UNIVERSIDADE DE BRASÍLIA

FACULDADE DE CIÊNCIA DA INFORMAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

CARLOS JACOBINO LIMA

DESCOBERTA DE CONHECIMENTO NO ACERVO DOCUMENTAL DO PRÊMIO PROFESSOR SAMUEL BENCHIMOL: PROSPECÇÃO E ANÁLISE DE INFORMAÇÕES SOBRE A REGIÃO AMAZÔNICA DE 2004 A 2015

Brasília

2016

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE CIÊNCIA DA INFORMAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

CARLOS JACOBINO LIMA

DESCOBERTA DE CONHECIMENTO NO ACERVO DOCUMENTAL DO PRÊMIO PROFESSOR SAMUEL BENCHIMOL: PROSPECÇÃO E ANÁLISE DE INFORMAÇÕES SOBRE A REGIÃO AMAZÔNICA DE 2004 A 2015

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de Mestre em Ciência da Informação.

Linha de pesquisa: Organização da Informação e do Conhecimento

Professora Orientadora: Dra. Lillian Maria Araújo de Rezende Alvares

Brasília

2016

FOLHA DE APROVAÇÃO

Título: "Descoberta de conhecimento no acervo documental do Prêmio Professor Samuel Benchimol: prospecção e análise de informações sobre a região amazônica de 2004 a 2015".

Autor (a): Carlos Jacobino Lima

Área de concentração: Gestão da Informação

Linha de pesquisa: Organização da Informação

Dissertação submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade em Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Mestre** em Ciência da Informação.

Brasília, 28 de julho de 2016.

Profª Drª Lillian Maria Araújo de Rezende Alvares
Presidente (UnB/PPGCINF)

Prof. Dr. Paulo Rogério Foina
Membro Titular (UNICEUB)

Prof. Dr. Renato Tarciso Barbosa de Sousa
Membro Titular (UnB/PPGCINF)

Profª Drª Kelley Cristine Gonçalves Dias Gasque
Membro Suplente (UnB/PPGCINF)

"*O mundo amazônico deve ser economicamente viável,*
ecologicamente adequado,
politicamente equilibrado,
e socialmente justo".

Samuel Isaac Benchimol
(1924-2002)

Dedico

este trabalho à memória do meu filho Lucas.

Agradeço

a Deus, pela força para perseverar e vencer os momentos de desânimo e cansaço;
a Cláudio e Francisca, meus pais, pelo exemplo de trabalho e caráter;
a Marília, minha base sólida, amiga, amante e cúmplice, que me completa em todos os aspectos;
aos meus filhos Lucas – in memorian –, Gabriel, Mariana e Carlos Eduardo, por suportarem com paciência os momentos da minha ausência.
À Doutora Líllian Maria Araújo de Rezende Alvares, minha professora e orientadora, grande mestra e mentora, pela cobrança, incentivo e seriedade na orientação desta pesquisa.

Obrigado.

RESUMO

A Região Amazônica tem sofrido com a desatenção da sociedade. Há décadas os problemas sociais, econômicos e ambientais são resolvidos de maneira assistemática, reativa e sem soluções definitivas. Na tentativa de reverter a situação, o Ministério do Desenvolvimento, Indústria e Comércio Exterior instituiu em 2003 o Prêmio Professor Samuel Benchimol, um certame que chama a sociedade a pensar os problemas da Amazônia sob suas perspectivas ambiental, econômica-tecnológica e social. Somaram-se, nos últimos doze anos, através deste prêmio, centenas de projetos, ideias e propostas que refletem as necessidades de desenvolvimento da Região. Este trabalho teve por objetivo identificar, classificar e organizar as propostas apresentadas, por meio da análise do acervo documental do Prêmio Professor Samuel Benchimol. Como metodologia deste estudo, foram aplicados processos de recuperação da informação em bases de dados não estruturadas, com a utilização de técnicas da Bibliometria e da Análise de Conteúdo, automatizadas em software de mineração de texto. Na perspectiva ambiental, as principais questões encontradas foram correlacionadas à implantação de alternativas para a educação ambiental nas comunidades locais e na escola, a valorização do meio ambiente, a preservação e conservação do ecossistema e da biodiversidade e a recuperação das áreas degradadas, em especial das matas ciliares. Na perspectiva econômico-tecnológica, os pontos principais estão atrelados ao desenvolvimento da Amazônia pelo empreendedorismo consciente, as propostas para uso de tecnologias sustentáveis para aumento da qualidade e da quantidade da produção, além da necessidade de promoção e ampliação da indústria sustentável da região. Já para a natureza “Social”, as questões-chave consistem nas propostas de iniciativas para inclusão social, nos problemas de vulnerabilidade social das famílias, em especial as crianças e os idosos, nos projetos e nos programas para geração de empregos e renda para os jovens, nos projetos de prevenção e controle de doenças e morbidades frequentes na Amazônia; e nos problemas de saneamento básico e baixa qualidade da água.

Palavras-chave: Amazônia. Análise de Conteúdo. Bibliometria. Mineração de Textos. Prêmio Professor Samuel Benchimol. Recuperação da Informação.

ABSTRACT

The Amazon region has suffered from the lack of society attention. For decades, social, economic and environmental problems are solved so unsystematic, reactive and no definitive solutions. In an attempt to reverse the situation, the Brazilian government established in 2003 the Professor Samuel Benchimol Award, an exhibition that calls society to think the Amazon problems in their environmental, economic, technological and social perspectives. The award received in the last 12 years hundreds of projects, ideas and proposals that reflect the region's development needs. This study aimed to identify, classify and organize the proposals submitted by the collection of Professor Samuel Benchimol Award. As methodology of this study, were used information retrieval processes in databases unstructured, using techniques of bibliometrics and content analysis automated in text mining software. On environmental perspective, the main issues found were related to the implementation of alternatives for environmental education in local communities and school, appreciation of the environment, preservation and conservation of ecosystems and biodiversity, and the recovery of degraded areas, especially of riparian forests. The economic and technological perspective, the main points are correlated to the development of the Amazon by conscious entrepreneurship, proposals for the use of sustainable technologies to increase the quality and quantity of production and the need for promotion and expansion of sustainable industry in the region. In social perspective, the key issues are the proposed initiatives for social inclusion. The social problems of vulnerability of families, especially children and the elderly, in projects and programs to generate jobs and income for young people in the projects prevention and control of common diseases and morbidities in the Amazon and sanitation problems and poor water quality.

