalho destes profissionais é cada vez mais importante. Assim, falar de Big Data é lembrar que nunca como hoje se precisou tanto de uma Ciência da Informação, especificamente dos seus modelos teóricos e práticos para a arquitetura da informação. Muito embora este conceito seja um tanto ou quanto ambíguo na literatura, poder-se-á entendê-lo simplesmente como um guia para estruturar e localizar qualquer tipo de informação (Victorino et al., 2017).

O enquadramento teórico que suporta o trabalho com Big Data é variado, no entanto uma teoria em especial tem merecido destaque, enquanto modelo para a estruturação e localização da informação: a teoria facetada. Ela advém dos sistemas de classificação bibliográfica, concretamente dos que se baseiam na análise facetada dos documentos. É a Shiyali Ramamrita Ranganathan (1892-1972), considerado o pai da Biblioteconomia na Índia, que geralmente se atribuem os méritos pela introdução da análise do conhecimento em facetas, concretizada na sua Colon Classification (Classificação Dois Pontos) e nos seus escritos teóricos (Ranganathan, 1933 e 1937). Garfield (1984) afirma que Ranganathan foi o primeiro a explicar completamente a teoria facetada, mas que as ideias de Sayers, Bliss e Richardson contribuíram para isso. Também Broughton (2006) aponta que vários teóricos anteriores a Ranganathan adotaram conceitos similares, ainda que de forma mais limitada. Muito embora tal seja verdade, os principais marcos teóricos da análise facetada são, sem dúvida, os trabalhos de Ranganathan e, posteriormente, os estudos do Research Classification Group, fundado em Inglaterra nos anos 50 do século XX, com o objetivo de desenvolver estudos teóricos e práticos no âmbito da classificação (Lima, 2004).

Os sistemas de classificação facetados dividem o assunto em propriedades homogéneos, isto é, em categorias e em facetas que lhe são inerentes, apresentando várias vantagens no domínio da organização do conhecimento. Num momento de análise, eles permitem o reconhecimento de vários aspetos num único assunto e, num momento de síntese, tentam sintetizar esses mesmos aspetos de maneira a melhor descrever o assunto, aclarando a multidimensionalidade e os diversos rumos que o conhecimento pode tomar (Lima, 2004). Entre as muitas vantagens que enumera, Broughton (2006) refere o facto de os sistemas facetados fornecerem ferramentas para visualizar e para pesquisar em tópicos do assunto, sendo um importante método para a organização e a apresentação dos conteúdos, bem como para a navegação em

ambiente web. A estrutura lógica e previsível dos sistemas facetados torna-os compatíveis com ou adaptáveis aos requisitos dos programas informáticos.

Assim, a teoria e os modelos facetados mostram-se adequados para serem aplicados na descrição, navegação e recuperação de informação em ambiente digital e vários são os investigadores que explicitamente o têm referido ao longo dos anos. Vickery (1965) vê a metodologia de Ranganathan como modelo para todos aqueles que trabalham com sistemas mecanizados. Kashyap (2001) declara que a técnica da análise facetada é, em alguns aspetos, superior a outras técnicas de recuperação de informação, nomeadamente em bases de dados on-line de sistemas de informação. Para Hudon (2006), os princípios facetados mostram vantagens não só para a representação do assunto, mas para a sua recuperação e aponta a organização dos documentos administrativos nas unidades governamentais do Quebec como um bom exemplo. Duarte e Cerqueira (2007) afirmam que o modelo facetado se apresenta como uma ferramenta auxiliar na representação de conceitos ideais em sistemas de hipertexto.

Mais recentemente, Castro, Cruz e Oddone (2013) defendem que a teoria de Ranganathan tem influência na modelagem dos sistemas de informação e que os desenvolvimentos informáticos recentes são subsidiários de uma realização manual. Gomes (2017) especifica que, em Prolegomena to Library Classification (Ranganathan, 1937), se encontram os elementos necessários para a organização de taxonomias e interfaces de navegação e que o método facetado é quase uma unanimidade na organização semântica da web. Satija (2017) enumera várias investigações que mostram que motores de busca e diretórios usam a abordagem da Ranganathan com bons resultados na recuperação da informação e comenta que é quase como se Ranganathan tivesse antecipado os ambientes web, uma vez que muito do que lá se passa depende de uma análise facetada.

