

ANÁLISE DE CONTEÚDO COM RECURSO A SOFTWARE

vs.

REVISÃO DE LITERATURA

ESTUDO COMPARATIVO DA RECUPERAÇÃO DA INFORMAÇÃO

Ana Gouveia Coelho¹

¹Universidade de Coimbra, uc1996021922@student.uc.pt, ORCID iD 0000-0001-8520-2522

Resumo:

A análise de conteúdo com auxílio de software é um método qualitativo que utiliza dados textuais transcritos como insumo da compreensão de determinado fenómeno ou contexto através da identificação das principais variáveis que o descrevem. Pretende-se comparar esse método face à revisão de literatura, no que respeita à recuperação da informação. Para o efeito, utilizou-se um método de tipo exploratório e comparativo. Procedeu-se à construção de um *corpus* com base na recolha de documentos sobre o tema “acesso aberto”, no RCAAP, no período 2016-2020. Com auxílio do software IRaMuTeQ encetou-se uma análise de similitude e construção do respetivo grafo de coocorrências, bem como uma análise fatorial de correspondência. Procedeu-se de seguida à revisão de literatura sobre o tema, utilizando a mesma fonte de informação. Compararam-se os resultados da recuperação da informação usando os dois métodos. Confirmou-se a efetividade dos dois métodos no que toca à recuperação da informação, sendo que, a análise de conteúdo com recurso a software, foi mesmo mais além em algumas situações, mas não ao ponto de excluir a revisão de literatura.

Palavras-chave: Análise de conteúdo, revisão de literatura, recuperação da informação, comparação, Iramuteq

Introdução

Conhecer o léxico de um domínio é o primeiro passo para o conhecimento desse domínio. Segundo Brandt & Brasher (2010) a organização do conhecimento “está relacionada com a análise conceitual de um domínio do conhecimento e a partir daí, a sua estruturação (a partir dos termos que o compõem), gerando uma representação de tal domínio”. Existem várias formas de estruturar (ainda que apenas mentalmente) o léxico de um domínio para recuperar a informação.

A mais utilizada é a revisão de literatura. Este método que não requer nenhum aparato especial, embora, para que seja bem constituída deva obedecer a um plano, consubstanciado em várias fases interligadas. (Machi & Mc Evoy, 2012). Independentemente de ser bem ou menos bem realizada, a informação, veiculada pelos sentidos, será sempre “filtrada” pela parte cognitiva do leitor.

Outra forma, consiste na análise de conteúdo com recurso a software. Esta utiliza várias técnicas automáticas para a descoberta de informação latente em dados textuais. Os dados tornam-se perceptíveis ao ponto de se transformar em informação a ser recuperada. Diversos autores: Camargo & Justo (2013) Hoffman, Bisset & Martim-Lahera, (2020); Vaynuskin & Grashenko, (2020); Ramos, Valderez & Amaral-Rosa, (2018); Pinto, Silva & Sena, (2010) consideram que os softwares, alguns *open-source*, como o IRaMuteQ¹, são ferramentas capazes para o processamento da linguagem natural transcrita de

¹ IRaMuTeQ - *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. Foi criado em 2009 por Pierre Ratinaud. Acresce que se trata dum software gratuito, de código fonte aberto que utiliza o ambiente estatístico do R, também *open-source*. Incorpora o mesmo algoritmo do Alceste©. Permite ao utilizador a análise de conteúdo (similitude), classificação hierárquica descendente, análise fatorial de correspondência, nuvem de palavras para além de outras ferramentas estatísticas, num leque bastante completo (Saviati, 2017)

forma correta, mormente, através do cálculo de um grafo de coocorrências ou outras ferramentas de índole estatística. Também, segundo Cervi (2018), tem a qualidade de ser menos subjetiva do que a produzida na mente do leitor, pela revisão de literatura, que precisa controlar os efeitos da interpretação pessoal nos significados apreendidos.

Dado o campo promissor oferecido pela análise automática de conteúdo, firmada em opiniões de diversos autores e também na experiência pessoal da própria autora, interroga-se se este método poderá mesmo substituir a revisão de literatura. Desta forma, presidiu à presente investigação a seguinte questão orientadora:

“Quais as diferenças entre o uso do método de análise de conteúdo com software face à revisão de literatura no que toca à recuperação de informação?”

