Biblos E-ISSN 2236-7594

O FLUXO TEMPORAL DE TERMOS RELEVANTES: uma análise em teses da UFMG de 2007 a 2018 nas ciências sociais aplicadas

Luiz Antônio Lopes Mesquita
Doutorando em Gestão e Organização do Conhecimento pela Escola de Ciência da Informação da Universidade Federal de Minas Gerais (UFMG). Mestre em Ciência da Informação pela UFMG, Belo Horizonte, Minas Gerais, Brasil.
mesquitaluiz@hotmail.com
https://orcid.org/0000-0002-0484-0117.

Renato Rocha Souza
Doutor em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG), Brasil. Professor e Pesquisador da Escola de Matemática Aplicada da Fundação Getúlio Vargas, Brasil. Pesquisador da Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Brasil. Bolsista de produtividade do CNPq.
renato.souza@fgv.br
https://orcid.org/0000-0002-1895-3905.

Célia da Consolação Dias
Doutora em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG).
Professora do Departamento de Organização e Tratamento da Informação da Escola de Ciência da Informação da UFMG.
celiadias@gmail.com
https://orcid.org/0000-0003-0891-6454.

RESUMO

O objetivo geral desta pesquisa foi analisar se há uma variação temporal característica da distribuição de valores de termos relevantes ao longo do tempo da produção de textos que possa contribuir como um critério para o processo de sua indexação automática. Foram analisadas as teses de doutorado dos programas de pós-graduação (PPGs) da área de Ciências Sociais Aplicadas da UFMG, considerando-se 7 PPGs distintos, sendo cada um deles um corpus, com um total de 641 teses defendidas período de 12 anos, de 2007 a 2018. Os termos considerados foram todos os sintagmas nominais contidos nos próprios textos das teses. Cada sintagma nominal recebeu um valor associado à sua relevância como descritor de acordo com os critérios de frequência do termo na própria tese (TF – Term Frequency) e com o inverso da frequência de ocorrência do termo no total de teses de cada PPG (IDF – Inverse Document Frequency). As teses foram divididas em 12 grupos em cada PPG para o cálculo da data média de defesa das teses e da média de pontuação consolidada dos termos relevantes nas teses. Como resultados, identificou-se o comportamento característico de cada PPG através de um gráfico de dispersão do nível médio de pontuação de relevância ao longo do tempo. Para cada gráfico de cada um dos 7 PPGs foi adicionada uma linha de tendência, considerando seu respectivo R^2, e feita sua análise específica. Todos os comportamentos de distribuição temporais foram caracterizados em equações polinomiais e podem ser aplicados como critério para indexação automática.

Palavras-chave: Recuperação da Informação Temporal. Indexação Automática. Sintagmas Nominais.

THE TEMPORAL FLOW OF RELEVANT TERMS: an analysis in UFMG theses from 2007 to 2018 in applied social sciences

ABSTRACT

This research's general objective was to analyze if there is a temporal variation characteristic of the distribution of values of relevant terms over the time of the production of texts that can contribute as a criterion for the automatic indexing process. The doctoral theses of the graduate programs (PPGs) in the area of Applied Social Sciences at UFMG were analyzed, considering seven different PPGs, each of which is a corpus, with 641 theses defended in a period of twelve years, from 2007 to 2018. The terms considered were all the noun phrases contained in the texts of the theses. Each noun phrase received a value associated with its relevance as a descriptor according to the term frequency criteria in the thesis itself (TF – Term Frequency)

and with the inverse of the frequency of occurrence of the term in the total of theses of each PPG (IDF – Inverse Document Frequency). The theses were divided into 12 groups in each PPG to calculate the average defense date of the theses and the average consolidated score of the relevant terms in the theses. As a result, each PPG's characteristic behavior was identified through a scatter plot of the average level of relevance score over time. For each graph of each of the 7 PPGs, a trend line was added, considering its respective R^2, and its specific analysis was made. All temporal distribution behaviors were characterized in polynomial equations and applied as a criterion for automatic indexing.

Keywords: Temporal Information Retrieval. Automatic Indexing. Noun Phrase.

Recebido em: 10/08/2020 Aceito em: 30/10/2020 Publicado em: 31/12/2020

1 INTRODUÇÃO

Em 1974 foram publicados 419 artigos relacionados à computação, cujas 3.812 referências em todos eles foram analisadas por Salton & Bergmark (1979), num dos primeiros estudos bibliométricos dessa ciência, para a qual concluíram existir naquele momento três principais áreas: fundamentos teóricos, *hardware* e *software*. Partindo da mesma base metodológica, 45 anos depois, Devarakonda *et al.* (2020) concluíram que a ciência da computação não somente expandiu em volume, mas também em subáreas interdisciplinares conforme cerca de 8 milhões de documentos analisados.

Bush (1945, pág.95) já predissera o crescimento da informação e suas consequências ao pontuar sobre a necessidade da interdisciplinaridade:

> "O investigador fica pasmo com as descobertas e conclusões de milhares de outros pesquisadores – conclusões que ele não consegue encontrar tempo para apreender, muito menos para lembrar, tal como aparecem. No entanto, a especialização se torna cada vez mais necessária para o progresso, e o esforço para estabelecer uma ponte entre as disciplinas é ainda superficial" (*ibidem,* tradução livre).

Para Saracevic (1996, pág.42), Bush (1945), como cientista do MIT e em plena Segunda Guerra Mundial, não só aponta o problema da "explosão informacional" como também sua possível solução com o uso das "tecnologias da informação", criando o cenário para o surgimento da Ciência da Informação (CI) nos anos 50. Mooers (1951) aponta um dos caminhos da C.I. que ele denominaria como Recuperação da Informação (RI) através de seu protótipo Zatocoding.

A Ciência da Computação também "desenvolve significativas pesquisas nessa área

[RI] com o objetivo principal de prover aos usuários de seus sistemas um fácil acesso à informação do seu interesse" (BAEZA-YATES; RIBEIRO-NETO, 2011, p.1, tradução do autor). Dentre muitas outras áreas que tornam a RI interdisciplinar, a Linguística contribui significativamente para o processamento de informações textuais em linguagem natural.

A contabilização de palavras isoladas pode ser feita facilmente com a identificação de delimitadores, como o espaço. Com os algoritmos de Processamento de Linguagem Natural (PLN) é possível utilizar cada vez mais estruturas linguísticas complexas. Uma dessas estruturas é o sintagma nominal (SN). Perini *et al*. (1996) apresenta que o SN possui maior valor semântico que a palavra isolada.

Os SNs podem ser extraídos automaticamente de textos. Os trabalhos de Kuramoto (1996), Souza (2005), Maia (2008), Corrêa *et al.* (2011), Mesquita *et al.* (2013; 2014) e outros apresentam como tema central a utilização de SNs através da sua extração em PLN de forma semiautomática e automática para a língua portuguesa. O crescente volume informacional somado à crescente capacidade de processamento, juntamente com avanços no PLN, abre espaço para novas pesquisas. Uma delas estaria no uso de sintagmas nominais em sistemas de recuperação da informação, como a indexação automática.

Os critérios para escolha de descritores para a indexação automática possuem suas raízes históricas nos conceitos de "frequência do termo", de Luhn (1957)[1], e de "especificidade", de Sparck Jones (1972). Posteriormente a eles, vários outros critérios foram apresentados e podem ser atribuídos a 8 classes distintas (BORGES e LIMA, 2015). Para Mathews e Kanmani (2012), um dos critérios mais recentes estaria ligado ao que denominam como *"Temporal Information Retrieval"* em virtude da quantidade imensa de dados disponíveis na Internet e que são fortemente dependentes do tempo.