De acordo com a literatura, o referencial teórico da análise facetada é, então, utilizado em contexto informatizado e apresenta potencial para fazer face aos atuais desafios da Ciência da Informação, perante o aumento informacional. Porém, será a teoria facetada uma resposta eficaz à real demanda por soluções efetivas (Victorino et al., 2017)? Será a teoria facetada a teoria para o momento Big Data?

Neste enquadramento e com estas interrogações, surge um estudo que coloca o enfoque desta teoria não no armazenamento e processamento da informação, mas na sua recuperação para consumo. Tem-se como objetivo verificar se a teoria facetada, perante grande volume e variedade de dados, é eficaz na recuperação veloz de informação verdadeira e com valor (isto é, aquela exata e singular). A resposta às questões irá ao encontro do que todos anseiam na atualidade, sejam eles profissionais da informação ou vulgares cidadãos.

2 METODOLOGIA

Para dar cumprimento ao objetivo do estudo, propõe-se um caso prático, especificamente a consulta a plataformas de comércio eletrónico, em busca de um determinado item da indumentária feminina. As páginas a servir de amostra foram selecionadas tendo em conta o grande volume e variedade de dados que comportam e sem qualquer intuito publicitário.

Trata-se de uma amostra por conveniência, uma vez que estas plataformas permitem a modelação da informação disponibilizada também em categorias e facetas e o que se pretende verificar é exatamente a eficácia do modelo facetado. Para se poder obter dados diversificados, optou-se por uma marca e página nacional, designadamente a Lanidor - https://www.lanidor.com/, e outra internacional, a saber Aliexpress - https://pt.aliexpress.com/ (na variante brasileira da língua portuguesa, por inexistência da variante europeia).

Apesar da conveniência, procuraram-se plataformas com grande volume e variedade de dados. A Lanidor foi fundada em 1966 e apresenta-se como a maior marca portuguesa de pronto-a-vestir feminino, com uma rede de 98 lojas, espalhadas por nove países[1]. A Aliexpress pertence ao grupo Alibaba, conhecido como o "gigante chinês do comércio eletrónico", com origem em 1999 e com um volume de negócios crescente e superior a empresas congéneres[2]. Em ambas, selecionou-se como objeto de busca de indumentária feminina "calças azuis". O item foi determinado pela sua real existência e vulgaridade (calças de senhora são bastante comuns e a cor é muito usual nesta peça de roupa), de forma a dificultar o alcance de recuperação da informação singular desejada.

Efetuaram-se duas pesquisas em cada uma das plataformas, em conformidade com o objeto de busca selecionado: uma primeira pesquisa em motor de busca geral e uma segunda direcionada pelos mecanismos de afunilamento da informação disponibilizados, designados por "filtro". Aqui, em cada uma das categorias, formam escolhidas as opções (facetas) consideradas mais vulgares, mais uma vez de forma a dificultar a recuperação de uma única e exata informação. Os resultados obtidos em ambas as páginas são explanados e comparados, por método e por plataforma comercial.

O caso prático realizou-se no dia 18 de fevereiro de 2018 e acredita-se que a data não terá uma influência considerável nos resultados e nas conclusões deste estudo, muito embora a constante atualização das bases de dados das páginas em análise ocasione resultados distintos em diferente período temporal e tal seja uma limitação do trabalho aqui desenvolvido.

3 RESULTADOS

Plataforma Lanidor

• Pesquisa 1 (em motor de busca)

calças azuis = 0 resultados

[1] 1 Ver em: https://www.lanidor.com/html/info/new/ajuda_16_pt.html

[2] Ver em: https://exame.abril.com.br/negocios/lucro-da-alibaba-cresce-35-no-3o-tri-e-empresa-revisa-previsoes/