Face à interrogação apresentada, decidiu-se empreender uma análise exploratória e comparativa que permitisse confrontar os dois métodos, aplicando-os a um caso prático. Nesse âmbito, foi escolhido o domínio do acesso aberto, conforme documentos recolhidos no RCAAP, no intervalo de tempo 2016-2020.

Objetivos

Como objetivo geral, comparar os resultados da recuperação da informação, utilizando os métodos de análise de conteúdo com software face à revisão de literatura, aplicados num domínio e recorte espaço-temporal: Acesso Aberto, RCAAP, 2016-2020.

Em termos específicos pretende-se descrever as vantagens/desvantagens de cada um dos métodos no que toca à recuperação da informação.

Método

Utilizou-se um método exploratório e comparativo. Foi construído um corpus resultante da recolha no RCAAP, 2016-2020, de documentos acerca do “acesso aberto”. Introduziu-se no campo “título” o termo “acesso aberto” (com aspas). A pesquisa resultou em 72 observações. Para a constituição do *corpus* considerou-se, o título, o resumo e as palavras-chave de cada documento. O conjunto foi operacionalizado em *grupos* de forma a ser compreendido pelo Iramuteq. Procedeu-se à análise de conteúdo, através do cálculo da similitude e análise fatorial de correspondência. Depois de sistematizar as inferências daí resultantes, empreendeu-se a revisão de literatura de forma a comparar os resultados.