Key-words: Amazon. Bibliometrics. Content Analysis. Information Retrieval. Teacher Samuel Benchimol Award. Text Mining.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AB – Bibliometria
AC – Análise de Conteúdo
AD – Análise de Discurso
ARPA – Advanced Research Projects Agency Network
ASK – Anomalous State of Knowledge
BD – Banco de Dados
BCE – Biblioteca Central da Universidade de Brasília
CDD – Classificação Decimal de Dewey
CDU – Classificação Decimal Universal
CI – Ciência da Informação
DOD – Department of Defense
GPS – Global Positioning System
HTML – Hyper Text Markup Language
IP – Protocolo de Internet
IR – Information Retrieval
ISACA – Information Systems Audit and Control Association
KDD – Knowledge Discovery in Database
KDT – Knowledge Discovery from Text
MDIC – Ministério do Desenvolvimento, Indústria e Comércio Exterior
NASA – National Aeronautics and Space Administration
NLM – National Library of Medicine
OC – Organização do Conhecimento
OCR – Optical Character Recognition
OI – Organização da Informação
OIC – Organização da Informação e do Conhecimento
PPGCINF – Programa de Pós-graduação em Ciência da Informação da Universidade de Brasília
RDF – Resource Description Framework
RI – Recuperação da Informação
ROC – Representação e Organização do Conhecimento
ROI – Representação e Organização da Informação
ROIC – Representação de Organização da Informação e do Conhecimento
SGBD – Sistema Gerenciador de Banco de Dados
SRI – Sistema de Recuperação da Informação
TCP/IP – Transfer Control Protocol/Internet Protocol
UnB – Universidade de Brasília
XML – eXtensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

O Prêmio Professor Samuel Benchimol é um certame que instiga empresários, acadêmicos e pesquisadores a pensarem acerca dos problemas da Amazônia sob as perspectivas ambiental, social, econômica e tecnológica. O acervo documental produzido gerou centenas de projetos, ideias e propostas que refletem as necessidades de desenvolvimento desta região.

Este trabalho tem por objetivo identificar, classificar e organizar as propostas apresentadas, por meio de análise do acervo documental do Prêmio Professor Samuel Benchimol. Para tanto, foram aplicadas técnicas de Análise de Conteúdo (AC) e Bibliometria automatizadas em software de mineração de texto.

A Ciência da Informação (CI) dispõe de ferramentas que possibilitam a análise de grandes volumes de dados, de fontes estruturadas ou não, para extração de conhecimento, o que permitirá o atingimento dos objetivos deste trabalho. Entre as ferramentas que foram aplicadas destacam-se a AC, sob a ótica de Laurence Bardin e a mineração de textos.

A Análise de Conteúdo é o nome genérico para técnicas de descrição do conteúdo das mensagens, que permitem traduzir a informação e o conhecimento a elas associados. Aplica-se à linguagem verbal e também a imagens, desenhos, pinturas, cartazes, vídeo e a toda comunicação não verbal: gestos, posturas, comportamentos e outras expressões culturais.

A Bibliometria, por sua vez, propõe o uso de métodos matemáticos e estatísticos (leis, fórmulas e teoremas) de análise e construção de indicadores para a mecânica da evolução da informação científica e tecnológica em campos multidisciplinares. Os estudos de Bibliometria ocupam-se da tentativa de quantificar os processos de comunicação escrita.

A mineração de textos utiliza mecanismos da Bibliometria para extração de informações, tendências, padrões e descoberta de conhecimento em grandes bases de documentos textuais, apoiada por software.

Em relação à estrutura textual, além desta introdução, a pesquisa apresenta o contexto do projeto, justifica a sua relevância para a Ciência da Informação e para as linhas de pesquisas de Organização da Informação e do Conhecimento e contextualiza o problema e os objetivos desta, encontrando-se ainda os seguintes capítulos:

a) Na segunda seção encontra-se a revisão de literatura, que aborda os temas relevantes à compreensão da natureza do projeto e a contextualização da pesquisa. No campo teórico, os principais temas abordados na revisão de literatura tratam da Interdisciplinaridade da Ciência da Informação com a Ciência da Computação; Organização da Informação e do Conhecimento; Recuperação da Informação; Bibliometria; Análise de Conteúdo; Bigdata e mineração de textos.
b) Na terceira seção, detalha-se a metodologia da pesquisa. Nesta parte do documento é demonstrada a sua caracterização, os processos e procedimentos, etapas e condições para a sua execução.
c) Já a quarta seção, traz os resultados e análises da pesquisa, destacando: os resultados quantitativos da mineração de dados; as estatísticas do prêmio; os elementos qualitativos extraídos pela Análise de Conteúdo; o perfil bibliométrico dos projetos e propostas do Prêmio Professor Samuel Benchimol.
d) Conclusões.
e) Por último, como apêndices I e II a IX, respectivamente, a Lista de *Stopwords* utilizada no processo de mineração de textos e as evidências dos recortes temáticos específicos (extraídos diretamente dos software) da mineração de textos, no acervo documental do Prêmio Professor Samuel Benchimol, analisados nesta pesquisa.

O problema de pesquisa encontra motivação nas inúmeras questões sociais, econômicas e ambientais da Amazônia que são tratadas assistematicamente, e de forma reativa, não resultando em soluções definitivas. A situação persiste há décadas, desde o último período de desenvolvimento da região, no Segundo Ciclo da Borracha. Ou seja, na década de 1910, empresários holandeses e ingleses entraram no lucrativo mercado mundial de borracha. Passaram a produzir, em larga escala e custos baixos, o produto na Ásia (Ceilão, Indonésia e Malásia). A concorrência fez com que, no

começo da década de 1920, a exportação da borracha brasileira caísse significativamente. Era o fim do ciclo da borracha no Brasil. Muitas cidades esvaziaram-se, entrando em plena decadência.

Para tentar reverter a situação o Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC) instituiu em 2003 um certame para chamar a população brasileira a “pensar” a Amazônia e, assim, foi criado o Prêmio Professor Samuel Benchimol, cujo objetivo perpassa a promoção, a reflexão e a proposição de ações no contexto econômico, científico-tecnológico, ambiental, social e de empreendedorismo para o desenvolvimento sustentável da Amazônia.

Assim, foi somada mais de uma década de contribuições, em pesquisas científicas, projetos, ideias e propostas que refletem as necessidades de desenvolvimento da Região Amazônica. Esse acervo materializou-se como uma grande oportunidade de estudo, ainda não realizado, para análise das pesquisas e trabalhos apresentados.

Espera-se identificar as questões prioritárias, por meio do levantamento dos temas mais frequentes nesses anos; a segmentação dos temas mais explorados; as instituições que emergiram como proponentes e ganhadores; e quais as principais linhas das propostas. Esses trabalhos são o que há de mais representativo em pesquisas, projetos e propostas para o desenvolvimento sustentável da Amazônia, uma vez que são originários das instituições que se dedicam direta ou indiretamente ao estudo da Amazônia.

Complementarmente a essas informações, pretende-se identificar qual é o perfil dos pesquisadores que os submetem; quais os aspectos mais relevantes, recorrentes e comuns dos trabalhos; quais são seus estados de origem, das propostas e dos temas mais submetidos por eles. Atualmente, não existem essas estatísticas.

Pode-se sintetizar o contexto apresentado na seguinte questão: Como, utilizando os instrumentais da bibliometria, da mineração de textos e da análise de conteúdo, é possível contribuir para a discussão e soluções referentes às questões-chave do desenvolvimento sustentável da Região Amazônica?