Com base nas técnicas de "Recuperação da Informação Temporal", Duchon *et al.* (2015) apresentam, sob forma de patente, um método de conversão de dados e fluxos de tópicos, combinados com métodos temporais, para prever objetivamente atividades de tópicos no futuro.

O objetivo geral desta pesquisa foi analisar se há uma variação temporal característica da distribuição de valores de termos relevantes ao longo do tempo da produção de textos que contribui como um critério para o processo de sua indexação

[1] Inicialmente, Luhn (1957) adotava terminologias como *auto-resumo* e *auto-indexação*. Posteriormente esses termos foram substituídos por indexação automática.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

automática.

Este artigo está organizado em 5 seções: Introdução, Fundamentação Teórica, Metodologia, Resultados e Conclusões.

2 FUNDAMENTAÇÃO TEÓRICA

A seguir são apresentadas cinco seções com as principais fundamentações teóricas utilizadas para essa pesquisa: sintagmas nominais; processamento de linguagem nominal; indexação automática; os critérios para pontuação de descritores e, por fim, recuperação da informação temporal.

2.1 Sintagmas nominais

Kuramoto (1999) apresentou em sua tese de doutorado uma das primeiras aplicações para computador utilizando o SN para recuperação de informação na língua portuguesa. Souza (2005), a partir desses estudos, propôs uma metodologia de escolha automática de SNs como descritores relevantes no processo de indexação automática. Maia (2008) desenvolveu uma ferramenta[2], a partir da metodologia de Souza (2005), que, dentre outras funcionalidades, extrai os SNs de forma automática. Mesquita *et al.* (2013, 2014), a partir da ferramenta de Maia (2008), verificaram comportamentos da pontuação de descritores relevantes em textos científicos.

Os SNs em um documento apresentam densidade informacional superior à palavras isoladas, mantendo maior proximidade do discurso contido nos documentos por eles descritos (KURAMOTO, 1996; SOUZA, 2005). “Palavras isoladas, como descritores, podem apresentar mais problemas de polissemia ou de plurisignificação” (LYONS, 1987, p.140). Além de apresentarem menos influência dos problemas acima, “os sintagmas nominais trazem em seu bojo o contexto semântico dos discursos” (SOUZA, 2005, p.136). Para Baeza-Yates e Ribeiro-Neto (2011) os substantivos, que compõem um SN, possuem maior valor semântico ao serem usados como termos de indexação. Portanto, o uso de SNs como termos de indexação pode apresentar melhores resultados que o uso de palavras isoladas.

Para Cintra (1983, p.9) um descritor pode ser analisado formalmente como um

[2] A ferramenta de Maia (2008) se chama Ogma e faz o processamento de linguagem natural para a língua portuguesa e com funcionalidade específica para a extração de sintagmas nominais.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

"sintagma de símbolos notacionais (números, letras, pontuação, marcas) isto é, unidades resultantes da combinação de formas menores em unidades de nível superior. Ex.: leit – eira ->• leiteira; o, vestido, verde, de, Lúcia ->* O vestido verde de Lúcia". Este último é um exemplo de sintagma nominal.

2.2 Processamento de linguagem natural

Cintra (1983, p.5) apresenta a importância da linguística para a indexação, que é definida por ela como:

> "a tradução de um documento em termos documentários, isto é, em descritores, cabeçalhos de assunto, termos-chave, que têm por função expressar o conteúdo do documento. A indexação assim definida é, pois; uma 'tradução lexical' das unidades lexicais da língua em que está escrito o documento, para unidades lexicais de uma linguagem documentária".

Existem várias ferramentas de processamento de linguagem natural, sendo algumas delas para a língua portuguesa. Dentre essas pode-se destacar o sistema Palavras de Bick (2000)[3], que é resultante de uma pesquisa de doutorado para a análise automática gramatical da língua portuguesa. O Palavras é o *parser* que permite o melhor desempenho na extração automática dos sintagmas nominais presentes em textos eletrônicos, quando comparado a outras ferramentas (SILVA & CORRÊA, 2017, p.15). Silva e Corrêa (*ibidem*) apontaram que o Palavras apresenta somente 6% de erro, enquanto outras ferramentas comparadas apresentaram em média 26% de erro. Além disso, o Palavras possui a vantagem de processar um texto de qualquer tamanho de uma única vez, enquanto outros *parsers* necessitam que se introduza frase por frase.

Em um PLN, um texto é analisado essencialmente por suas palavras. As fases do processamento são essencialmente: (1) conversão do documento em um formato de texto puro; (2) retirada opcional de quebras de linhas ; (3) análise léxica com o tratamento de acentuações, espaços, pontuações, números, hifens, etc.; (4) marcação estrutural, como títulos, por exemplo; e (5) retirada de palavras de baixa relevância (*stopwords*) contida em uma lista (*stoplist*).

[3] Esta ferramenta encontra-se disponível na Universidade Federal de Minas Gerais – UFMG em uma parceria entre a sua Escola de Ciência da Informação e sua Faculdade de Letras.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

2.3 Indexação automática

A indexação pode ser definida como:

> "[...] o processo de analisar o conteúdo informacional dos registros do conhecimento e sua expressão na linguagem do sistema de indexação. Ele implica: a) Selecionar os conceitos indexáveis de um documento; e b) Expressar esses conceitos na linguagem do sistema de indexação". (BORKO; BERNIER, 1978, p.8)

Além da inviabilidade do tratamento de grandes quantidades de documentos, os problemas práticos da atividade de indexação manual encontram-se também na inconsistência praticada pelos indexadores (DIAS; NAVES, 2007), que podem ser interindexadores e intraindexadores (BORKO, 1977). A inconsistência interindexadores ocorre quando dois ou mais indexadores elegem ou atribuem descritores diferentes para um mesmo documento. A inconsistência intraindexadores ocorre quando um mesmo indexador atribui descritores diferentes para um mesmo documento em momentos diferentes.

A indexação automática se justifica então pela sua capacidade de atender ao crescente volume de documentos eletrônicos e de forma mais consistente que a manual. As pesquisas em indexação automática ganharam força após a Segunda Guerra Mundial, quando o espírito pragmático e o apoio em pesquisa tecnológica dos Estados Unidos geraram um grande avanço, permitindo várias implementações (ORTEGA, 2004).

2.4 Critérios para pontuação de descritores

Conforme Sayão (1985), a indexação automática começou a ganhar notoriedade com as publicações de Luhn (1957). Muitos autores contribuíram para a evolução dessa área de pesquisa nas suas primeiras décadas: Baxendale (1958 *apud* SAYÃO, 1985), Swanson (1963 *apud* SAYÃO, 1985), Borko (1968 *apud* SAYÃO, 1985), Salton (1967, 1971a, 1971b *apud* SAYÃO, 1985), Salton e Lesk (1968 *apud* SAYÃO, 1985), Van Rijsbergen (1971 *apud* SAYÃO, 1985), Sparck Jones (1972), Sparck Jones (1973, 1978, 1979 *apud* SAYÃO, 1985), Field (1975, 1977 *apud* SAYÃO, 1985), Dillon (1982 *apud* SAYÃO, 1985), Robredo (1980, 1982a, 1982b *apud* SAYÃO, 1985) e outros. Atualmente

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

existe uma grande quantidade de critérios para a indexação automática, sendo que ainda prevalece aqueles apontados no início de sua história, como o uso da frequência de palavras isoladas.