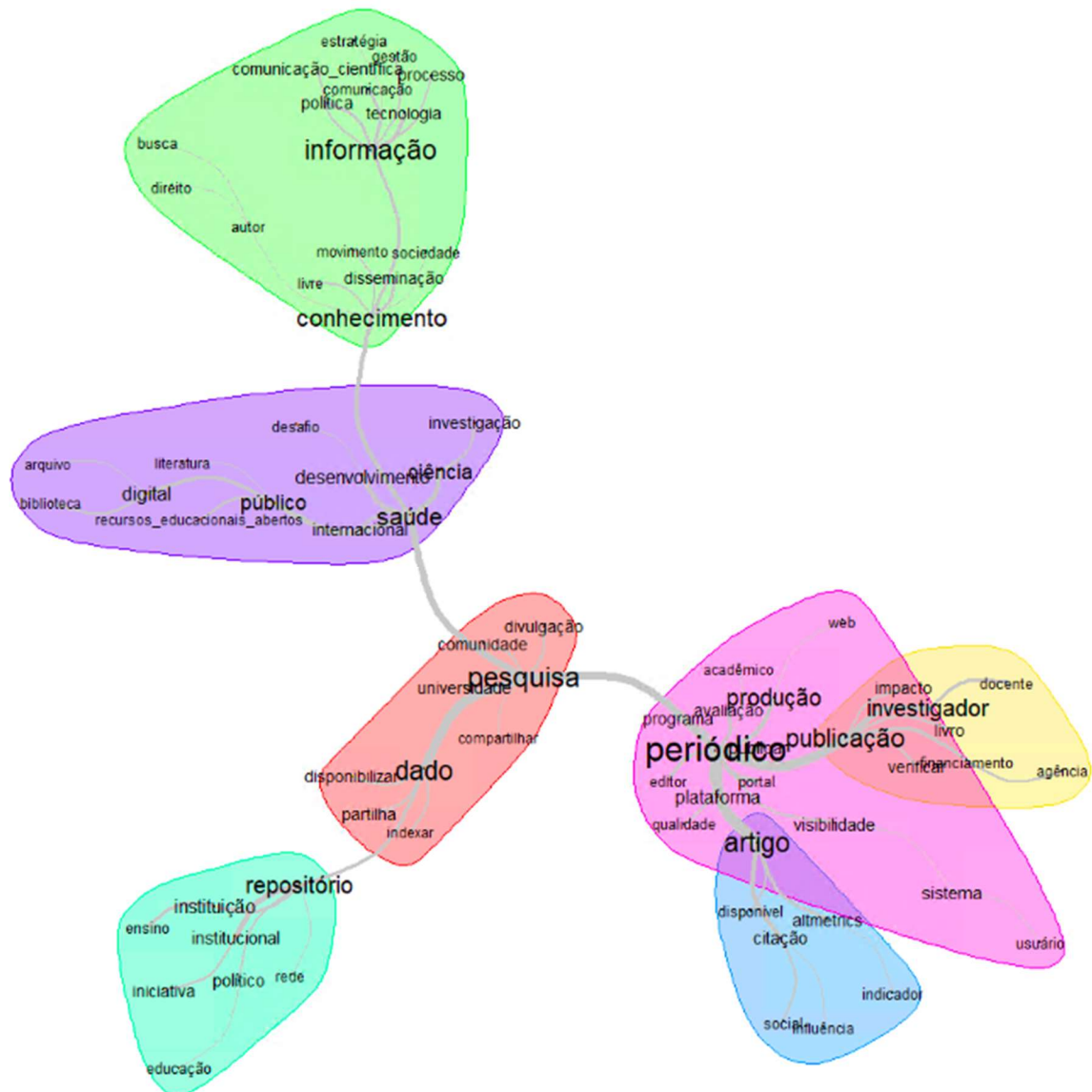
Resultados

1- Análise de conteúdo com recurso a software

1.a) Análise de similitude

A análise de similitude (coocorrências) entre os diferentes documentos que compõem o *corpus*, resultou no grafo abaixo. Salientaram-se as “comunidades” lexicais com um halo, para melhor visualização.

Grafo 1. Coocorrências lexicais



Nota: Layout de *Freuchmann-Reingold*, com comunidades e halos.. Fonte: Output do Iramuteq.

Morfologicamente, a espinha dorsal do grafo é constituída por “**Informação**”, “**Conhecimento**”, “**Saúde**”, “**Pesquisa**”, “**Periódico**”, “**Artigo**”.

Existem relacionamentos laterais importantes tais como, “**Pesquisa-Dado-Repositório**” e “**Periódico-Publicação-Investigador**”:

- A principal comunidade lexical foi gerada pelo termo “**periódico**”. É intercetada por duas comunidades, decorrentes dos termos “**publicação**” e “**artigo**”. Este conjunto contém os termos: “**financiamento**”, “**impacto**”, “**agência**”, “**investigador**”, “**avaliação**”, “**portal**”, “**desafio**”, “**critério**”, “**qualidade**”, “**análise**”, “**citação**”, “**altmetrics**”, “**indicador**”, “**visibilidade**”, “**social**”. Remete para as questões económicas e de mercado relacionadas com a avaliação científica, os seus desafios e contradições. Refira-se que a presença do termo “**livro**” na comunidade “**publicação**” é um indicador que outro tipo de documentos, neste caso livros, fazem parte deste paradigma.

- A comunidade lexical “**pesquisa-dado**” encerra “**partilha**”, “**compartilhar**”, “**comunidade**”, “**universidade**”, que remete para o paradigma de acesso à informação/dados e o papel das universidades. A ligação da comunidade com a gerada pelo termo “**repositório**”, que agrega os termos “**público**”, “**instituição**”, “**institucional**”, “**rede**”, remete para os canais preferenciais da recuperação da informação