Modernamente, é fácil verificar na internet a existência de diversos aplicativos (alguns gratuitos) para a realização de Análise de Conteúdo qualitativa e quantitativa,

os quais poderão ser pesquisados e testados pelo leitor de acordo com suas necessidades. Nesta pesquisa, especificamente, foram utilizadas as funcionalidades de Bibliometria, Análise de Conteúdo Automatizada, mineração de texto e Análise Semântica de Conteúdo, da plataforma Tropes Zoom software.

Os software Tropes são desenvolvidos, evoluídos e mantidos pela Semantic Knowledge, empresa multinacional, criada conjuntamente pela francesa ACETIC (especializada em Análise Semântica de Textos e Processamento de Linguagem Natural), pela portuguesa CYBERLEX (especialista em Recuperação da Informação e Análise de Linguagem) e por investidores e pesquisadores autônomos. Desde o ano de 2002, o Tropes Zoom software já foi distribuído e licenciado para mais de cem mil usuários em todo o mundo.

Para responder ao problema apresentado, os objetivos podem ser assim definidos:

OBJETIVO GERAL

- Identificar, classificar e analisar as propostas submetidas ao Prêmio Professor Benchimol, durante os anos de 2004 a 2015, por meio da mineração de textos, para definição das questões-chave de desenvolvimento da Região Amazônica sob a ótica desse prêmio.

OBJETIVOS ESPECÍFICOS

Complementarmente às informações, pretende-se identificar qual é o perfil dos pesquisadores que os submetem; quais os aspectos mais relevantes, recorrentes e comuns dos trabalhos; quais os seus estados de origem, das propostas e dos temas mais submetidos. Atualmente, não existem essas estatísticas. Nesse sentido, os Objetivos Específicos são:

a. Identificar, quantificar, e qualificar os temas, propostas e projetos apresentados, quanto à recorrência das questões;

b. Segmentar as propostas de trabalhos apresentados nas categorias ambiental, econômico-tecnológica e social, bem como analisar as ligações entre eles;

c. Analisar a base de conhecimento revelada pelas perspectivas quantitativa e qualitativa.

Considerando o grande acervo científico e documental do Prêmio Professor Samuel Benchimol, em mais de uma década de existência, vislumbrou-se a oportunidade de realizar pesquisa científica para apoiar a análise das questões relacionadas ao desenvolvimento sustentável da Amazônia, tendo em vista que a pesquisa científica lidou diretamente com um grande volume de informações, assumiu-se que deveria ser realizada no âmbito da Ciência da Informação, que é a disciplina ocupada com a investigação das propriedades, do comportamento e do fluxo informacional, bem como dos meios para processar a informação, com o objetivo de atingir acessibilidade e utilidade ótimos (BORKO, 1968).

O estudo foi realizado na linha de pesquisa “Organização da Informação” do Programa de Pós-Graduação em Ciência da Informação, da Universidade de Brasília. Entende-se, assim, que essa é a melhor linha para este trabalho, uma vez que se trata de pesquisa aplicada para Recuperação da Informação (RI) em um grande volume documental heterogêneo. Segundo Lima e Alvares (2012, p. 35), o objetivo central da Organização da Informação (OI) “é permitir a recuperação e o acesso à informação por meio da estruturação dos elementos de organização do conhecimento”. Já a Organização do Conhecimento (OC) tem na representação do conhecimento “uma tentativa de se apropriar dos elementos informacionais existentes nas estruturas e processos mentais que compõem o conhecimento individual, para que o saber possa ser socializado” (LIMA; ALVARES, 2012, p. 33).

Organizar a informação e o conhecimento do grande acervo científico e documental do Prêmio Professor Samuel Benchimol contribuiu para a identificação dos temas mais relevantes para a Região Amazônica, seus pesquisadores, os estudos mais proativos, as instituições que se destacam.

A pesquisa também justificou-se pelo ponto de vista teórico, com a aplicação de conceitos bibliométricos para a imersão e aprofundamento nos aspectos quantitativos e qualitativos da Região Amazônica.

Do ponto de vista prático, tornou-se viável pela realização de Análise Automatizada de Conteúdo e mineração de textos na aplicação de leis clássicas da Bibliometria (LOTKA, BRADFORD e ZIPF) ao acervo do prêmio, para investigação sistemática de autores, trabalhos e palavras, a fim de estabelecer fatos e chegar a novas conclusões, descobrir novos fatos ou agrupar antigos, por meio de estudo científico do tema Amazônia, permitindo uma abordagem quantitativa. Especificamente, a Bibliometria contou com um conjunto de abordagens e técnicas baseado em software de mineração de texto.

Esperava-se que os resultados da pesquisa possam apoiar iniciativas e caminhos em busca do desenvolvimento da Região Amazônica. Tratava-se de uma proposta oportuna, pois pretendia contribuir com o mapeamento, a identificação, a análise e a priorização de propostas de desenvolvimento da Região Amazônica. Fez-se também urgente pela relevância que o tema Amazônia possui nos debates atuais.

2 REVISÃO DE LITERATURA

2.1 ORGANIZAÇÃO DA INFORMAÇÃO E DO CONHECIMENTO (OIC)

Com a evolução do conceito de informação, surgem inúmeros desafios contemporâneos relativos às diversas etapas do seu ciclo de vida: geração, documentação, armazenamento, recuperação, acesso, representação e organização da informação e do conhecimento. A problemática contemporânea da OIC origina-se na inquietação humana pelo saber. Desde o advento da produção científica, o homem cria e desenvolve mecanismos para classificação dos seres, objetos, informação, saberes, modos de fazer e do conhecimento, em uma busca pelo entendimento do mundo e do próprio homem (POMBO, 1998).

As fronteiras deste campo de pesquisa ainda não são claras ou bem definidas. Alguns autores tratam da OIC de maneira unificada, outros separam Representação e Organização da Informação (ROI ou apenas OI) de Representação e Organização do Conhecimento (ROC ou apenas OC). Segundo Lima e Alvares (2012, p.35), o objetivo central da OI "é permitir a recuperação e o acesso à informação por meio da estruturação dos elementos de organização do conhecimento", já a OC, aduz que "Representar o conhecimento é uma tentativa de se apropriar dos elementos informacionais existentes nas estruturas e nos processos mentais que compõem o conhecimento individual, para que o saber possa ser socializado" (LIMA e ALVARES, 2012, p. 33).

A ORC trouxe subsídios como as Teorias do Conceito e da Classificação, além da Análise Documentária, inspirada na Lógica, na Filosofia, na Linguística e na Teoria Geral das Terminologias. Concorda Navarro, ao afirmar que "Organização do Conhecimento apresenta-se como uma plataforma de integração das ciências documentais" (NAVARRO, 1995). O desafio é criar mecanismos para organizar a representação do conhecimento, concernente ao pensamento de Lima e Álvares: "Dentre seus limites de atuação, tenta responder a como se representa o conhecimento; se as áreas do conhecimento são representadas da mesma maneira o

que pode ser representado; e se tudo pode ser representado" (LIMA; ALVARES, 2012, p. 27).