Os conceitos básicos para os modelos de recuperação da informação surgem com Luhn (1957) assumindo a frequência do termo (*term frequency – TF*)[4] como critério para atribuição de pesos em um documento.

> **Definição**: *Frequência do Termo.* O valor, ou peso, de um termo k_i que ocorre em um documento d_j é simplesmente proporcional à frequência do termo $f_{i,j}$. Isto é, quanto mais o termo k_i ocorre em um texto do documento d_j, mais alto é seu peso por frequência de termo $TF_{i,j}$ (LUHN, 1957, tradução do autor).

Sparck Jones (1972) apresentou o conceito de *especificidade do termo* que foi denominado como frequência inversa do documento e se baseou nas noções de exaustividade e especificidade dos termos.

> **Definição**: *Exaustividade e Especificidade*. Exaustividade é uma propriedade de descrição do documento, especificidade é uma propriedade dos termos de indexação. A exaustividade da descrição do documento é interpretada como a sua cobertura para os principais tópicos do documento. A especificidade de um termo de indexação é interpretada como o quão bem o termo descreve um tópico do documento (BAEZA-YATES; RIBEIRO-NETO, 2011, p.70, tradução livre).

O nível de exaustividade adotado é considerado como a principal decisão da política de indexação e vai determinar estatisticamente a quantidade de termos de indexação usada em média para cada documento. Uma indexação exaustiva elege/atribui termos de indexação para todos os assuntos de um documento, por outro lado, a indexação seletiva elege/atribui uma quantidade limitada de termos de modo a representar somente os assuntos principais de um documento (LANCASTER, 2004).

A *exaustividade ótima* considera que o número de termos de indexação deva ser otimizado de modo que a probabilidade de relevância do documento recuperado seja maximizada (BAEZA-YATES; RIBEIRO-NETO, 2011). Ou seja, para uma provável consulta, a quantidade de termos de indexação deve possibilitar uma máxima recuperação de documentos considerados relevantes por um usuário.

[4] Zipf (1932) caracterizou a ordenação decrescente das frequências dos termos de um documento como uma função exponencial. Logo, para obter um peso com variação linear em função da frequência, pode ser usada uma escala logarítmica da frequência de cada termo. Esse mesmo recurso matemático pode ser usado para o cálculo dos pesos relacionados à especificidade.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

A especificidade é a propriedade semântica do termo que depende do seu significado. Por exemplo, *moradia* é menos específico que *casa* ou *apartamento*. A especificidade pode ser ainda definida através da estatística em substituição da propriedade semântica do termo de indexação. Ou seja, o valor de especificidade de um termo pode ser calculado através do inverso da quantidade de documentos nos quais ele ocorre. Se um termo ocorre em todos os documentos, sua especificidade é baixa ou nula.

Baeza-Yates e Ribeiro-Neto (2011) apresentam três recomendações[5] de equações para o cálculo de pesos para termos em um documento. No Quadro 1, as três equações utilizam as seguintes expressões:

- $f_{i,j}$ → frequência do termo *i* no documento *j* (TF);
- N/n_i → número total de documentos dividido pelo número de documentos nos quais ocorre o termo *i* ao menos uma vez (especificidade ou IDF).

Quadro 1 – Recomendações de equações para o cálculo de pesos de termos.

Peso do termo em um documento
$f_{i,j} \log N/n_i$
$1 + \log f_{i,j}$
$(1 + \log f_{i,j}) \log N/n_i$

Fonte: Adaptado de BAEZA-YATES; RIBEIRO-NETO, 2011, p. 74.

A última equação acima foi utilizada na seção metodologia dessa pesquisa.

2.5 Recuperação da Informação Temporal

Bolour (1982) apresenta que Mattison (1967 *apud* Bolour, 1982) foi um dos primeiros a dar uma significativa contribuição para o critério temporal em processamento de informação. Até então, o critério temporal era usado essencialmente de forma binária, ao apontar se um valor estaria dentro de um intervalo de tempo ou não. Mattison (*ibidem*) apresentou uma função matemática, f(t)[T], na qual t seria um termo e T seria o ano, possibilitando verificar se um termo correria em vários momentos.

Moulahi *et al.* (2015) apresenta que na Recuperação da Informação Temporal há dois recursos temporais principais: marca de data / hora do documento e a relevância temporal do conteúdo do documento. A classificação temporal pode ser usada, por

[5] A terceira recomendação utilizada nesta pesquisa e é apresentada no capítulo sobre a metodologia.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

exemplo, para especificar a natureza da consulta. Mathews e Kanmani (2012), ao analisar 9 métodos distintos em Recuperação da Informação Temporal, afirmam que a inserção do critério temporal aumenta o desempenho dos sistemas de recuperação da informação.

Duchon *et al.* (2015, pg.7) descrevem um método que armazena o histórico de acesso a tópicos de um usuário, com base nesse histórico, e suas respectivas informações temporais, um modelo é gerado com base na informação de vários usuários. A partir do modelo criado, para cada comportamento de uso de acesso a tópicos numa determinada sequência e intervalos temporais, um usuário pode receber sugestões de novos tópicos que ele possa desejar no futuro.

3 METODOLOGIA

Em virtude da necessidade de um *corpus* com textos que caracterizassem um aspecto temporal, buscou-se por teses de doutorado, como textos concebidos por um período de tempo relativamente mais longo (cerca de 4 anos) e acessíveis digitalmente. Foi escolhido o Repositório Institucional da UFMG (RI-UFMG), uma vez que o mesmo, por se tratar da mesma instituição na qual foi desenvolvida essa pesquisa, permitiria avaliar mais informações transversais nas análises de dados.

Para uma tese, que “possui aproximadamente entre cem e quatrocentas páginas relacionadas a uma área de estudos” (ECO, 2007, p.27), acredita-se que o seu tempo de desenvolvimento, em virtude de um vínculo institucional de cerca de quatro anos pode favorecer ao estudo da distribuição temporal dos SNs como descritores. Essa hipótese é baseada nos seguintes aspectos: as repetições de um mesmo SN tendem a aumentar conforme o crescimento da quantidade de palavras em um texto que trata de uma mesma área; com uma quantidade maior de repetições de um mesmo sintagma, pode-se avaliar com mais detalhes suas variações da distribuição ao longo do tempo em que as teses são defendidas.

A escolha aqui de teses como elementos de pesquisa implica em maior *custo computacional* de processamento da extração dos SNs, em comparação a artigos, uma vez que estes últimos, geralmente, possuem um tamanho da ordem de dez vezes menor. No entanto, com o desempenho dos recursos computacionais atuais em relação aos mais

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

antigos[6] usados em outras pesquisas, que se basearam em artigos, o processamento de teses mostrou-se viável.

Além do recorte do tipo de documento, também, para essa pesquisa, foi definido um recorte temporal de um total de 12 anos e relacionados ao período contínuo de 2007 a 2018. O ano inicial de 2007 está relacionado a uma portaria da CAPES de 2006 que passa a regular que teses e dissertações sejam disponibilizadas através de repositório digital. O limite superior do ano de 2018 foi utilizado pois os dados foram coletados em 2019, sendo que as publicações deste mesmo ano ainda estavam sendo publicadas.

Ainda para justificar o recorte, foram considerados somente os PPGs que tiveram ao menos 12 teses no referido recorte temporal anterior. Essa limitação mínima quantitativa é referente a uma possibilidade de análise de dados considerando ao menos uma tese por cada um dos 12 intervalos temporais.