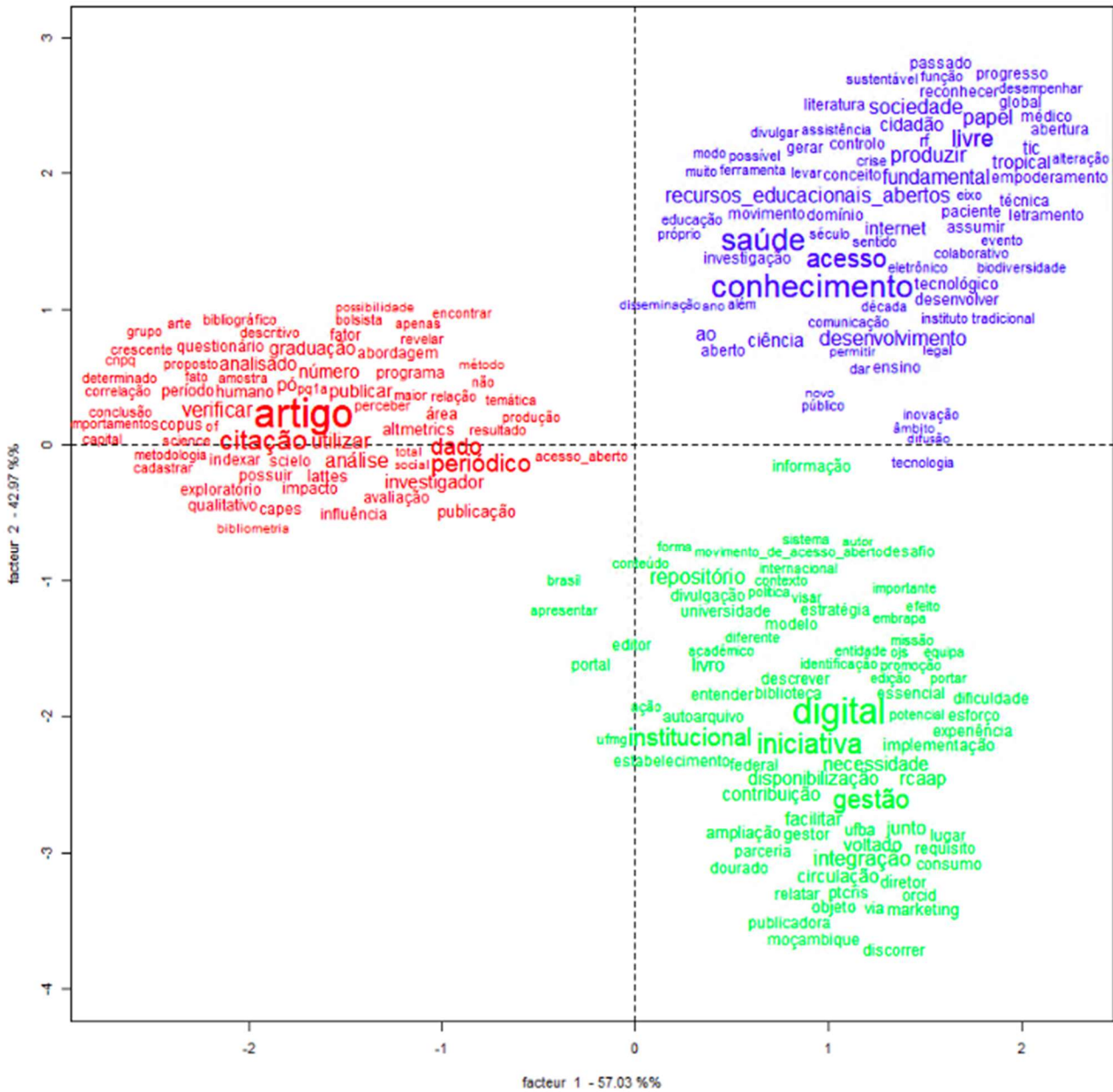
- O termo “**saúde**” produz uma comunidade que agrega “**público**”, “**ciência**”, “**desenvolvimento**”, “**investigação**”, “**digital**”, “**biblioteca**”, “**recursos educacionais abertos**” e “**desafio**”. O acesso aberto é estratégico no ensino/investigação naquele campo do saber e as bibliotecas desempenham um papel importante nesse processo. A presença da palavra desafio sugere problemas.

- A comunidade lexical “**conhecimento**” - “**informação**”, contém os termos, “**disseminação**”, “**movimento**”, “**sociedade**”, “**livre**”, “**gestão**”, “**iniciativa**”, “**estratégia**”, consigna os princípios epistemológicos do modelo de acesso aberto. Um ramo desta comunidade remete-nos para a questão dos “**direitos**” de “**autor**”.

1 b) Análise fatorial de correspondência

A análise fatorial de correspondência permitiu a visualização, no plano cartesiano, do “jogo de oposições” (Veiga & Paiva, 2013). Proporciona uma perspectiva fundada nas semelhanças e nas diferenças. Tal é perceptível na arrumação do léxico pelos quatro quadrantes.

Gráfico1. Análise fatorial de correspondência



O software calculou três classes lexicais bem definidas e praticamente estanques, que se distribuem da forma acima, no plano cartesiano. O conjunto explica praticamente 100% da variância, sendo, portanto, bastante relevante

Fonte: Output Iramuteq

A disposição do léxico indica que a principal oposição (na diagonal), decorre da classe “artigo” que encerra termos como “citação”, “dado” “periódico” “avaliação” “altmetrics”. Inere-se que a forma atual da avaliação científica (artigos) constitui um óbice, face à generalidade do modelo do acesso aberto.