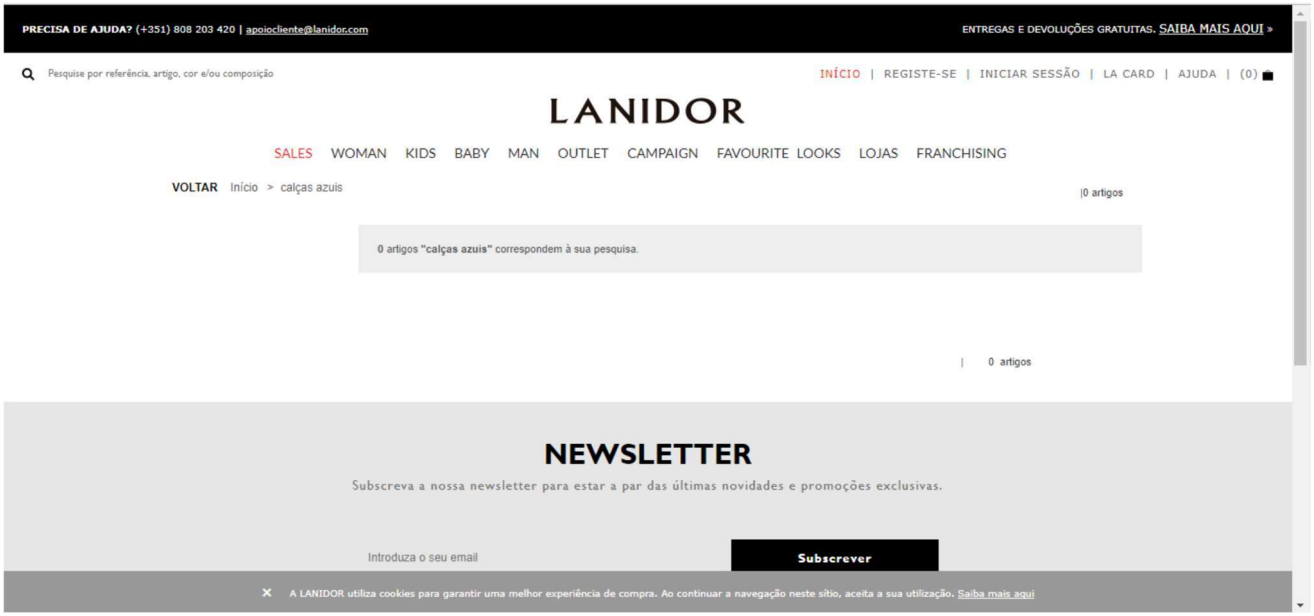

Figura 1. Resultado da pesquisa 1 em Lanidor

• Pesquisa 2 (com recurso a filtro)

Seleção da Categoria Woman > Calças = 11 resultados

Filtro: Categorias Características (Calças), Cor (Azul), Tamanho (M 36) e Ordenar (Preço Ascendente) = 1 resultado