Para essa pesquisa foram analisadas 641 teses em 7 PPGs distintos. A Tabela 1, a seguir, apresenta o quantitativo total por PPG, sendo cada um deles considerado como um *corpus* para esta pesquisa.

Tabela 1 – Quantidade de teses analisadas por PPG (*corpora*).

Nº	Programa	Teses
1	Pós-Graduação em Administração	159
2	Pós-Graduação em Arquitetura e Urbanismo	51
3	Pós-Graduação em Ciência da Informação	125
4	Pós-Graduação em Comunicação Social	40
5	Pós-Graduação em Demografia	85
6	Pós-Graduação em Direito	98
7	Pós-Graduação em Economia	83
	Total Geral	**641**

Fonte: Elaborados pelos autores.

A metodologia consistiu essencialmente em três etapas, cujas partes são descritas a seguir:

1. Extração dos SNs:
   1.1. Obtenção dos documentos originais (em formato PDF);
   1.2. Conversão dos documentos originais para o formato texto (TXT);
   1.3. Etiquetagem e extração dos SNs usando o Palavras ;

[6] Souza (2005) utilizou um computador com processador AMD Athlon XP 2600+ com 256MB de memória RAM. Dentre os recursos computacionais usados para essa pesquisa, um deles foi uma *instância* de uma máquina virtual na *Google Cloud Plataform* com 16 núcleos de processamento e 16GB de RAM.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

1.4. Tratamento dos SNs (retirada das partes determinantes[7], e a retirada de SNs de acordo com uma *stoplist*);

2. Seleção e pontuação dos SNs candidatos a descritores em cada tese:
   2.1. Processamento do atributo de frequência dos SNs por documento (*TF*);
   2.2. Processamento do atributo de especificidade dos SNs por *corpus* (*IDF*);
   2.3. Pontuação ($P_{Tese-SN}$) dos SNs candidatos a descritores (através do *TF* e *IDF*) e utilizando a terceira fórmula do Quadro 1;
3. Distribuição da pontuação dos SNs candidatos a descritores:
   3.1. Consolidação da pontuação dos SNs: nessa etapa, todo SN em cada tese recebeu uma pontuação $P_{Tese-SN}$ de acordo com seu TF/IDF. Em cada tese, a pontuação $P_{Tese-SN}$ de seus SNs foram somadas ($P_{Tese-Soma-SNs}$).
   3.2. Normalização da pontuação dos SNs: para evitar que teses com maior número de páginas gerassem um viés nos dados, o valor de $P_{Tese-Soma-SNs}$ de cada tese foi reduzido a 1 dividindo-se $P_{Tese-Soma-SNs}$ por ele mesmo. Logo, cada pontuação de cada SN na tese, $P_{Tese-SN,}$ também foi dividido por $P_{Tese-Soma-SNs}$, resultando na pontuação do sintagma nominal normalizada, $P_{Tese-SN-Normalizada.}$
   3.3. Pontuação do SN no PPG: após a normalização, cada SN recebeu uma pontuação geral para o PPG somando-se as respectivas pontuações normalizadas em cada tese, $P_{Tese-SN-Normalizada}$, e resultando na sua pontuação geral no PPG, P_{PPG-SN}.
   3.4. Distribuição da pontuação geral de cada SN no PPG: a pontuação geral no PPG de cada SN, $P_{PPG-SN,}$ foi redistribuída de acordo com o TF de cada SN em cada tese, resultando na $P_{Tese-SN-Normalizada.}$
   3.5. Consolidação final por tese no PPG: após a redistribuição da pontuação normalizada de cada SN, $P_{Tese-SN-Normalizada}$, foi possível calcular a pontuação acumulada em cada tese somando-se esses mesmos valores em cada tese, $P_{PPG-Tese.}$
   3.6. Consolidação por dimensão temporal: cada tese foi posicionada por ordem da data de sua defesa. As teses foram divididas igualmente (exceto por

[7] Os determinantes são os artigos, os pronomes possessivos, os pronomes demonstrativos, os adjetivos interrogativos, relativos e indefinidos e, ainda, os numerais que constituem o SN e dependem do substantivo, cabeça ou constituinte principal do SN (DUBOIS *et al*, 2013, pg.180).

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

necessidade de arredondamento) em 12 partes[8]. Cada uma das 12 fatias de tempo recebeu uma data calculada pela média das datas de defesas das suas respectivas teses, assim como recebeu uma média das pontuações obtidas em cada tese, $P_{PPG\text{-}Tese}$.

3.7. Caracterização das distribuições de pontuações por funções matemáticas: foram utilizadas técnicas de regressão polinomial até no máximo o grau 6, para caracterizar o comportamento da distribuição temporal dos valores de pontuações médias em cada PPG.

A seguir são apresentados os gráficos de comportamento da distribuição de pontuações por dispersão temporal, assim como suas equações de regressão polinomial e a análise desses resultados.

4 RESULTADOS

A seguir são apresentados os valores de pontuação para cada PPG, sendo estes distribuídos por dispersão com base nas datas médias em cada uma das suas 12 fatias temporais.

Em cada gráfico é possível observar que cada ponto representa uma fatia temporal de uma sequência de N/12, onde N é o total de teses no PPG. Cada uma das 12 fatias tem, portanto, uma quantidade igual de teses, exceto quando a divisão não é inteira e foi então arredondada. Para este caso a diferença entre um grupo e outro, quando houve, foi de uma tese somente.

Horizontalmente, a proximidade de pontos representa que houve uma maior densidade de defesas de teses em tal momento. Um maior espaçamento, significa o contrário.

Para sistematizar a análise dos gráficos, foram utilizadas regressões polinomiais com seu respectivo R^2, que representa, numa variação de 0 a 1, o quanto a regressão passa próxima aos pontos originais (1 para o melhor caso). Usualmente considera-se acima de 0,95 uma boa regressão, embora esse valor possa esconder outras possíveis regressões, como a exponencial, por exemplo.

A seguir são apresentadas duas seções, sendo a primeira uma análise detalhada individual de cada PPG e a segunda análise geral das Ciências Sociais Aplicadas.

[8] A divisão em 12 partes é referente ao total de anos, de 2007 a 2018, do recorte usado na pesquisa.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

4.1 Análise de distribuição temporal por PPG por regressão polinomial

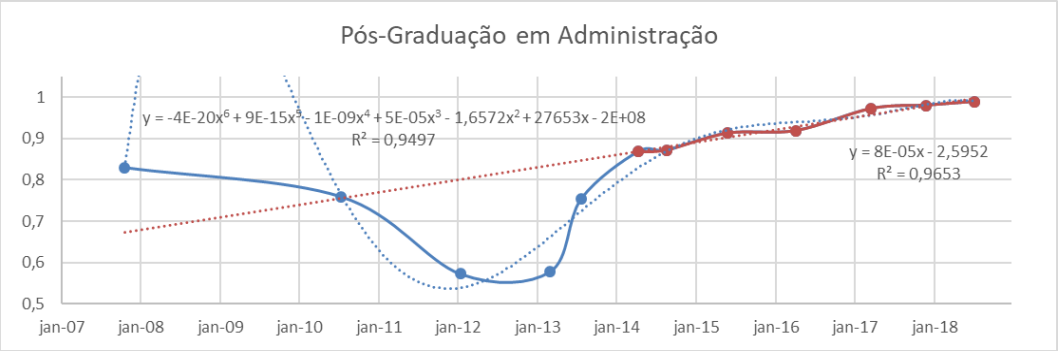
As curvas de cada PPG com suas respectivas linhas de tendência, obtidas por regressão linear ou polinomial, são apresentadas a seguir. Para cada PPG foi testada uma regressão linear simples e cinco polinomiais (variando de grau 2 ao 6).

