APROXIMAÇÃO DA BIBLIOMETRIA E RECUPERAÇÃO DE INFORMAÇÃO NA BRAPCI

Rene Faustino Gabriel Junior (UNESP)
renefgj@gmail.com

EIXO TEMÁTICO: Métodos, Técnicas e Ferramentas para Estudos Bibliométricos e Cientométricos
MODALIDADE: Apresentação oral

1 INTRODUÇÃO

É muito comum o usuário receber uma grande quantidade de documentos quando realiza buscas em Sistemas de Recuperação de Informação (SRI), sendo necessário refinar os termos usados na consulta objetivando reduzir o número de documentos para leitura e análise. Porém o refinamento da estratégia de busca faz com que muitos documentos importantes acabem sendo omitidos, ficando invisíveis para o usuário. Os tradicionais SRI são normalmente baseados em um modelo reducionista, utilizando a lógica booleana para identificação de documentos, este estudo visa apresentar uma nova dimensão Recuperação de Informação (RI), aplicando a bibliometria e cientometria como elementos de expansão de busca, por meio de interações entre o usuário e o SRI, de forma a apresentar ao usuário as estruturas e atividade científicas referentes ao domínio ao qual está buscando informações.

Segundo Wormell (1998) as primeiras aproximações da bibliometria com a RI sugiram na década de 1950, quando a demanda por informação era alta e os profissionais da informação não conseguiam produzir *abstracts* e índices de acordo com a necessidade dos pesquisadores. Dimensionando este problema, Eugene Garfield em 1955 propôs à utilização de métodos bibliométricos, com o uso das citações, para gerar índices que corroborassem a RI, de forma a ser uma alternativa aos sistemas baseados na linguística e indexação de termos.

Desde então, segundo Araújo (2009), na literatura vários SRI já foram testados, tendo como premissa, as propriedades da revocação e precisão. Ainda segundo o autor, os SRI bibliométricos utilizam técnicas estatísticas para estabelecimento de padrões de regularidade em itens informacionais, tendo como indicadores estatísticos as citações, número de livros, de edições, de autores dentre outros elementos bibliográficos. Estes indicadores são utilizados empregando os modelos desenvolvidos principalmente por meio de leis empíricas estabelecidas desde a década de 1920 (Lotka, Bradford, Zipf). Para Araújo (2009), o campo só ganhou notoriedade após a década de 1960, com as possibilidades de automação e com a criação do campo de estudos de análise de citação. Entretanto, Mutschke e outros autores

(2011) ressaltam que a aproximação da bibliometria e RI tem uma proposta de avaliação, de forma a definir e identificar um conjunto de documentos, e ou autores de referência com base em uma metodologia e utilização de indicadores padronizados.

Em 1998, Wormell destacou que a análise de citação, originalmente proposta por Garfield, ainda não foi percebida pela maior parte dos profissionais da informação como instrumento nos SRI, afirmando que a principal vantagem da indexação de citações é a capacidade de transpor o uso de formas linguísticas comuns como palavras do título, palavras-chave ou cabeçalhos de assunto, sendo "o papel simbólico representado pela citação na representação do conteúdo de documentos é uma dimensão ampla da recuperação da informação" (WORMELL, 1998). Ou seja, os autores sugerem que a aplicação de indicadores bibliométricos e índices de citação combinados com expressões da linguagem natural do usuário podem melhorar consideravelmente as buscas exaustivas de literatura.