Nesse sentido, os estudos de ROI e ROC trouxeram à pauta a necessidade de um aprofundamento interdisciplinar, principalmente em relação às Teorias da Classificação, que sustentam o arcabouço metodológico-teórico necessário à divisão e à organização do conhecimento, conforme sustenta Burke, citado por Araújo:

> o autor destaca esse momento histórico como especificamente relevante, quando se verificou um esforço sistematizado de divisão e organização do conhecimento, desde as "árvores do conhecimento", no século XVI, até os três subsistemas que serviram para a classificação do conhecimento no âmbito das universidades europeias: a organização dos currículos, a ordem das bibliotecas e a estrutura das enciclopédias. (BURKE, 2003, p. 79 apud ARAÚJO, 2006).

Segundo Lima e Alvares, "Representar o conhecimento é uma tentativa de se apropriar dos elementos informacionais existentes nas estruturas e processos mentais que compõem o conhecimento individual, para que o saber possa ser socializado" (LIMA e ALVARES, 2012, p. 33). A força motriz dos estudos de ROIC está na necessidade de Recuperação da Informação e do conhecimento, para que possa haver o intercâmbio informacional mesmo para pessoas separadas geograficamente ou temporalmente. Os autores defendem, ainda, que a qualidade da Recuperação da Informação depende do processo de organização e representação da informação.

No que diz respeito à Classificação, é um processo definido como "dividir em grupos ou classes, segundo as diferenças e semelhanças. É dispor os conceitos, segundo suas semelhanças e diferenças, em certo número de grupos metodicamente distribuídos" (PIEDADE, 1977). Araújo (2006) concorda com esta linha, afirmando que o principal elemento para a caracterização do processo de classificação é a formação metódica e sistemática de grupos. Trata-se do ordenamento sistemático de um conjunto de registros informacionais em partes menores, por meio das semelhanças e características comuns que os incluem dentro de determinado grupo e, ao mesmo tempo, não compartilhadas pelos demais registros. Esse processo define critérios de divisão, classificações, distinções e aproximações para agrupamentos dos registros (ARAÚJO, 2006, p. 2).

Classificação é um processo hierárquico e finito de subdivisão sucessiva classificatória de domínios até o nível ideal (APOSTEL, 1963, *apud* POMBO, 1998).

Assim, a Classificação deve comportar um número finito de divisões e um número finito de classes internas a cada divisão (o que se chama de finitude da classificação) e em cada nível subsequente (POMBO, 1998). Não deve admitir conjuntos ou subconjuntos idênticos a outros níveis anteriores ou posteriores (o que se chama de progressividade da classificação).

Essas divisões não devem estar vazias e nem sobrepostas, mesmo que parcialmente, devendo ser exaustivas, isto é, cobrir toda a extensão do domínio classificado. Segundo Apostel (*apud* POMBO, 1998), existem cinco características gerais de toda classificação: i) Cada classificação usa uma determinada estrutura classificadora que executa, com melhor ou pior eficácia, as operações necessárias à classificação; ii) Cada classificação visa uma multiplicidade sistemática de fins que são determinantes à sua estrutura; iii) Cada classificação participa de um domínio da realidade em que as estruturas internas influenciam o nível de dificuldade das operações inerentes ao processo de classificação; iv) Cada classificação pertence a um contexto das classificações precedentes do mesmo domínio, cuja historicidade provoca subdivisões onde novos critérios de classificação são gerados; v) Para cada classificação tem-se uma interface externa com a atividade classificadora à qual está vinculada, representando uma árvore genealógica da classificação. É o processo de estabelecimento de hierarquias entre subclasses no interior das classes previamente estabelecidas (APOSTEL 1963, p. 195, apud POMBO, 1998).

Para Araújo (2006), existem inúmeras manifestações da Classificação, desde as classificações tidas como sociais, integrantes da vida humana e cotidiana (por exemplo, “classe média”, “classe média alta”, “classe média baixa”; “música erudita”, “música popular”, “música da cultura de massa”; “políticos de centro”, de “centro-esquerda”, “de centro-direita”), até aquelas especializadas e, entre essas, destacam-se as classificações bibliográficas. Para as classificações bibliográficas existem várias facetas possíveis, conforme sustenta Araújo, citando Burke que reconhece a existência de várias formas de classificação do conhecimento ao longo da história humana, em que as distinções mais comuns consistem em “conhecimento teórico x prático; público x privado; legítimo x proibido; alto x baixo; liberal x útil; especializado x geral; dos livros x das coisas; e conhecimento quantitativo x qualitativo [...]” (BURKE, 2003, p. 79, apud ARAÚJO, 2006).

A teoria da classificação de Aristóteles apoia-se em cinco predicados dos arranjos lógicos: i) Gênero: classe ou grupo de indivíduos ou objetos que comungam de certo número de características; ii) Espécie: indivíduo que possui uma diferença específica que o diferencia de seu gênero mais próximo (gênero + diferença); iii) Diferença: é a característica distingue uma nova espécie; cada nova diferença gera uma nova espécie; iv) Propriedade: algo exclusivo a cada elemento de uma classe, todavia, não é imprescindível à definição da classe; v) Acidente: ocorrência esporádica em elementos de uma classe (ARAÚJO, 2006).

As noções de classificação hierárquica (SHERA; EGAN, 1969, p. 55, apud ARAÚJO, 2006) foram essenciais no desenvolvimento e na formulação dos primeiros sistemas de classificação bibliográfica, normalmente conhecidos como sistemas de classificação hierárquicos devido à forma de organização dos conceitos em estruturas de gênero e espécie, identificando atributos essenciais e acidentais. Assim, obteve-se uma estrutura conceitual pela aplicação sucessiva de características de divisão (ARAÚJO, 2006). Dentre os primeiros sistemas de classificação bibliográfica, os mais representativos são os de Cutter, a Classificação Decimal de Dewey (CDD), a Classificação Decimal Universal (CDU) e a classificação da *Library of Congress* (MENDES, 1995, p. 41, apud ARAÚJO, 2006).

Em meados do século XX, Ranganathan (1967) propôs uma nova forma de classificação bibliográfica (facetada). O diferencial do sistema classificatório proposto por ele é a utilização de uma estrutura dinâmica, com o ingresso do termo faceta, “que ficou sendo, nos modernos estudos sobre teoria da classificação, o substituto de característica” (BARBOSA, 1969, p. 16, apud ARAÚJO, 2006). O ponto de partida foi outra ideia de Aristóteles, a demarcação das dez categorias do ser, ou seja, as formas sob as quais os seres e objetos apresentam-se:

> Substância ou matéria (homem, cachorro, pedra, casa, etc.); qualidade (azul, virtuoso, etc.); quantidade ou extensão (grande, comprido, dois quilos, etc.); relação (mais pesado, escravo, duplo, mais barulhento, etc.); tempo ou duração (ontem, 1970, de manhã, etc.); lugar ou localização (aqui, Brasil, no pátio, etc.); ação ou atividade (correndo, cortando, falando, etc.); paixão ou sofrimento da ação (derrotado, cortado, etc.); maneira de ser (saudável, febril, etc.); posição (horizontal, sentado, etc.) (DODEBEI, 2002, p. 96-97 apud ARAÚJO, 2006).