Para cada PPG foi testada à regressão mais simples e que obteve o R² satisfatório para as condições de análise aqui empregadas. A seguir, juntamente com os gráficos, são apresentadas para cada PPG da grande área da Ciências Sociais Aplicadas as regressões escolhidas com base nesses critérios.

- **Pós-Graduação em Administração**

Esse foi um dos PPGs que mais apresentou ondulações em relação à relevância Numa regressão linear simples, o seu R² apresentou o mais baixo coeficiente, significando que uma linha reta conseguiria representá-lo de forma insuficiente. Foram testadas novas regressões, e somente na polinomial de grau 6 a tendência foi satisfatória.

Gráfico 1 – Pontuação X Tempo para PPG em Administração com linha de tendência.



Fonte: Elaborado pelos autores.

Assim como o PPG em Demografia, a ser analisado adiante, é possível perceber claramente que as oscilações foram significativas somente até um momento, especificamente até 2013. A partir de 2014 é possível perceber uma estabilidade até o final. Tal estabilidade apresentou uma regressão linear simples[9] com R² 0,965. Podemos inferir que o PPG em Administração teve notoriamente duas fases, sendo a última delas de clara estabilidade de crescimento nas pontuações médias.

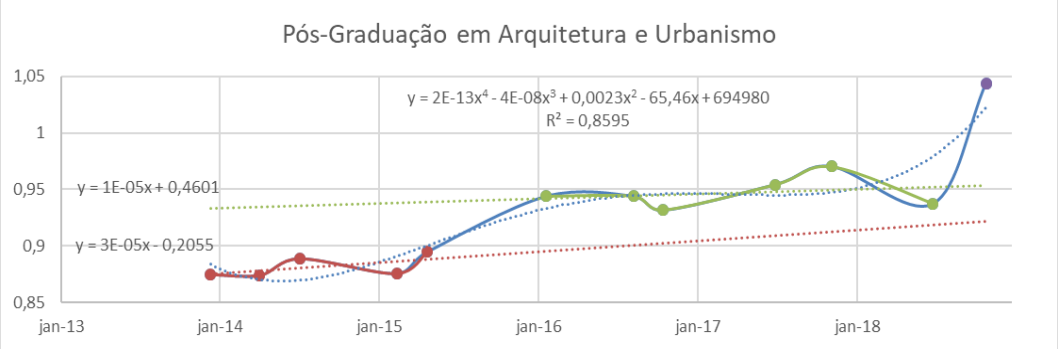
[9] Embora o ponto correspondente ao ano de 2010 não tenha sido usado para a regressão linear de 2014 a 2018, o mesmo pode ser percebido no gráfico como se fizera parte de tal conjunto.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

- **Pós-Graduação em Arquitetura e Urbanismo**

Por ser um PPG novo, este apresentou um comportamento com bastante ondulações. Ao se analisar as regressões polinomiais, percebeu-se um aumento significativo quando foi testado o grau 4. Isso reflete que esse PPG poderia ser analisado em três diferentes momentos na sua história, pois a quantidade de inflexões é o grau do polinômio subtraído de 1.

Gráfico 2 – Pontuação X Tempo para PPG em Arquitetura e Urbanismo e linha de tendência.



Fonte: Elaborado pelos autores.

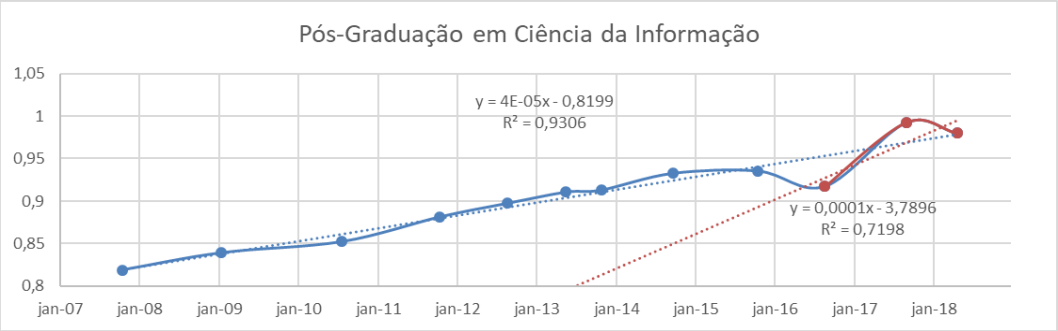
Para esse PPG é possível perceber notoriamente três fases: até início de 2015, até aproximadamente julho de 2018 e a última, com somente um ponto, em outubro de 2018. Como trabalhos futuros para esse PPG especificamente, recomenda-se analisar esses três momentos de modo a avaliar os principais fatores do contexto que influenciaram tal resultado.

- **Pós-Graduação em Ciência da Informação**

O PPG em Ciência da Informação foi o único que permitiu uma análise por pura regressão linear, com um R^2 de 0,93. As demais regressões apresentaram incrementos no R^2 pouco expressivos. Logo, para esse PPG, foi considerado o R^2 um pouco abaixo de 0,95, como satisfatório.

Gráfico 3 – Pontuação X Tempo para PPG em CI com linha de tendência.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395



Fonte: Elaborado pelos autores.

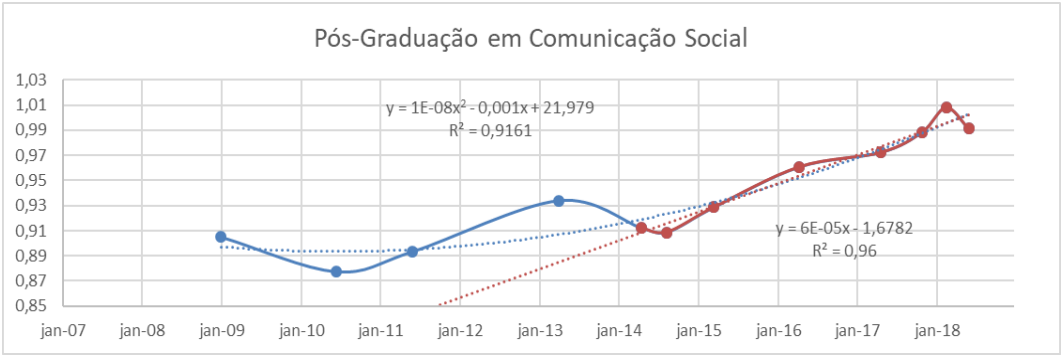
Para esse PPG, embora tenha sido caracterizado aqui de forma linear, é possível perceber notoriamente uma ruptura em nos últimos 3 pontos com os seguintes valores: 25/08/2016 com 0,917; 04/09/2017 com 0,992; e 29/04/2018 com 0,980. A Escola de Ciência da Informação até 2015 tinha somente um PPG (o de Ciência da Informação), quando foi criado então o PPG em Gestão e Organização do Conhecimento. Destaca-se que parte de seu corpo discente e docente advindos por um processo de absorção a partir do PPG em Ciência da Informação. Esse fato, pode ter contribuído para a significativa oscilação vista nos anos de 2016 a 2018, nos quais o universo terminológico apresenta um comportamento de maior especificidade. Como trabalhos futuros para esse PPG especificamente, recomenda-se analisar esses três momentos finais de modo a avaliar os principais fatores de contexto que possam o ter influenciado.