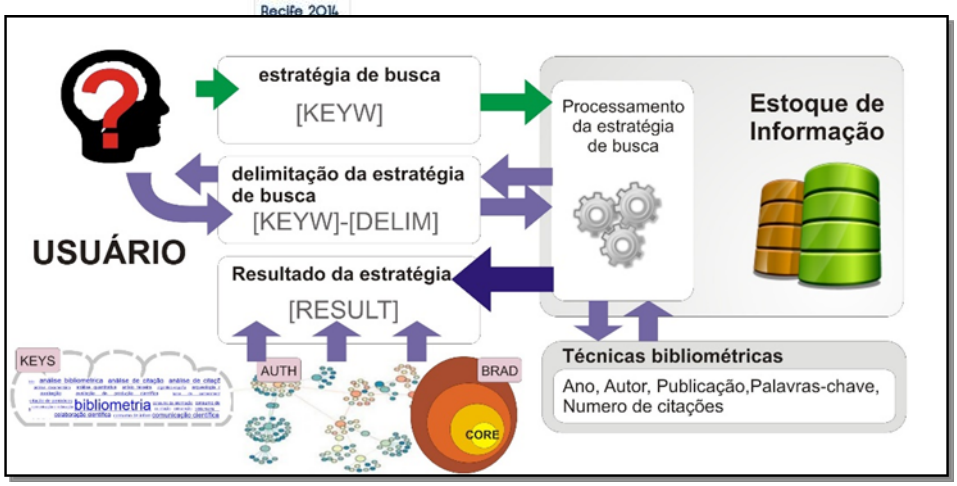
No processo metodológico, a bibliometria moderna, segundo Glänzel (2003), trabalha com três objetivos principais, ou grupos alvo que determinam tópicos e subáreas de pesquisa. O primeiro grupo, consistindo na bibliometria para bibliométricos, focando na metodologia, com trabalhos voltados para a pesquisa básica. O segundo grupo, com a aplicação em disciplinas científicas, formando o maior grupo e interesses mais diversificados, delimitada principalmente em seu domínio, e se relaciona com as pesquisas em RI e SRI. O último grupo utiliza a bibliometria na avaliação da pesquisa, desenvolvendo e analisando a gestão científica e as políticas em ciência e tecnologia (MUELLER, 2013). Nesta perspectiva, este estudo se qualifica no segundo grupo, com estudo aplicado em SRI.

Na literatura brasileira em periódicos são poucos os estudos que buscam a aproximação da bibliometria, bases de dados e RI. Em uma busca exploratória em artigos das revistas em CI, destaca-se o estudo de Oliveira (1984), que apresentou práticas bibliométricas para a implementação e operação de SRI; as reflexões e estudos de Wornell (1998) realizados no Centro de Estudos Informétricos de Copenhague; a proposta de Mugnaini (2003) com a associação da bibliometria, linguística e indexação na filtragem de informações e RI; e estudo avaliativo de Meireles e Cendón (2011) sobre a eficiência e a viabilidade do uso de Redes Neurais Artificiais para categorizar e classificar documentos utilizando as referências bibliográficas por elas citadas.

As metodologias empregadas na RI tem evoluido nos ultimos anos, principalmente com o desenvolvimento das tecnologias da informação, com a ampliação do espaço de armazenamento, velocidade dos computadores e da Internet . Estas evoluções fizeram com que as bases de dados, principalmente as *online,* pudessem processar um volume de informação cada vez maior em uma velocidade menor. Desta forma não sendo mais o hardware o limitador das buscas de informação, mas a metodologias de recuperação e os modelos de organização da informação dentro das bases de dados os limitadores.

Este trabalho tem como objetivo realizar um estudo preliminar sobre a aplicação da bibliometria em SRI, de forma a expandir a área de domínio a qual o usuário está realizando suas buscas. De forma mais específica, o estudo parte de uma busca exploratória na literatura para identificação de estudo similar, e na construção de dois modelos bibliométricos que serão aplicados em uma base de dados concreta, de forma a testar os modelos.

Procurando um modelo expansivo na literatura, localizou o a proposta de Mutschke e outros autores (2011) com a implementação de três tipos diferentes de modelos científicos para os serviços de busca, de forma a agregar valor a diferentes aspectos da atividade acadêmica, que podem ser utilizado de forma homogenea ou combinadas, abordando três diferentes dimensões, conforme apresentado no Quadro 1. O primeiro modelo é baseado no acoplamento de palavras-chave em uma estrutura cognifiva de um campo, descrevendo as relações com outros termos (KEYS). O segundo modelo proposto pelos autores tem relação com a *Bradfordizing* (BRAD), originalmente descrita por White em 1981, sendo uma simples utilização da lei de Bradford. O terceiro modelo proposto é baseado na centralidade de autores, em análise de rede social, onde os mesmo são recuperados na estratégia de busca e organizados em estruturas sociais, representados por meio de um grafo, possibilitando ao usuário a identificação de autores centrais em um domínio (AUTH).