2-Revisão de Literatura

Procedeu-se à revisão de literatura. Consideraram-se os artigos constantes do conjunto recuperado no RCAAP, os mesmo que subjazeram a análise de conteúdo acima.

A massificação das TIC ocorrida nos finais do século XX, abriu novas possibilidades na recuperação da informação. Segundo Borges (2016) “A Internet e a WWW trouxeram o ambiente tecnológico exato para garantir o acesso e a disseminação do conhecimento científico, dando um contributo decisivo para o que conhecemos hoje como Acesso Aberto”

O movimento do acesso aberto foi formalizado em três declarações de princípios, Budapeste (2002), Bethesda (2003) e Berlim (2003). Estas consignaram a importância da partilha de informação e dados, como um fator importante para avanço da ciência.

Segundo Borges (2016), o acesso aberto à informação científica (...) visa estimular a sua produção e impacto, através da promoção da sua ampla acessibilidade, mas também da sua reutilização em outros contextos, ampliando, deste modo, a sua repercussão na produção científica”. Ou seja, se por um lado, existe uma “visão utilitarista da abertura científica, mensurável em termos de produtividade e competitividade”, existe outra perspectiva consubstanciada em outros valores como, “coesão, direitos e justiça social que se levantaram na sua edificação “(Clínio, 2019).

O movimento do acesso aberto apesar de ter inerente os valores de liberdade e justiça no acesso à informação, não tem impedido tentativas de viragem desses valores, mormente, por iniciativas do mercado editorial. Em resposta à publicação aberta, “manifestam-se esforços de contrarreação por parte de agentes económicos, pelo desenvolvimento de novos modelos de negócio a fim de preservar a sua hegemonia, principalmente através de estratégias de integração vertical de plataformas de pesquisa como forma de controle das publicações.” (Appel, 2019).

Segundo Quintanilha (2019), cedendo às fraquezas idiossincráticas a comunidade do acesso aberto deixa-se abarcar pela indústria com fins lucrativos, pondo a ênfase nas métricas e replicando o modelo original de comunicação científica. Se é verdade que a informalidade pode prejudicar a qualidade, também o reverso promove o fechamento.

Segundo Ricarte (2018), toda a pesquisa científica é constantemente avaliada, seja na fase de conceção durante a busca por financiamento, seja na divulgação dos resultados. Os próprios cientistas são também constantemente avaliados, o que impacta as suas próprias carreiras. Mas o modelo requer novas métricas que permitam avaliar o alcance da divulgação pública e o próprio processo de pesquisa. Estas métricas passam, na opinião de diversos autores (Cintra, 2017); (Gouveia & Carvalho, 2017) (Araújo, Oliveira & Lucas), por redes sociais e pela web.

Segundo Henriques (2020), na área da saúde, o acesso aberto é considerado, por vários autores, como vital. Este possibilita o acesso, por parte dos profissionais de saúde, a resultados de investigação, sem necessidade de investimento financeiro e possibilita a integração desse conhecimento na melhoria dos cuidados prestados aos pacientes. Assina ainda, um dos perigos da subversão económica do acesso aberto na saúde: O declínio dos padrões de qualidade, motivado pelas revistas em acesso aberto, aceitarem artigos sem revisão por pares, para beneficiarem das taxas pagas pelos autores. Esta oposição ao modelo acaba por ter consequências a nível dos cuidados de saúde.

Concomitantemente, surgem novos desafios. O uso em massa das tecnologias digitais para a recolha, tratamento e análise de dados tem vindo a criar, o que alguns autores designam como o “quarto paradigma em ciência” (Silva, Pinto, Carvalho, Pereira & et al, 2017), caracterizado por informação muitas vezes organizada de forma caótica e uma produção massiva de dados. Os princípios FAIR na gestão de dados (*findable, accessible, interoperable, reusable*), “visam a normalização na sua gestão, garantindo a sua interoperabilidade, sendo crítico para a expansão mundial da abertura dos dados de pesquisa” (Henning, Ribeiro, Silva-Santos, et al. 2019).