- **Pós-Graduação em Comunicação Social**

O PPG em Comunicação Social apresentou um salto significativo quando foi testada uma regressão polinomial de segundo grau, saltando de 0,800 para 0,916. As demais regressões apresentaram incrementos no R² pouco expressivos. Logo, para esse PPG, foi considerado também o R² um pouco abaixo de 0,95 como satisfatório.

Gráfico 4 – Pontuação X Tempo para PPG em Comunicação Social com linha de tendência.

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395



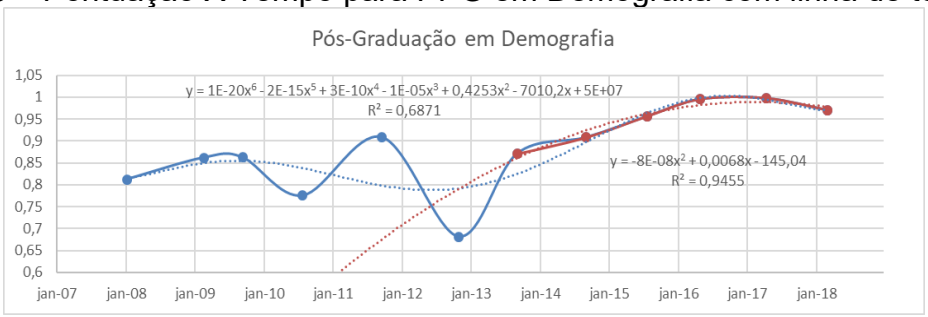
Fonte: Elaborado pelos autores.

Para esse PPG, percebeu-se uma variação significativa na produtividade quantitativa de teses a partir de 2014. Aa partir desse ano apresenta um crescimento linear, cuja curva de regressão apresentou R² de 0,96. Os seus últimos três momentos temporais são bastante densos e com oscilação. Como trabalhos futuros para esse PPG especificamente, recomenda-se analisar um possível novo contexto em 2014, assim como as temáticas em 2018.

- **Pós-Graduação em Demografia**

O PPG em Demografia apresentou um dos resultados com maior quantidade de oscilações. Mesmo com regressão polinomial de grau 6, o seu R² chegou somente a 0,687. No entanto é possível perceber claramente que tais oscilações foram significativas até 2012.

Gráfico 5 – Pontuação X Tempo para PPG em Demografia com linha de tendência.



Fonte: Elaborado pelos autores.

Para esse PPG, é possível perceber uma estabilidade do final de 2013 em diante. Tal estabilidade apresentou uma característica de uma suave ondulação, com R² de 0,946, que atinge seu máximo ao final de 2016 e, a partir de então, começa a declinar levemente. Como trabalhos futuros para esse PPG especificamente, recomenda-se

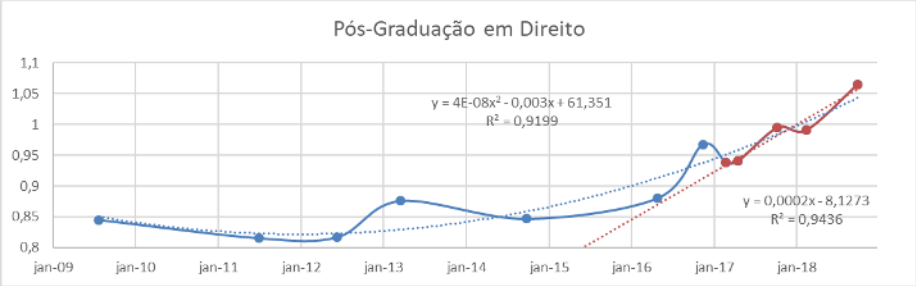
Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

analisar as três faixas temporais mencionadas, com atenção maior aos anos de 2011 e 2012 nos quais há maior amplitude de variação.

- **Pós-Graduação em Direito**

Assim como no PPG em Comunicação Social, o PPG em Direito apresentou um salto significativo quando foi testada uma regressão polinomial de segundo grau, saltando de 0,741 para 0,920. As demais regressões apresentaram incrementos no R^2 pouco expressivos. Logo, para esse PPG, foi considerado também o R^2 um pouco abaixo de 0,95 como satisfatório.

Gráfico 6 – Pontuação X Tempo para PPG em Direito com linha de tendência.



Fonte: Elaborado pelos autores.

Para esse PPG, é possível perceber uma oscilação no início de 2013 e um salto evolutivo a partir do ano de 2017, cuja regressão linear apresentou um R^2 de 0,944. Esse mesmo período é marcado por uma produtividade quantitativa bastante expressiva. Observou-se uma grande aproximação de duas fatias temporais, o que foi bastante incomum em comparação aos outros PPGs. Ela ocorreu precisamente nas fatias 8 e 9 com suas respectivas médias de datas em 28/02/2017 e 24/04/2017. Como trabalhos futuros para esse PPG, especificamente, recomenda-se analisar a oscilação no ponto 4, na data média de 20/03/2013, e os anos de 2017 e 2018.

- **Pós-Graduação em Economia**

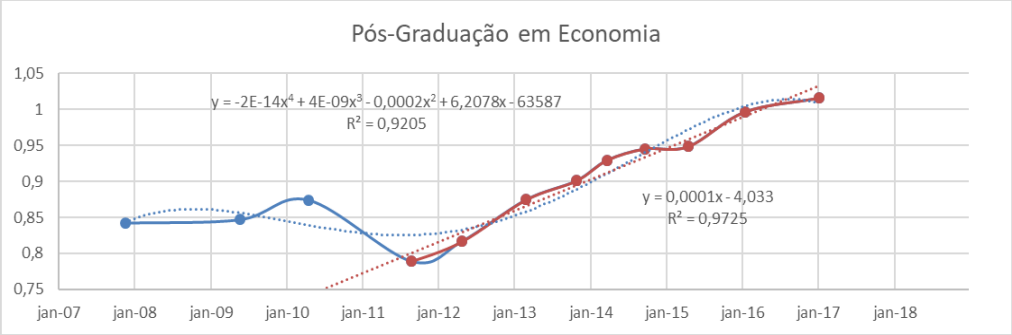
Assim como nos PPGs em Comunicação Social e em Direito, o PPG em Economia apresentou um salto significativo quando foi testada uma regressão polinomial de segundo grau, saltando de 0,658 para 0,854. E também apresentou um salto significativo quando foi testada uma regressão polinomial de quarto grau, saltando de 0,854 para 0,921. As demais regressões apresentaram incrementos no R^2 pouco expressivos. Logo, para esse PPG, foi considerado também o R^2 um pouco abaixo de 0,95 como satisfatório.

Para esse PPG, é possível perceber dois momentos: o ano de 2010 registra uma

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

nova fase que, a partir de então, apresenta um crescimento linear com R² de 0,973, com uma baixa de produtividade quantitativa no ano de 2015. Como trabalhos futuros para esse PPG especificamente, recomenda-se analisar o contexto de 2010/2011 e 2015.

Gráfico 7 – Pontuação X Tempo para PPG em Economia com linha de tendência.