Quadro 1 – modelo expansivo do uso da bibliometria na estratégia de busca do usuário
Fonte: Autor baseado no modelo de Mutschke e outros (2013).

O Quadro 1 é a representação dos modelos (KEYS), (BRAD) e (AUTH) com interações constantes do usuário no mecanismos de busca, qualificando os resultados com o contexto do domínio.

Para aplicar os modelos foi utilizado a Brapci como repositório informacional, o que facilita a construção e validação do protótipo dos modelos.

2 BASE DE DADOS REFERENCIAL DE ARTIGOS DE PERIÓDICOS EM CIÊNCIA DA INFORMAÇÃO (BRAPCI)

A Brapci, cita-se sua criação em 1996, tem como objetivo subsidiar estudos e propostas na área de Ciência da Informação, fundamentando-se em atividades planejadas institucionalmente, sua construção está subsidiando a construção de um observatório para estudos analíticos e descritivos sobre a produção editorial e organização do conhecimento de uma área em crescente desenvolvimento - a Ciência da Informação.

Disponibilizada ao público desde 2008 como produto do projeto de pesquisa “Opções metodológicas em pesquisa: a contribuição da área da informação para a produção de saberes no ensino superior”. Em seu estado atual, conta com 35 publicações, sendo destas 27 ativas e nove descontinuadas. Com mais de 13.641 documentos indexados, destes 9.321 de artigos, dossiês e comunicações científicas. Desde 2012 estão sendo coletadas as referências bibliográficas dos artigos, de forma a proporcionar estudos bibliométricos e indicadores sobre a CI no Brasil.

A Brapci é o instrumento de avaliação e validação das metodologias propostas pelo grupo de estudo G3PI da UFPR, que desenvolve pesquisa das mais diversas perspectivas dentro da base de dados.

3 TRAJETÓRIA METODOLÓGICA

Este estudo partiu de uma busca exploratória nas bases de dados Brapci e Scopus para identificar trabalhos que tenham como tema a aproximação entre a bibliometria, cienciometria, infometria e a RI. Foram recuperados 144 documentos da Scopus e sete na Brapci sobre o assunto. Após uma leitura rápida, observou-se que o tema é tratado desde a década de 1950, porém a maioria dos trabalhos tem relação associada a filtro de informação, que não é a proposta deste estudo, reduzindo para 18 trabalhos relacionados ao contexto.

Com fundamentos nos artigos consultados, que exploraram o modelo expansivo do uso da bibliometria na estratégia de busca de usuários, foram desenvolvidas propostas seguindo os modelos dos autores que pudessem ser aplicadas na Base Brapci, principalmente em estudos referente à ampliação do repertório documental em SRI. Neste contexto foi desenvolvido um modelo com as palavras-chave (KEYS) e outro com as relações de coautoria (AUTH) conforme modelo proposto por Mutschke e outros (2011).

4 RESULTADOS

O primeiro modelo foi testado no protótipo com a construção automática de uma nuvem de *tags,* conforme os documentos recuperados na estratégia de busca do usuário. Para esta construção, identificam-se todos os trabalhos recuperados na busca, ordenando todas as palavras-chave em ordem alfabética mantendo sua frequência de incidência. A nuvem de *tags* apresenta de forma gráfica, o resultado desta ordenação, aplicação um algoritmo matemático, baseado na frequência das palavras-chave, de forma a atribuir um tamanho de letra maior aos termos com maior incidência.

Para testar o comportamento do protótipo buscou-se um conjunto de termos que representasse a "inclusão", no contexto de "medidas para eliminar ou dinimuir as desigualdades sociais", para este fim, foram utilizados os termos "inclusão", "desigualdade",

"exclusão" e "social" de forma individual e ou combinada. Em uma análise *a priori* no *corpus* da Brapci, sabe-se que exitem 283 documentos relacionados ao tema, denomida de base de controle. Os resultados das buscas são apresentadas no Quadro 2. A percentagem de recuperados é calculada pelo total do resultado dividido pelo resultado esperado, enquanto a precisão é mensurada pelos trabalhos que deveriam ser recuperado.