Essas categorias podem ser utilizadas como um grande conjunto de características classificatórias, ou seja, para a separação entre os seres e a

delimitação de grupos. A evolução desses estudos e conceitos possibilitou o desenvolvimento dos sistemas facetados, que foram construídos para atender a diferentes objetivos. O mais comum deles é a organização de documentos objetivando proporcionar formas ativas e distintas de acesso aos conteúdos. Por fim, tem-se o fato de que os processos relacionados à Representação e Organização da Informação são centrais no escopo do ciclo de vida da informação. O maior impacto é percebido quando da necessidade de Recuperação da Informação (RI), em que a efetividade dos métodos de representação e a organização estão diretamente relacionadas à qualidade dos conteúdos recuperados. As práticas e os métodos de Organização da Informação e do Conhecimento contribuirão com esta pesquisa no tocante à categorização e à classificação das propostas, projetos, pesquisas, temas e subtemas do acervo do Prêmio Samuel Benchimol.

2.2 RECUPERAÇÃO DA INFORMAÇÃO (RI)

A preocupação em registrar e recuperar a informação é uma inquietação latente do ser humano, inicialmente como forma de transmissão de experiências e registro de fatos históricos. A CI, em sua acepção, é ocupada da produção, seleção, organização, interpretação, armazenamento, recuperação, disseminação, transformação e uso da informação (GRIFFITH, 1980 apud CAPURRO, 2003, p. 4).

Em 1968, Harold Borko, em seu artigo *Information Science: What Is It?*, foi o primeiro a organizar os limites para esta nova disciplina e área de conhecimento. Borko criou uma definição para a Ciência da Informação que ainda é aceita até hoje:

> Ciência da Informação é a disciplina que investiga as propriedades e o comportamento da informação, as forças que regem o fluxo informacional e os meios de processamento da informação para a otimização do acesso e uso. Está relacionada com um corpo de conhecimento que abrange a origem, coleta, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação [...]. Tem tanto uma componente de ciência pura, que indaga o assunto sem ter em conta a sua aplicação, como uma componente de ciência aplicada, que desenvolve serviços e produtos. [...] A biblioteconomia e a documentação são aspectos aplicados da ciência da informação (BORKO, 1968).

Dentre os vários conceitos presentes na definição de Borko para a CI destaca-se a Recuperação da Informação. Sem os processos de RI o conceito de CI não

estaria completo, pois as informações registradas que não podem ser recuperadas e utilizadas pouco ou nada têm a contribuir com a ciência da informação.

Segundo Choo (2003) “A informação é um componente intrínseco de quase tudo que uma organização faz. Sem uma clara compreensão dos processos organizacionais e humanos pelos quais a informação transforma-se em percepção, conhecimento e ação, as empresas não são capazes de perceber a importância de suas fontes e tecnologias de informação”. O autor ainda define três arenas de uso da informação: i) Criar significado; ii) Construir conhecimento; e iii) Tomar decisões (CHOO, 2003, p. 27-28).

Nesse contexto, Choo (2003) deixa claro que as organizações que forem capazes de integrar efetivamente os processos de criação de significado, construção do conhecimento e tomada de decisões poderão ser consideradas organizações do conhecimento. Em plena era do Bigdata, as dificuldades para acesso e Recuperação da Informação apontam para que um vasto caminho ainda seja percorrido, tanto no campo acadêmico, quanto no âmbito prático do desenvolvimento de software e ferramentas computadorizadas para acesso e Recuperação da Informação.

A expansão maciça da Internet, da Web Semântica e, principalmente, da produção de informação em múltiplas mídias (fotos, vídeos, textos, sons, ondas, fractais, etc.), culminou na produção de um “caos” informacional que já não mais pode ser administrado pelas ferramentas tradicionais de RI, isso inaugurou uma nova época, baseada em software de Bigdata. Junto com a nova plataforma tecnológica também estão os desafios para manipular, processar e gerar informações úteis a partir de múltiplas bases de dados: estruturadas, semiestruturadas e não estruturadas.

Em 1945, Vannevar Bush, avançando na temática, introduz o conceito do interesse específico da CI nos processos de Recuperação da Informação (RI), destacando que esta deveria ser operacionalizada por associação de elementos conceituais. Nos anos de 1950, Calvin Mooers cunhou o termo *Information Retrieval* (Recuperação da Informação) e definiu a sua ocupação nos aspectos intelectuais de descrição da informação e sua especificação para busca e também quaisquer sistemas, técnicas ou máquinas que são utilizadas para executar a operação. Nos dias de hoje, em plena explosão informacional, percebemos uma clara evolução dos métodos, processos e ferramentas, todavia, também com novos desafios. Entre os

novos paradigmas discutidos no campo da Recuperação da Informação está o conceito de Bigdata. Este trabalho discutirá aspectos históricos, acontecimentos, fatos, autores e marcos da evolução dos sistemas e ferramentas de RI até a nova era do Bigdata.

2.2.1 Mecanismos de recuperação da informação

Bons mecanismos para RI são tão importantes para os usuários quanto o próprio conteúdo, pois sem esses meios não se acessa os registros informacionais. Os usuários de informação científica, incluindo estudantes de graduação e pós-graduação, deveriam ser ensinados também sobre como recuperar a informação, facilitando e tornando mais eficientes os processos de RI (GARFIELD, 1967).

Em 1979, Rijsbergen descreve as limitações e a problemática dos processos de armazenamento e Recuperação da Informação, principalmente no tocante ao dilema da grande quantidade de informação disponível e da dificuldade no acesso correto e rápido a ela. O fenômeno, conhecido como explosão da informação, caracterizado pelo “irreprimível crescimento exponencial da informação e de seus registros, particularmente em ciência e tecnologia” (SARACEVIC, 1996, p. 42), corroborou com o desafio de recuperar informação. Ingwersen (1992) também relaciona a Recuperação da Informação aos processos de armazenamento da informação, assim como a processos de representação e busca. O autor enfatiza que a informação presente nesses processos deve ser relevante ao usuário, segundo os seus desejos de informação.

Desde então, com o advento de novas tecnologias, de meios de comunicação inovadores, da inflexível dependência da Ciência da Informação e da Tecnologia da Informação (SARACEVIC, 1995) e do importante papel da informação na sociedade contemporânea (CAPURRO, 2007), a Recuperação da Informação torna-se umas das principais ocupações da Ciência da Informação (SARACEVIC, 1995).