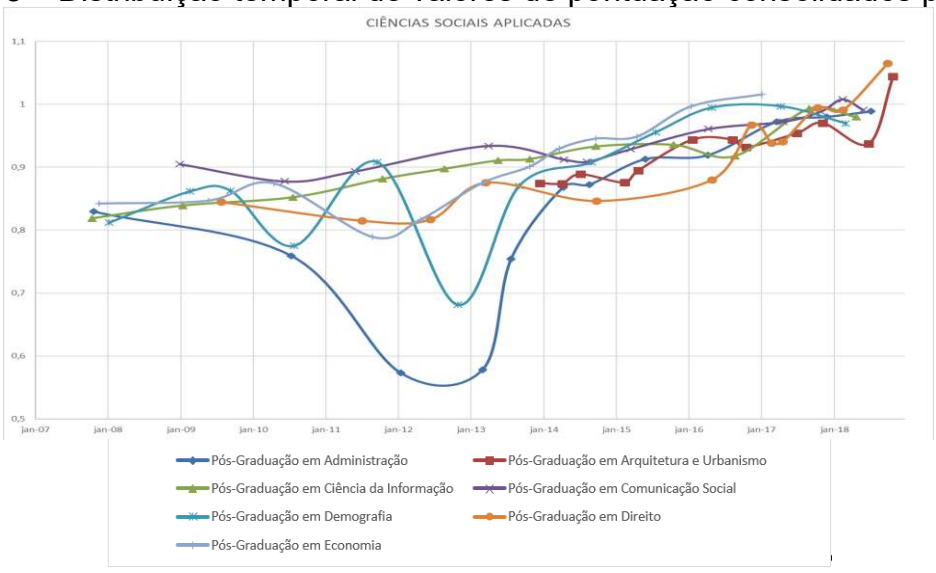


Fonte: Elaborado pelos autores.

4.2 Análise geral de distribuição temporal por PPG

A mais significativa contribuição dessa pesquisa está na análise da variação da amplitude em cada gráfico. É possível perceber uma tendência de crescimento para quase todos os PPGs. Destaca-se que esse fato pode ser considerado como uma possível evolução de cada PPG ao adotar novas terminologias. Somado a isso, há o uso das terminologias pregressas. Cabe para trabalhos futuros, e em cada PPG, uma análise das terminologias específicas e suas ocorrências ao longo do tempo. Todos os gráficos apresentados podem ser vistos reunidos no Gráfico 8.

Gráfico 8 – Distribuição temporal de valores de pontuação consolidados por PPG.

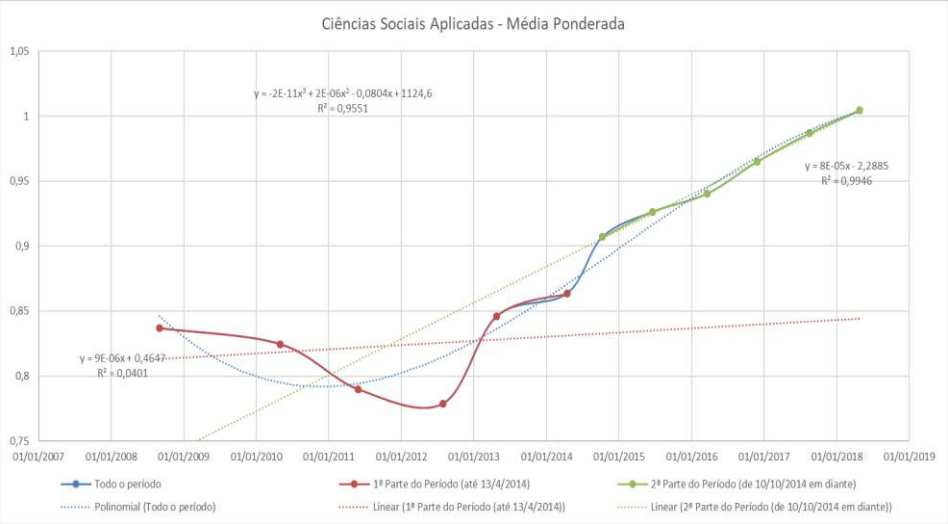


Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

Fonte: Elaborados pelos autores.

Realizando uma média ponderada em cada uma das 12 fatias temporais dos 7 PPGs das Ciências Sociais Aplicadas aqui analisados, é possível perceber claramente que o mesmo apresenta uma tendência estável de crescimento de suas pontuações a partir do segundo semestre de 2014, com uma linha de tendência linear com R^2 de 0,995.

Gráfico 9 – Média da Pontuação X Tempo entre todos PPGs das Ciências Sociais Aplicadas.



Fonte: Elaborados pelos autores.

A qualidade desse último resultado foi muito além das expectativas, pois dentre todas as regressões, a linear é a mais simples. E somado a isso, um R^2 de 0,995 é excelente, mesmos para somente 6 pontos (2º semestre de 2014 em diante).

Conforme as Normas Gerais de Pós-Graduação da UFMG, "o Doutorado tem por objetivo desenvolver a capacidade de propor e conduzir, de forma autônoma, pesquisas originais em área específica ou interdisciplinar do conhecimento". Uma vez que parte da pontuação dos termos ocorre pela sua especificidade, podemos inferir que o crescimento ocorre pela adequação condução de "pesquisas originais" nas Ciências Sociais Aplicadas da UFMG, sobretudo a partir de 2014 em diante.

5 CONSIDERAÇÕES FINAIS

O critério temporal se mostrou bastante efetivo para ser considerado na pontuação de descritores, que, por sua vez, pode ser aplicada na indexação automática. Podemos concluir que, a partir de 2014, foi possível encontrar uma variação temporal característica da distribuição de valores de termos relevantes ao longo do tempo da produção de textos

Biblos: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande
v. 34, n. 02, p. 344-365, jul./dez. 2020. ISSN 2236-7594. DOI: https://doi.org/10.14295/biblos.v34i2.12395

que contribui como um critério para o processo de sua indexação automática.

A metodologia descrita aqui, com seus 14 passos, apresenta-se atualmente como inédita na literatura, sendo fruto parcial de uma pesquisa de doutorado em andamento. Igualmente a esse recorte, os autores já possuem dados para analisar as outras 8 áreas de conhecimento da UFMG: Ciências da Saúde; Ciências Humanas; Linguística, Letras e Arte; Ciências Exatas e da Terra; Engenharias; Ciências Biológicas; Ciências Agrárias e a denominada como Multidisciplinar. Espera-se publicar tais resultados em outras pesquisas como essa, assim como um comparativo entre todas elas.

REFERÊNCIAS

ARAÚJO, Ronaldo Ferreira; ALVARENGA, Lidia. A bibliometria na pesquisa científica da pós-graduação brasileira de 1987 a 2007. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 16, n. 31, p. 51-70, 2011.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: ACM Press, 1999. 511p.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**: the concepts and technology behind search. 2. ed. London: Pearson Education Limited, 2011. 913 p.

BAXENDALE, P. B. Machine-made index for technical literature: an experiment. **IBM Journal of Research and Development**, v. 2, n. 4, p. 354-361, 1958.

BICK, E. **The Parsing System Palavras***: a*utomatic grammatical analysis of portuguese in a constraint grammar framework. Aarhus: Aarhus University Press, 2000.

BOLOUR, Azad et al. The role of time in information processing: a survey. **ACM SIGART Bulletin**, n. 80, p. 28-46, 1982.

BORGES, Graciane Silva Bruzinga; LIMA, Gercina Ângela Borém de. O. Desenvolvimento de softwares de indexação automática: breve avaliação dos principais critérios. **Informação & Tecnologia**, v. 2, n. 2, 2015. Disponível em: https://periodicos.ufpb.br/index.php/itec/article/view/33926 Acesso em: 05 jan. 2019.