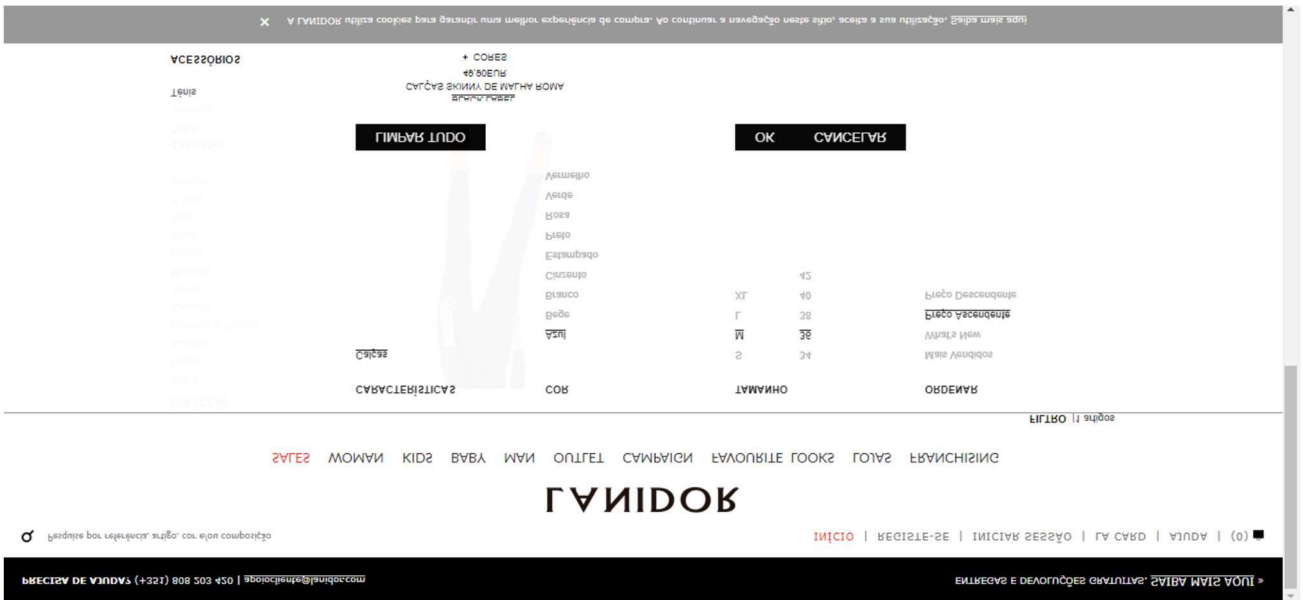

Figura 2. Filtro de Lanidor

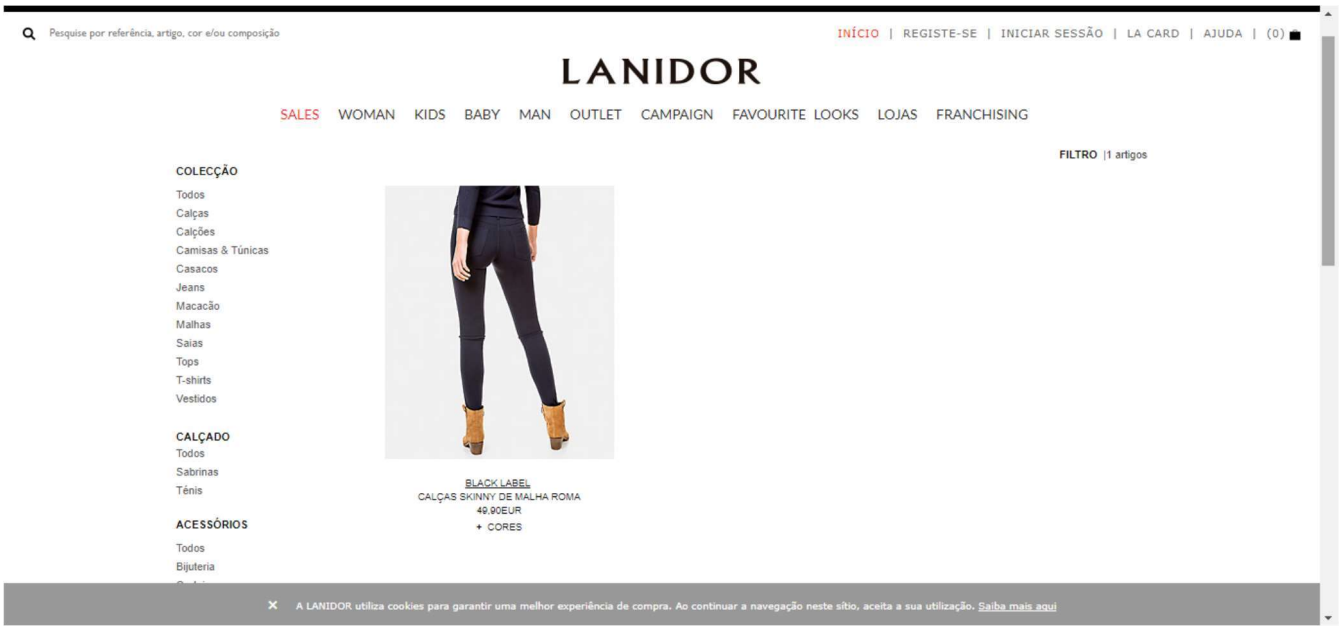

Figura 3. Resultados da pesquisa 2 em Lanidor

Plataforma Aliexpress

• Pesquisa 1 (em motor de busca)