Belkin e Croft (1987) definem a busca e a recuperação de informação como um processo de localização de documentos e de itens de informação previamente

armazenados, visando permitir ao usuário o acesso a eles. Portanto, a Recuperação da Informação ocorre pela comparação do que foi solicitado com o que está armazenado, bem como com o conjunto de procedimentos que tal processo envolve. Belkin (1982) havia afirmado que um elemento fundamental nos processos de RI é a necessidade de informação, gerada a partir de um Estado Anômalo de Conhecimento (*Anomalous State of Knowledge – ASK*). Os estudos sobre ASK indicam que a necessidade de informação surge com o reconhecimento de um estado anômalo de conhecimento do usuário a respeito de um assunto e, após contato com o conteúdo recuperado, o usuário passa a um novo estado de conhecimento. Portanto, para que a Recuperação da Informação seja bem sucedida, é importante que as informações presentes em um sistema de RI sejam representadas em termos apropriados aos seus usuários.

Respaldando o pensamento de Belkin e Croft (1987), Pignatari (1993) conceitua a Recuperação da Informação com um conjunto de instruções seletivas que possibilitam ao usuário satisfazer suas necessidades, dúvidas e anseios. Rodrigues e Crippa (2011) também consideram as necessidades do usuário ao afirmar que as questões da RI estão relacionadas àquilo que é importante ao usuário. Por isso, um Sistema de Recuperação da Informação (SRI) deve levar em conta a relevância informacional. Lancaster (2004), ao tratar de questões relacionadas à indexação de documentos, afirma que o problema de Sistemas de Recuperação da Informação está na dificuldade em recuperar todos os documentos úteis a um usuário, sem recuperar documentos não pertinentes.

Bastos (1994, apud FIGUEREIDO, 2006) ampliando o conceito, define RI como um subprocesso de comunicação no qual emissor e receptor interagem para atender a uma necessidade de informação, inserido em um processo multidisciplinar que envolve conhecimentos lógicos, tecnológicos e linguísticos.

A Recuperação de Informação é uma vertente tecnológica da Ciência da Informação, consequência da interdisciplinaridade com a Ciência da Computação. Um dos intuitos da Ciência da Informação é prover acesso efetivo aos usuários de Sistemas de Recuperação da Informação (SRI) às informações que lhes sejam relevantes. Porém, existe grande subjetividade na definição de informação e no

conceito de relevância, o que dificulta a realização deste objetivo (SARACEVIC, 1995).

Nesse sentido, Rodrigues e Crippa (2011) destacam que as discussões referentes à RI estão intrinsecamente ligadas à noção de relevância informacional, em que um SRI não se dedica a indexar qualquer coisa, mas, sim, aquilo que é importante e relevante para o usuário e, em consequência disso, aquilo que será recuperado posteriormente (RODRIGUES; CRIPPA, 2011, p. 4). Observar este preceito básico dos processos de RI obriga a aplicação de abordagens para aumentar a relevância informacional dos registros recuperados.

Outra questão que maximiza o problema observado por Mooers (1950) no que tange a Recuperação da Informação (RI), é a indefinição do usuário dos sistemas de Recuperação da Informação. Estes sistemas tentam atender às necessidades informacionais dos usuários. Todavia, existe a percepção de problemas nesses sistemas. Um deles é o "caos" informativo (CATARINO; BAPTISTA, 2007). Os usuários não utilizam expressões e termos de busca em linguagem documentária e não há uma normalização do uso das palavras para as pesquisas. Cada usuário pesquisa seguindo seu próprio modelo mental e cognitivo, ou seja, a expressão que faz sentido para ele. As palavras e expressões utilizadas deixam lacunas à polissemia ou mesmo à polifonia, além da sinonímia.

Para cada registro informacional existe uma "oferta de sentidos" (CAPURRO, 2003; LARA, 2008) e os usuários, leitores e consumidores da informação, selecionam com base em seu modelo mental, formação cultural, influência social, vivências históricas e ideologias. Tudo dependente da capacidade e habilidade de interpretação de cada indivíduo (CAPURRO; HJORLAND, 2007), dificultando o trabalho dos profissionais da informação, principalmente daqueles que se ocupam dos processos e ferramentas para RI.

Apoiando esse pensamento, Bates (1999) distinguiu o processo de busca em duas partes distintas: uma realizada por indivíduos e a outra pelo sistema de recuperação. Observou-se que o processo de busca efetuado por pessoas é constituído por diferentes movimentos, táticas, estratagemas e estratégias. Por outro lado, o processo de busca realizado pelo sistema de recuperação seria composto,

entre outros, pela linguagem de busca do sistema e pela estrutura de informação da base de dados.

Face ao exposto, tem-se que aquilo que é considerado como informação por um indivíduo pode não o ser para o outro. Quando é iniciado um processo de busca por determinada informação em um sistema de recuperação, o que se recupera é o registro físico desta (RODRIGUES; CRIPPA, 2011). Segundo Ferneda (2003, p. 11), “os sistemas não recuperam “informação”, mas sim documentos ou referências cujo conteúdo poderá ser relevante para a necessidade de informação do usuário”. A informação “só vai se consubstanciar a partir do estímulo externo-documento, se também houver uma identificação (em vários níveis) da linguagem desse documento, e uma alteração, uma reordenação mental do receptor-usuário” (BRAGA, 1995, p. 86, apud FERNEDA, 2003).

É imprescindível destacar que Bates (1999, p. 35) conceituou a estratégia de busca como o “estudo da teoria, princípios e prática de planejar e executar táticas e estratégias de busca”. A autora destaca-se na literatura por ter sido a primeira a definir teoricamente o conceito de estratégia de busca e a tática para a sua execução. Adaptando as definições ao ambiente de Recuperação da Informação, deu ênfase ao conceito de comportamento de busca, indagando: “o que as pessoas fazem, e como pode ser determinado o que elas pensam quando estão executando uma busca de informação?”.

2.2.2 Advento da internet e novas tecnologias de RI

Com a chegada da década de 80, após a *Advanced Research Projects Agency Network* (ARPA) adotar protocolo *Transfer Control Protocol* / *Internet Protocol* (TCP/IP), que possibilitou a comunicação entre redes de computadores, os problemas relacionados à Recuperação da Informação tomaram novas proporções. O TCP/IP tornou possível a integração de diversas instituições de pesquisa em uma grande rede, permitindo que milhares de usuários compartilhassem suas informações.

No início da década de 90, a Internet já conectava mais de um milhão de computadores, expandindo suas fronteiras para outras áreas, como o comércio eletrônico, jogos, cultura e entretenimento, extrapolando o universo acadêmico. Os registros informacionais passaram a ser integrados possibilitando o avanço nos mecanismos e processos de Recuperação da Informação. Nessa linha, surgiram então o sistema Archie, que possibilitava a busca de arquivos, e o sistema Gopher, que buscava informações por meio de menus e diretórios criados pelo usuário.

Tim Berners-Lee, um dos pais da Internet, no final da década de 1980 trabalhou de maneira árdua no desenvolvimento de tecnologias que possibilitassem o compartilhamento de informação usando arquivos de texto. Estes arquivos eram documentos textuais que se referenciavam por meio de ligações em nível de metadados.