BORKO, H. Information science: what is it? **American Documentation**, v. 19, n. 1, p. 3-5, jan. 1968.

BORKO, H.; BERNIER, C. **Indexing concepts and methods**. New York: Academic Press. 1978.

BUSH, Vannevar et al. As we may think. **The atlantic monthly**, v. 176, n. 1, p. 101-108, 1945.

CINTRA, Anna Maria Marques. Elementos de linguística para estudos de indexação. **Ciência da informação**, v. 12, n. 1, 1983.

DEVARAKONDA, S. et al. Viewing computer science through citation analysis: Salton and Bergmark Redux. **Scientometrics**, v. 125, n. 1, p. 271-287, 2020.

DIAS, E. W.; NAVES, M. M. L. **Análise de assunto**: teoria e prática. Brasília: Thesaurus, 2007. 116 p.

DILLON, M. Thesaurus-based automatic book indexing. **Information Processing & Management**, v. 18, n. 4, p. 167-78, 1982.

DUBOIS, J. *et al*. **Dicionário de lingüística**. São Paulo: Cultrix, 1973. 657p.

DUCHON, Andrew P. et al. **Method and system to predict the likelihood of topics**. U.S. Patent n. 9,165,254, 20 out. 2015.

ECO, U. **Como se faz uma tese em ciências humanas**. 13. ed. Lisboa: Presença, 2007. 238 p.

FIELD, B. J. Towards automatic indexing: automatic assignment of controlled-language indexing and classification from free indexing. **Journal of Documentation**, v. 31, n. 4, 1975.

KURAMOTO, H. **Proposition d'un système de recherche d'Information assistée par ordinateur avec application à la langue portugaise**. 1999. Tese (Doutorado em Ciências da Informação e da Comunicação) – Université Lumière Lyon 2, Paris, França, 1999.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.

LUHN, H. P. A statistical approach to mechanized enconding and searching of literature information. **IBM Journal of Research and Development**, v. 1, n. 4, p. 309-317, oct. 1957.

LYONS, J. **Linguagem e Lingüística**: uma introdução. Rio de Janeiro: Livros Tecnicos e Científicos, 1987. 322 p.

MAIA, Luiz Cláudio Gomes; SOUZA, Renato Rocha; , . Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**, [S.l.], v. 15, n. 1, p. 154-172, mar. 2010. ISSN 19815344. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/875. Acesso em: 05 jan. 2020.

MATHEWS, Litty K.; KANMANI, S. Deepa. A survey on temporal information retrieval systems. **International Journal of Computer Applications**, v. 58, n. 4, 2012.

MATTISON, Robert. **A formal system for the logical analysis of temporal relationships between intervals of time**. RAND CORP SANTA MONICA CALIF, 1967.

MESQUITA, Luiz Antônio Lopes; SOUZA, Renato Rocha; PORTO, Renata Maria Abrantes Baracho. Características de Teses de oito áreas de conhecimento: uma análise para o desempenho de indexação automática através de sintagmas nominais. In.: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Santa Catarina. **Anais...** Florianópolis, SC, 2013.

MESQUITA, Luiz Antônio Lopes; SOUZA, Renato Rocha; PORTO, Renata Maria Abrantes Baracho. *Noun phrases in automatic indexing: A structural analysis of the distribution of relevant terms in doctoral theses*. Advances In KNOWLEDGE ORGANIZATION, v. 14, 327-34. 2014. Polônia. **Anais...** Cracóvia, 2014.

MOOERS, Calvin N. Zatocoding applied to mechanical organization of knowledge. **American documentation**, v. 2, n. 1, p. 20-32, 1951.

MOULAHI, Bilel; TAMINE, Lynda; YAHIA, Sadok Ben. When time meets information retrieval: Past proposals, current plans and future trends. **Journal of Information Science**, v. 42, n. 6, p. 725-747, 2016.

ORTEGA, C. D. Relações históricas entre Biblioteconomia, Documentação e Ciência da Informação. **DataGramaZero**, v. 5, n. 5, out. 2004. Disponível em: http://www.dgz.org.br/out04/Art_03.htm Acesso em 21 jan. 2020.

PERINI, M. A. *et al*. O SN em português: a hipótese mórfica. **Revista de Estudos de Linguagem – UFMG**, Belo Horizonte, p. 43-56, jul./dez. 1996.

ROBREDO, J. A. Indexação automática como mecanismo básico no processo de transferência da informação. In: CONGRESSO LATINO-AMERICANO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 1., Salvador, 1980. **Anais...** Salvador: FEBAB, 1980, 19 p.

ROBREDO, J. A. Indexação automática de textos: o presente já entrou no futuro. In: Machado, U. D. (Ed). **Estudos avançados em Biblioteconomia e Ciência da Informação**, Brasília: ABDF, 1982b. p. 236-74

ROBREDO, J. A. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico. **Ciência da Informação**, v. 11, n. 1, p. 3-18, 1982a. Disponível em: http://revista.ibict.br/ciinf/article/view/175. Acesso em 21 fev. 2020.

SALTON, G. Automatic indexing using bibliographic citations. **Journal of Documentation**, v. 27, n. 2, p. 98-110, jun. 1971a.

SALTON, G. Designing automatic information system; results obtained with the SMART programs. *Social Science Information.* Vol. 6(2):111-17, Feb 1967.

SALTON, G. **The SMART retrieval systems**: experiments in automatic document processing. New York: Prentice-Hall, Englewood Cliffs, 1971b.

SALTON, G.; LESK, M. E. Computer evaluation of indexing and text processing. **Journal of the ACM**, v. 15, n. 1, p. 8-36, jan. 1968.

SALTON, Gerard; BERGMARK, Donna. A citation study of computer science literature. **IEEE Transactions on Professional Communication**, n. 3, p. 146-158, 1979.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação.** Belo Horizonte, v.1, n.1, p. 41-62, jan./jun. 1996. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/235. Acesso em 21 jan. 2020.

SAYÃO, L. F. **SALF**: um algoritmo para indexação automática utilizando vocabulário controlado. 1985. 177 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, IBICT, Rio de Janeiro, 1985.

SILVA, Tiago José da; CORRÊA, Renato Fernandes. Ferramentas para indexação automática: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX-Parser e a extração manual de sintagmas nominais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 18., 2017, Marília. **Anais**... Marília: UNESP, 2017.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. 197 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

SPARCK JONES, K. A statistical interpretation of term specificity and its application to retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11-20, 1972.

SPARCK JONES, K. Collection properties influencing automatic term classification performance. **Information Storage and Retrieval**, v. 9, p. 499-513, 1973.

SPARCK JONES, K. Experiments in relevance weighting of search terms. **Information Processing & Management**, v. 15, n. 13, p. 133-144, 1979.

SPARCK JONES, K. The role of automatic indexing in oprational on-line retrieval systems. In: FID Congres, 38, Edinburg, 1978. **Proceedings…** London: ASLIB, 1980, p. 33-38

SWANSON, D. R. Automation indexing and classification. In: Nato Advanced Study Institute on Automatic Analysis, 1963, Venice. **Proceedings…** New York: [s.n.], 1963. p. 125-128.

VAN RIJSBERGEN, C. J. **Information Retrieval**. London: Butterwords, 1979.

ZIPF, G. K. **Selected studies of the principle of relative frequency in language**. Cambridge, USA: Havard University Press, 1932.