calças azuis = 14 562 resultados

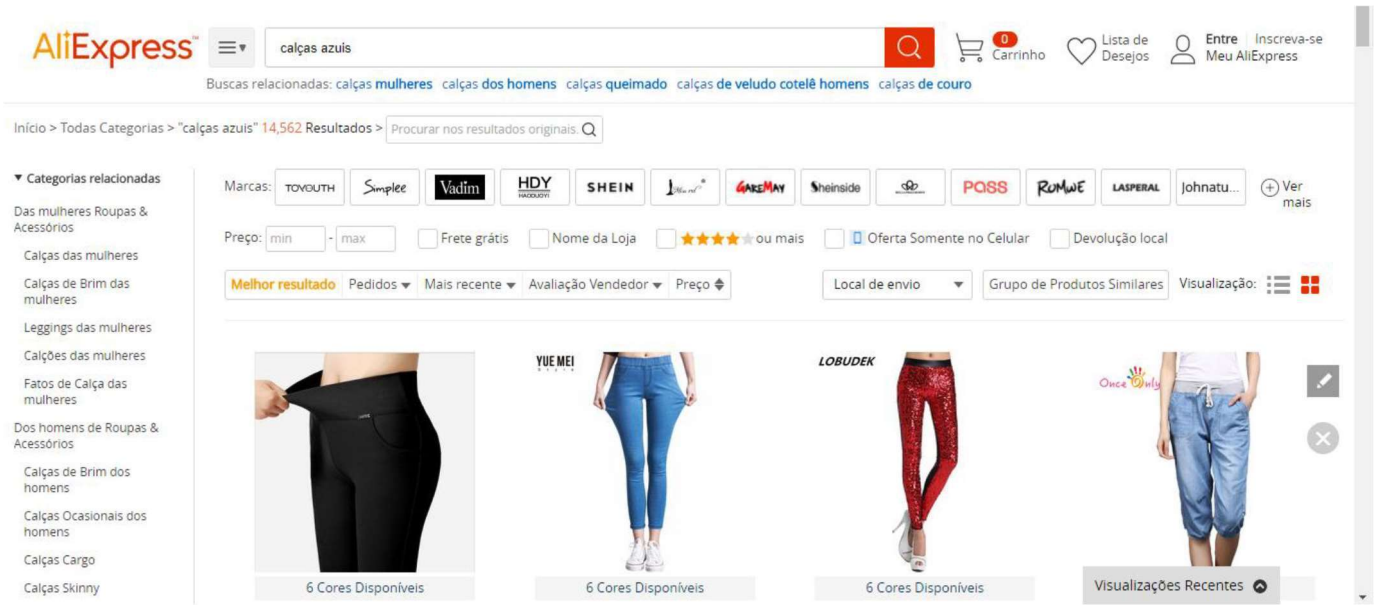

Figura 4. Resultado da pesquisa 1 em Aliexpress

• Pesquisa 2 (com recurso a filtro)

Seleção da Categoria Moda Feminina > calças e capris = 92 289 resultados

Filtro vertical: Categorias Material (algodão), Estilo (casual), Tipo de estampa (sólida), Tecido (sarja), Acabamento/Decoração (bolsos), Ajuste (normal), Cintura (média) e Tipo de fecho (zíper) = 15 resultados