Estratégia de busca	Resultados	Resultados esperado	% recuperado	% precisão
Inclusão	336	283	118.7%	78%
"Inclusão" e "desigualdade"	17	283	6,0%	5,6%
"Inclusão" e "social"	207	283	73,1%	66,1%
"Igualdade" "social"	35	283	12,3%	10,9%
"Exclusão" "social"	33	283	11,7%	10,9%
"Desigualdade" "social"	34	283	12,1%	7,1%

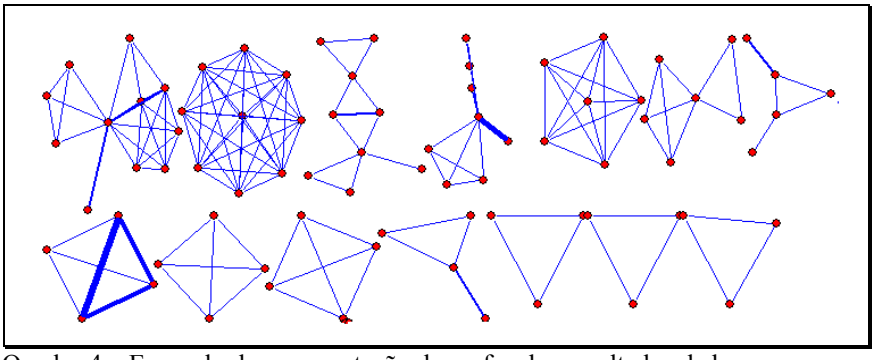
Quadro 2 – Resultado da estratégia de busca por termos.

Analisando o Quadro 2 observa-se que conforme a variação e combinação dos termos os resultados das recuperações são afetadas diretamente, seja pelo excesso de documentos, pela precisão ou por omissão. A consulta com maior precisão (78%) está indicada ao termo "inclusão", porém foram recuperados 336 trabalhos, dos quais, 220 relacionados à base de controle. Quanto maior a inclusão de outros termos, menor a precisão e percentual de recuperação, delimitando os resultados da busca.

Entretanto, quando utilizado a mesma estratégia de busca e visualizando os resultados por uma nuvem de *tags*, observa-se que a resposta é mais ampla, sugerindo ao usuário a busca por outros termos relaciona ao tema. Esta forma possibilita ao usuário redefinir sua estratégia por constantes interações. Nos testes, foi possível identificar os 283 trabalhos em 12 operações. Ou seja, nenhum documento ficou oculto do usuário, porém foram necessárias diversas interações até suas exaustão. No Quadro 3 é apresentado um exemplo de interação utilizando o termo "Inclusão", demonstrado as relações bibliométricas de coocorrência dos termos, destacando-se os termos "inclusão digital", "inclusão social" , "cidadania" pelo tamanho de sua fonte.

Quadro 3 – Resultado da busca por "inclusão" no protótipo
Fonte: Autor (2014).

O segundo modelo aplicado ao SRI foi a utilização de grafos para representar estruturas científicas, apresentando autores centrais e a composição de grupos de pesquisa referentes ao tema. Para esta proposta, foi criado um novo algoritmo que gera internamente uma matriz de coautoria de autores, conforme trabalhos recuperados, identificando as relações entre autores. Na proposta inicial, todos os autores eram para ser representados conforme resultado do SRI, formando círculos de tamanhos diferentes conforme frequência de publicação, e linhas de espessuras diferentes conforme incidência de coautoria. Porém, no teste, quando recuperado 336 trabalhos, houve a identificação de 329 autores e 858 ligações de coautoria. Esta quantidade de autores, fez com que o grafo explicitasse muitas relações, ficando praticamente ilegível. Sendo necessário aplicar filtros de visualização, para apresentar somente as ligações com três ou mais autores, resultando no Quadro 4.



Quadro 4 – Exemplo da representação de grafos dos resultados da busca
Fonte: Autor (2014).