O acesso aberto é inseparável de competências de pesquisa, recuperação, utilização e sentido crítico da

informação. Isso constitui um desafio. No ensino superior, os bibliotecários e também professores, têm vindo a desenvolver estratégias para apoiar o desempenho dos estudantes na recuperação da informação, e promovendo a literacia da informação (Antunes, Sanches, Lopes e Rodrigues, 2018)

Face ao exposto é fácil constatar que o acesso aberto é um domínio complexo, que inclui múltiplos conceitos. Esse domínio não é estático, vai evoluindo. Contudo, o reconhecimento do atual “estado da arte”, pode ser determinante para a própria preservação do modelo, mormente, por meio de ações que contrariem tendências que atualmente o hostilizam.

CONCLUSÕES

A recuperação da informação através da análise de conteúdo com software e através da revisão de literatura, mostraram um elevado nível de coerência. A revisão de literatura permitiu aprofundar cognitivamente o conhecimento veiculando antes pela análise de conteúdo.

Para melhor elucidação das diferenças relativas e das vantagens/desvantagens de cada um dos métodos, procedeu-se à construção duma tabela que resume as inferências resultantes da comparação efetuada:

**Tabela 1. Vantagens/Desvantagens relativas:
Análise de conteúdo com recurso a software vs Revisão de literatura**

	Análise de conteúdo textual com recurso a software	Revisão de literatura
Vantagens Relativas	-Correlação automática entre os léxicos com evidência dos mais importantes para a composição do domínio	-Não requer conhecimentos especializados
	-Demarcação automática de comunidades lexicais	-Compreensão aprofundada sobre o objeto
	-Rápida	-Fundamental na construção de um discurso fluente acerca do objeto
	-Não subjetiva: Não influenciável por fatores cognitivos	-Pormenorizada
	-Traça um sumário do estado da arte do objeto em estudo (considerando um recorte temporal)	-Se exaustiva, proporciona a descoberta de documentos únicos ou originais
	-Pode ser utilizado como guião para a escolha de subtemas considerados relevantes pelo leitor	
	-Através da análise fatorial de correspondência consegue-se perceber o jogo de oposições entre os léxicos e consequentemente as diferentes facetas conflituantes do objeto	
	-Não influenciada por métricas ou outros fatores externos	
Desvantagens relativas	-Requer conhecimentos especializados, quer sobre o software, quer sobre análise de conteúdo/teoria dos grafos/estatística multivariada para a respetiva interpretação	-A mente do leitor é um filtro cognitivo: Subjetiva
	A análise de similitude, por ser baseada em coocorrências tende (dependendo dos valores mínimos de coocorrência escolhidos pelo investigador) a ocultar assuntos originais ou únicos (que não geram ocorrências com outros documentos constituintes do <i>corpus</i>).	-A escolha dos documentos pode ser influenciada por métricas ou outras distorções externas
	Em caso de <i>corpus</i> muito extensos os gráficos gerados podem ser confusos e não legíveis	-Tende a não ser totalmente abrangente o que implica uma potencial dificuldade na correlação exaustiva dos assuntos
	A mera visualização de léxicos, mesmo que relacionados, não ajuda à construção de um discurso fluente, acerca do objeto em estudo	-Lenta -Tende a ser desorganizada

Fonte: A autora

Face ao exposto é possível afirmar que os dois métodos se complementam e por tal, devem, idealmente, ser utilizados em conjunto. Ou seja, a análise de conteúdo com software não substitui a revisão de literatura e vice-versa. A utilização conjunta permite uma assimilação mais célere dos conceitos do domínio, pois, na passagem de um método para o outro, estes passam a ser “reconhecidos”, invés de apenas “conhecidos” o que contribui largamente para a sua compreensão.

Recomendações

A análise de conteúdo com recurso a software deveria ser mais utilizada na ciência da informação para a recuperação de informação, mormente, porque a existência de softwares para a análise de *corpus* textuais, alguns *open-source*, (IRaMuteQ, Gephi) permitem um acesso facilitado a este tipo de análise. Mas não ao ponto de excluir a revisão de literatura.

Limitações do estudo

A análise foi realizada no domínio do acesso aberto, conforme documentos recolhidos no RCAAP (2016-2020). Serão precisas mais análises da mesma natureza para confirmar as inferências realizadas.