Mais refinamento: Tamanho (M) e Cor (azul) = 6 resultados

Filtro horizontal3: Categorias Preço (7,66€-11,72€4) = 2 resultados

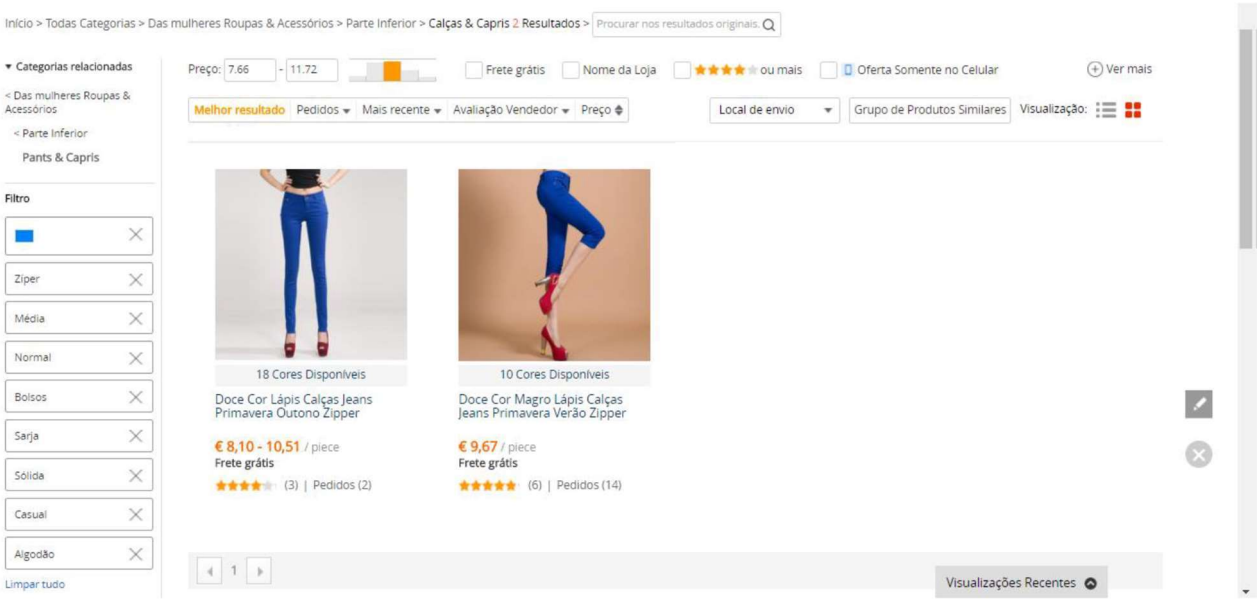

Figura 5. Filtro e resultados da pesquisa 2 em Aliexpress

Através desta metodologia, apurou-se que, em ambas as plataformas comerciais, a pesquisa por motor de busca existente não é eficaz, pois, se no caso nacional não se recupera qualquer informação (apesar de o item procurado estar efetivamente lá), no caso internacional o volume de informação obtida continua bastante significativo e extremamente longe da singularidade pretendida. Já na pesquisa com recurso ao filtro, em ambos os casos, a informação recuperada é singular (Lanidor – 1 resultado) ou está muito perto disso (Aliexpress – 2 resultados), o que indica valor para um possível consumidor.

A informação recuperada com recurso a filtro nas duas plataformas também pode ser considerada exata, portanto verdadeira, uma vez que os resultados obtidos representam efetivamente calças azuis. Regista-se, todavia, uma variante: no caso da página Aliexpress, as imagens mostram que se obteve um resultado com calças e outro com capri (calças curtas). Logo, se existisse uma categoria para filtrar o comprimento do objeto de pesquisa, obter-se-ia um resultado único,

por conseguinte ainda mais exato. O mesmo poderia acontecer, caso se incluíssem facetas relativas à tonalidade da cor pretendida, por exemplo, azul claro ou azul escuro.

O tempo para a obtenção de resultados foi quase imediato para todas as pesquisas, em qualquer uma das plataformas. Assim sendo, embora difiram no volume e na variedade dos dados, não existem diferenças consideráveis entre as páginas Lanidor e Aliexpress, no que toca ao valor, à veracidade e à velocidade da recuperação da informação.

Comparativamente com um motor de busca, verificou-se a eficácia de uma análise facetada na recuperação de informação. Posto isto, determina-se que uma arquitetura de sistema que recorra à teoria facetada se revela altamente eficaz para a recuperação da informação, mesmo tratando-se de ambientes Big Data.

3 Não são selecionadas as categorias que não se prendem com as características efetivas do objeto de busca; frete grátis, avaliação com 4 estrelas ou mais, oferta somente no celular, nome de loja, avaliação vendedor… referem-se ao envio ou ao consumo e não ao item de indumentária.

4 Intervalo de preço referido como mais comum pela própria página.

4 CONCLUSÕES

Num contexto de Big Data com grande volume e variedade de informação e em que se deseja recuperar informação com valor, veracidade e velocidade, testam-se duas páginas de comércio eletrónico, através de um caso prático. Concluiu-se que, em acordo com a literatura da área, a teoria facetada tem aplicação informática efetiva e é utilizada em modelos reais, nomeadamente nas plataformas nacional Lanidor e internacional Aliexpress.