A ideia de Berners-Lee era desenvolver uma ferramenta de comunicação baseada na recém-criada Internet, que possibilitasse o compartilhamento de informações, documentos e registros entre as universidades em todo o mundo. Assim, surgiu uma linguagem de marcação hipertextual denominada por ele de HTML, bem como os protocolos de comunicação da Web.

Com o advento da Internet e as grandes possibilidades de integração de bases de dados antes isoladas, houve o aparecimento de inúmeros sistemas e mecanismos de busca sem finalidade específica, como por exemplo o Yahoo, líder no segmento de *search engines* (motores de busca) até a entrada da Google neste mercado no início dos anos 2000. A lógica geral desses mecanismos era o ordenamento dos links para as páginas que, em teoria, guardavam os registros passíveis de serem recuperados. A lista apresentada era ranqueada em função da expressão fornecida pelo usuário ao SRI, desprezando os registros considerados de baixa relevância.

A linguagem HTML (*HyperText Markup Language*) foi muito bem aceita pelos desenvolvedores de páginas Web, principalmente nas versões 1.0, 4.0 e, mais recentemente, na versão 5.0, tornando-se um padrão. Todavia, mesmo com as atualizações de versão, surgiram novas exigências e demandas tecnológicas para desenvolvimento de sites e Recuperação da Informação que não puderam ser atendidas pelas limitações da HTML.

Este cenário foi propício para o surgimento de uma nova linguagem que resolvesse as limitações da HTML, a linguagem XML (*eXtensible Markup Language*). XML tem como principal característica a flexibilidade, pois um desenvolvedor de páginas Web pode definir suas próprias tags, libertando-se da marcação da HTML padrão. Em XML a estrutura e a semântica da linguagem integram o interior de um documento. Segundo Ferneda (2003), a linguagem *Resource Description Framework* (RDF) fornece um meio de agregar semântica a um documento sem se referir à sua estrutura, eis que “A RDF visa oferecer uma forma eficiente de descrever metadados na Web, possibilitando a interoperabilidade entre aplicações que compartilham metadados” (FERNEDA, 2003, p. 111).

Este avanço tecnológico possibilitou uma melhora considerável nos processos de Recuperação da Informação, pois bases de dados e sistemas implementados em linguagens de programação diferentes puderam ser integrados por meio da linguagem XML. A ideia era aumentar a eficiência dos mecanismos de busca e de outros tipos de ferramentas de processamento automático de documentos por meio de linguagens de definição de dados e regras da Web Semântica (DACONTA, OBRST, SMITH, 2003).

Para viabilizar a Web Semântica é necessário um conjunto de linguagens que permitam tanto a definição de dados, através de marcações (HTML), quanto possibilitem também descrever formalmente estruturas (XML) conceituais que possam ser utilizadas pelos robôs de indexação dos motores de busca. Por outro lado, somente as linguagens não são suficientes para viabilizar a interoperabilidade de conteúdo. São necessários protocolos e infraestrutura de comunicação que possa prover a integração entre as fontes de informação.

A interoperabilidade de conteúdos e metadados entre diferentes bases de dados e Sistemas de Recuperação da Informação (SRI) ocorre mediante uso de protocolos, tais como Z39.50 e OAI-PMH. Esses protocolos definem padrões para procedimentos e funcionalidades de busca e Recuperação da Informação. É possível utilizar estes protocolos em diferentes plataformas, como por exemplo: DSpace, software utilizado para desenvolver repositórios de informação.

Segundo Oliveira e Carvalho (2009), o protocolo OAI-PMH foi criado pela *Open Access Initiative* (OAI), com o objetivo de facilitar a coleta de dados entre repositórios

digitais, possibilitando o compartilhamento de metadados. O protocolo Z39.50 define padrões de interoperabilidade para diversos sistemas de informação em uma única interface. Esse protocolo permite a busca e a Recuperação da Informação em diversos formatos (OLIVEIRA; CARVALHO, 2009).

Nos dias atuais, a interoperabilidade de conteúdo não é apenas um conceito, mas uma necessidade. Com o avanço das tecnologias da informação e, principalmente, com a explosão informacional trazida pela Internet, é indispensável o desenvolvimento de linguagens, protocolos e sistemas cada vez mais interoperáveis.

Os desafios das novas tecnologias para RI não param de crescer. Não é suficiente trabalhar na Recuperação da Informação textual. Novas mídias e formas de armazenar informações têm se tornado cada vez mais presentes: imagens, sons, vídeos, figuras multidimensionais, fractais, webpages e outras formas de armazenamento demandam tratamento e Recuperação da Informação diferenciada e efetiva (BURKE, 1999).

Segundo Ferneda (2003), particularmente no contexto da web, uma das principais mudanças é a “desterritorialização do documento e a sua desvinculação de uma forma física tradicional como o papel” o que possibilita uma integração entre diferentes suportes (texto, vídeo, imagem, som) e uma mudança de paradigma nas formas de acesso aos documentos na web. A utilização de técnicas da Inteligência Artificial surgiu por consequência da evolução dos modelos matemáticos aplicados ao tratamento semântico dos textos. É a tecnologia possibilitando o aperfeiçoamento de antigas ideias. Como é o caso do modelo booleano estendido, em uso nos motores de busca da web, implementado para potencializar os processos de Recuperação da Informação na Internet.

Para Araujo-Junior (2006), a Internet e a web possibilitaram um rompimento parcial das divergências de interesse nas pesquisas voltadas à Recuperação da Informação, com a democratização do acesso à informação. A web promoveu um rápido direcionamento nos esforços de pesquisa dos mais variados campos científicos para os problemas relacionados à recuperação de informação. Se muitas vezes a obra de Paul Otlet é criticada por seu centralismo autoritário e seu monumentalismo, o que vemos na web são problemas gerados por uma exagerada “democracia informacional” em uma dimensão que supera o “monumental” (FERNEDA, 2003, p.13).

Os usuários têm necessidades de informação e constroem, por meio dos motores de busca, expressões que as representem. Essas necessidades de informação podem ser especificadas em linguagem natural ou por meio de uma linguagem artificial, e devem resultar na recuperação de um número de documentos que possibilite a verificação daqueles que são úteis.

2.2.3 Sistemas de Recuperação da Informação (SRI)

Um sistema de recuperação de informação (SRI) pode ser definido como um conjunto padronizado de dados armazenados em meio eletrônico, utilizados para identificar informação e fornecer a localização de informações (ORTEGA, 2002). Reescrevendo, o objetivo de um SRI é permitir que um usuário possa recuperar documentos por meio das características específicas do próprio documento como: palavras-chaves, autor, título, assunto e combinação de expressões.

No princípio, sistemas de RI baseavam-se no cálculo da frequência de palavras contidas no texto e também na eliminação de palavras de pouca relevância (ARAÚJO-JUNIOR, 2006). Na década de 1960 o pensamento geral era que os métodos puramente estatísticos seriam suficientes para tratar das questões relativas à recuperação de informação. No entanto, percebeu-se a necessidade de novos métodos de busca que possibilitassem uma análise semântica mais precisa. Salton (1983) tem se mostrado interessado, desde seus primeiros trabalhos, pela utilização de processos de tratamento da linguagem natural na recuperação de informação.