No grafo, houve ainda a necessidade de retirar o nome dos autores para evitar a intercalação dos mesmos com as linhas de ligação. Na forma que foi testado o protótipo, foi possível identificar onze grupos (com mais de três autores) que trabalham com o tema, dos

quais seis apresentam maior atividade. A representação por grafos não se apresentou muito eficaz na ampliação do repertório para o usuário. Porém os dados do grafo possibilitaram a identificação dos autores com maior centralidade, com o destaque aos pesquisadores mais articulados e significativos em seu conjunto. Deste cálculo, foi possível identificar alguns pesquisadores "chave" com maior relevância sobre o tema.

Em conjunto com a centralidade de autores, foi possível identificar as revistas que mais publicam sobre o tema e o ano que tema teve maior quantidade de artigos publicados.

5 CONSIDERAÇÕES

A aproximação da bibliometria com a RI na base Brapci se demonstrou promissora nos primeiros estudos efetivados, porém ainda é necessário o refinamento dos modelos, e aplicação de outros estudos em domínios diferentes. Uma higienização dos dados da base ainda é necessária, com menos erros de indexação para que os relacionamentos se apresentem mais eficazes, principalmente com a criação e utilização de vocabulários controlados. Um dos maiores problemas no SRI foi originado da falta de padronização dos termos utilizados, que em muitos casos são definidos pelo autor do trabalho e estes são replicados na indexação da base.

O protótipo, na busca por termos, não se apresentou mais eficiente que o processo tradicional, porém possibilitou ao usuário, por meio de interações, uma busca mais exaustiva na base de dados, recuperando documentos que na busca booleana seriam ocultados. O refinamento da busca ocorre pela constante troca de informações entre o SRI e o usuário, possibilitando a recuperação de todos os documentos do tema por meio de algumas interações.

Os resultados apresentados neste estudo ainda são considerados preliminares, necessitando concretizar o protótipo e testar o modelo com usuários em condições de necessidade de informação. Nenhum dos modelos propostos no protótipo apresentou precisão na RI em suas primeiras interações, porém possibilitaram a ampliação da visão do usuário em seu processo de busca por informação. Os resultados sugerem que existe potencial na proposta, possibilitando a ampliação e qualificação do repertório, suprindo o usuário não somente com fontes de informação, mas apresentando estruturas e fontes qualificadas de

informação, tendo no usuário um elemento ativo no processo de RI. Novos modelos estão sendo articulados, combinando principalmente o referencial teórico (referências) dos artigos, identificando afinidades temáticas e tendências.

REFERÊNCIAS

ARAÚJO, C. A. A. Correntes teóricas da ciência da informação. **Ciência da Informação**. v. 38, n. 3, 2009.

GARFIELD, E. Citation indexes for science: A new dimension in documentation through association of ideas. **Science**, Washington, v. 122, n. 3159, p. 108-111, July 1955.

GLÄNZEL, W. **Bibliometrics as a research field: a course on theory and application of bibliometric indicators**. [S. l.]: Courses Handout, 2003.

MEIRELES, M. R. G.; CENDÓN, B. V. Categorização e classificação de documentos a partir de suas citações: uma proposta baseada em redes neurais artificiais. **DataGramaZero**, v. 12, n. 5, 2011.

MUELLER, S. P. M. Estudos métricos da informação em ciência e tecnologia no Brasil realizados sobre a unidade de análise artigos de periódicos. **Liinc em Revista**, v. 9, n. 1, p. 6-27, 2013.

MUGNAINI, R. A bibliometria na exploração de base de dados: a importância da Linguística. **Transinformação**, v. 15, n. 1, p.45-52, 2003.

MUTSCHKE, P.; et. al. Science models as value-added services for scholarly information systems. **Scientometrics**, v. 89, n. 1, p.349-364, 2011.

OLIVEIRA, S. M. Aplicações e Limitações dos Processos Bibliométricos. **Revista Brasileira de Biblioteconomia e Documentação**, v. 17, n. 1/2, p. 55-65, jan./jul. 1984.

White, H. D. 1981. 'Bradfordizing' search output: how it would help online users. **Online Information Review**, v. 5, n. 1, p-47-54, 1981.

WORMELL I. Informetria: explorando bases de dados como instrumentos de análise. **Ciência da Informação**, v. 27, n. 2, p. 210-216, 1998.