Os resultados mostram que os sistemas facetados trazem bastante eficácia na recuperação da informação, mostrando exatidão, singularidade e rapidez quase imediata, mesmo em grandes e diversificados volumes de dados. Conclui-se que quanto maior o número de categorias e de facetas, maior será a eficácia do sistema. Considera-se até que, mediante a quantidade de categorias do filtro apresentado, os sistemas facetados podem funcionar como instrumentos cognitivos orientadores para as possíveis opções de compra pelos cidadãos, na medida em que os guiam e mostram um caminho na multiplicidade e multidimensionalidade da informação.

Desta feita, muito embora se trate de um pequeno caso prático e com limitações de ordem quantitativa (quatro pesquisas, duas plataformas em análise e um item de busca), concluiu-se que a teoria facetada é (a) adequada ao momento Big Data.

REFERÊNCIAS

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings New Information Perspectives, 58(1/2), pp. 49-72. Retrieved from https://doi.org/10.1108/00012530610648671

Castro, F., Cruz, F., & Oddone, N. (2013). O paradigma da orientação a objetos, a linguagem unificada de modelagem (uml) e a organização e representação do conhecimento: um estudo de caso de um sistema para bibliotecas. Informação & Informação, 18(1), pp. 82-105. Retrieved from https://doi.org/10.5433/1981-8920.2013v18n1p82

Duarte, E. A., & Cerqueira, R. (2007). Análise Facetada: Um olhar Face a Modelagem conceitual. Revista Digital de Biblioteconomia e Ciência da Informação, 4(2), pp. 39-52. Retrieved from https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/2020

Garfield, E. (1984). A Tribute to S. R. Ranganathan, the Father of Indian Library Science. Part 1. Life and Works. Essays of an Information Scientist, 7, pp. 37-44. Retrieved from http://garfield.library.upenn.edu/essays/v7p037y1984.pdf

Gomes, H. E. (2017). Marcos históricos e teóricos da organização do conhecimento. Informação & Informação, 22(2), pp. 33-66. Retrieved from https://doi.org/10.5433/1981-8920.2017v22n2p33

Hudon, M. (2006). Le passage au XXIe siècle des grandes classifications documentaires. Documentation et Bibliothèques, 52(2), pp. 85-97. Retrieved from https://www.erudit.org/fr/revues/documentation/2006-v52-n2-documentation01812/1030012ar/

Kashyap, M. M. (2001). Similarity Between the Ranganathan's Postulates for Designing a Scheme for Library Classification and Peter Pin-Sen Chen's Entity Relationship Approach to Data Modelling and Analysis. DESIDOC Bulletin of Information Technology, 21(3), pp. 3-16. Retrieved from http://publications.drdo.gov.in/gsdl/collect/dbit/index/assoc/HASH7eef.dir/dbit2103003.pdf

Lima, G. (2004). O modelo simplificado para análise facetada de Spiteri a partir de Ranganathan e do Classification Research Group (CRG). Información, cultura y sociedad, 11, pp. 57-72. Retrieved from http://www.scielo.org.ar/pdf/ics/n11/n11a03.pdf

Ranganathan, S. R. (1933). Colon Classification. [First edition]. Madras: Madras Library Association.

Ranganathan, S. R. (1937). Prolegomena to Library Classification. Madras: Madras Library Association.

Satija, M. P. (2017). Colon Classification (CC). Knowledge Organization 44(4), pp. 291-307. Retrieved from http://www.isko.org/cyclo/colon_classification#ref

Souza, R., Almeida, M. & Baracho, R. (2015). Ciência da informação em transformação: Big Data, nuvens, redes sociais e Web Semântica. Ciência da Informação, 42(2), pp. 159-173. Retrieved from http://revista.ibict.br/ciinf/article/view/1379

Vickery, B. C. (1965). Ranganathan's work on classification. Library Science Today, 1. Kaula, P. N. (Ed.). Bombay: Asia Publishing House.

Victorino, M., Shiessl, M., Oliveira, E., Ishikawa, E., Holanda, M. & Hokama, M. (2017). Uma proposta de ecossistema de Big Data para a análise de dados abertos governamentais conectados. Informação & Sociedade: Estudos, 27(1), pp. 225-242. Retrieved from http://www.ies.ufpb.br/ojs/index.php/ies/article/view/29299