Referências Bibliográficas

- Antunes, Sanches, Lopes, Rodrigues, *et al* (2017). *Information literacy in the portuguese university contex*. [Paper presentation]. ECIL. <https://bit.ly/308zAlb>
- Appel, A. L. (2019). Dimensões tecnopolíticas e econômicas da comunicação científica em transformação [Tese de Doutorado] Universidade Federal de Rio de Janeiro]. <https://www.rcaap.pt/detail.jsp?oai:ridi.ibict.br:123456789/1024>
- Borges, M. M. (2016). ‘Christmas is over’... Is spring coming? *A publicação da ciência em acesso aberto*. [Paper presentation] XII Jornadas APDIS, Lisboa. <https://bit.ly/37Lamgg>
- Brandt, M., Brascher, M. (2010). Folksonomia: esquema de representação do conhecimento? *TransInformação*, 22(2), 111–121.
- Camargo, B. V., Justo, A. M. (2013). IRAMUTEQ: um software gratuito para análise de dados textuais. *Temas em Psicologia*, 21(2)
- Cervi, E. (2018). *Análise de conteúdo automatizada para conversações em redes sociais online: uma proposta metodológica*. [Paper presentation]. 42º Encontro ANPOQS, Mato Grosso. <https://bit.ly/38UkYtz>
- Clinio, A. (2019). Ciência Aberta na América Latina: Duas perspectivas em disputa. *Transinformação*, 31. <https://doi.org/10.1590/238180889201931e190028>
- Henriques, A (2020). *A edição de revistas de saúde portuguesas: visibilidade e perspectivas sobre Acesso Aberto*. APDIS
- Henning, P. C., Ribeiro, C. J. S., Santos, L. O. B. da S., & Santos, P. X. dos. (2019). GO FAIR e os princípios FAIR: O que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. *Em Questão*, 25(2), 389–412. <https://doi.org/10.19132/1808-5245252.389-412>
- Hoffmann, Y. T., Bisset Alvarez, E., & Martí-Lahera, Y. (2020). Análise textual com IRaMuTeQ de pesquisas recentes em História da educação matemática no Brasil: Um exemplo de Humanidades Digitais. *Investigación Bibliotecológica: archivonomía, bibliotecología e información*, 34(84), 103. <https://doi.org/10.22201/iibi.24488321xe.2020.84.58097>
- Lopes, C., Antunes, M. da L., & Sanches, T. (2019). *Ciência Aberta & literacia da informação: Estudo de caso num centro de investigação*. CONFOA 2019 – 10a Conferência Luso-Brasileira de Ciência Aberta. <https://bit.ly/3qG9XnR>
- Machi & McEvoy (2008). *The literature review: Six steps for success*. University de La Verne.
- Pinto, A. L., M, Silva, A. M. & Sena, A. B. (2010). Ontologias baseadas na visualização da informação das redes sociais. *PRISMA*. 13, 5–24.

- Quintanilha, T. L. (2019). Os quatro grandes desafios ao modelo de Ciência Aberta: (Des)acreditação, informalidade, comodificação e predação. *Texto Livre: Linguagem e Tecnologia*, 12(2), 202–213. <https://doi.org/10.17851/1983-3652.12.2.202-213>
- Ramos, M.G, Valderez, M, Amaral-Rosa, M. (2018). *Contribuições do software IRAMUTEQ para a Análise Textual Discursiva*. [Paper presentation]. AtasCIAIQ, Fortaleza <https://bit.ly/2X30Dgc>
- Ricarte, I. (2018). Métricas em Ciência Aberta. *XIII Jornadas APDIS*, 1–11. <https://bit.ly/3gueMfi>
- Saviati, M.E. (2017). *Manual do aplicativo Iramuteq*. Planaltina.
- Silva, D. M., Pinto, E. M., Carvalho, É. R. S. de, Pereira, P. R., & Leite, F. C. L. (2017). Comunicação científica sob o espectro da ciência aberta: Um modelo conceitual contemporâneo. *Revista eletrônica, comunicação e informação em saúde*, 11. <https://doi.org/10.29397/reciis.v11i0.1414>
- Silva, F. C. C. da, & Silveira, L. da. (2019). O ecossistema da Ciência Aberta. *Transinformação*, 31.
- Vanyushkin, A. & Grashenko, L. (2018) Analysis of Text Collections for the Purposes of Keyword Extraction Task. *JIOS*, 44(1), 171-184. <https://doi.org/10.31341/jios.44.1.8>