Em 1983, Salton e McGill apresentaram um artigo intitulado “*Future directions in Information Retrieval*”, que tratava da aplicação do processamento da linguagem natural e da lógica fuzzy na recuperação de informação, indicando a realização de futuras pesquisas sobre Inteligência Artificial (FERNEDA, 2003).

Face ao exposto, destaca-se que a eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que utiliza e que a grande maioria dos modelos de recuperação de informação é de natureza quantitativa.

2.2.4 A evolução do SRI

Segundo Ferneda (2003), um processo de RI eficiente está mais preocupado com a satisfação da necessidade do usuário do que com a correta resposta à instrução ou expressão de busca inserida no SRI: “o processo de recuperação de informação consiste em identificar, no conjunto de documentos (corpus) de um sistema, que atendem à necessidade de informação do usuário” (FERNEDA, 2003, p. 14).

Nesse contexto, o usuário de um SRI está muito mais interessado em, de fato, recuperar a informação, independentemente do assunto tratado, do que na recuperação de dados que simplesmente satisfaçam a sua expressão de busca, mesmo que seja um grande conjunto de dados. Essa é a diferença básica entre um SRI e um Sistema Gerenciador de Bancos de Dados (SGBD) e, apesar de ambos armazenarem grandes conjuntos de dados e documentos, estruturados ou não, têm aplicações e características completamente distintas.

Os sistemas de Recuperação da Informação, segundo Lancaster (2004, p. 202, apud LOPES, 2002), evoluíram em duas grandes linhas. A primeira tem origem nos grandes sistemas e bancos de dados norte-americanos: *National Library of Medicine* - NLM, *Department of Defense* - DOD e da NASA (*National Aeronautics and Space Administration*), que indexavam os registros informacionais das bases de dados por meio de modelos específicos de cada área temática. Já outra grande linha desenvolveu-se no âmbito do direito para a organização de grandes volumes de textos das leis, acórdãos, decisões e processos oriundos das ações judiciárias.

Para uma eficiente Recuperação da Informação, é necessária uma análise acurada de assunto e sua consequente organização, de forma a tornar possível seu exame por meio de princípios sistemáticos e sob diferentes pontos de vista (FERNEDA, 2003). Os SRI precisam recuperar “unidades de informação” que são materializadas por combinações de vários aspectos. Assim, uma análise precisa da informação possibilitará que classificações, organizações, ideias, noções, juízos, que constituem o conhecimento, sejam extraídos da leitura, identificando, caracterizando e organizando uma composição adequada.

Portanto, são duas linhas distintas, uma relacionada aos bancos de dados referenciais e indexados e a outra às bases de dados textuais, o que demandará um planejamento acurado das estratégias de busca e Recuperação da Informação. A dificuldade se acentua na medida em que estas estratégias de RI requerem flexibilidade suficiente para atender às necessidades de informação singulares de cada usuário.

Nesse prisma, torna-se imprescindível aos requisitos de um SRI efetivo a possibilidade de criar inter-relações semânticas entre conceitos e ideias contidas nos textos. Os antigos sistemas que buscam, contabilizam e recuperam conjuntos de dados apenas pela quantidade de ocorrências nos registros não são mais aceitáveis. Na mesma linha pensa Datta (1977, p. 1): “Os velhos esquemas não são satisfatórios, sendo necessário um novo tipo de sistema ou classificação que possa mostrar claramente, de uma forma analítica, a complexidade do conhecimento”. A autora complementa que um SRI “deverá refletir a maneira pela qual o conhecimento é realmente adquirido e estruturado, baseando-se na evidência científica de seu desenvolvimento. O esquema deverá representar uma organização adequada de conceitos" (p. 1).

Os sistemas de recuperação de informação devem representar o conteúdo dos documentos da base de dados e apresentá-los ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfaçam total ou parcialmente a sua necessidade de informação, formalizada por meio de uma expressão de busca. O desafio atual dos sistemas de Recuperação da Informação é a integração do conhecimento por meio da interoperabilidade de conteúdo, seja dentro de uma única organização ou integrando bases de dados espelhadas geograficamente. Os SRI funcionam de acordo com modelos lógicos. Estes modelos são classificados em diferentes categorias e podem influenciar a eficiência dos sistemas (FERNEDA, 2003).

No ano de 2010, os inventores Onno Zoeter, Michael J. Taylor, Edward Lloyd Snelson, John P. Guiver, Nicholas Craswell e Martin Szummer, registraram a patente “US 8037043 B2” (também publicada como “US 201000769 49”), cedida à Microsoft Corporation, uma das líderes globais no segmento de tecnologias da informação. A patente trata da invenção de um Sistema de Recuperação da Informação preditivo,

para recuperar uma lista de documentos, tais como páginas web, arquivos de texto, PDF ou outros itens de uma base indexada em resposta a uma consulta do usuário.

Foi desenvolvido um motor de predição que é usado para prever tanto a informação relevante explícita, tais como etiquetas de julgamento, como também a informação implícita relevante, identificadas pelos dados dos cliques dos usuários. De forma prática, a informação relevante predita é aplicada a uma função de determinação da utilidade da informação e posterior armazenamento, que descreve a satisfação do usuário utilizador do objeto informacional recuperado, por meio de uma sessão de pesquisa. Isso produz pontuações e ranqueamento da utilidade para a proposição de listas de documentos.

Atualmente (2016), esses princípios são utilizados em larga escala nos motores preditivos de recuperação da informação. O uso das notas e ranking de utilidade da informação é fundamental à listagem dos documentos que serão selecionados pelo usuário do SRI. Dessa forma, diferentes fontes de informações relevantes são combinadas em um único Sistema de Recuperação da Informação de modo eficaz, garantindo ao usuário o melhor desempenho.

2.2.4.1 Modelos Quantitativos de SRI

A maioria dos modelos de Sistemas de Recuperação da Informação são classificados como quantitativos. Modelos deste tipo são baseados em disciplinas como lógica, estatística, matemática e teoria dos conjuntos (FERNEDA, 2003).

O entendimento dos princípios básicos dos modelos quantitativos de um SRI é primordial para a compreensão e posterior aplicação dos processos de recuperação da informação utilizando as ferramentas de mineração de textos, facilitando o alcance dos objetivos dessa pesquisa.

2.2.4.2 Modelo Booleano

A lógica aristotélica baseava-se na diferenciação entre verdadeiro e falso para explicar a realidade. Este pensamento foi consolidado como a Lógica Booleana, que possibilitou o desenvolvimento de vários campos científicos e provocou avanços tecnológicos como a linguagem binária, baseada em estágios de zero e um, ligado e desligado, falso e verdadeiro. Não teria sido possível o desenvolvimento da eletrônica e da computação sem a compreensão desta lógica.

Segundo Camargo (2009), um sistema de RI booleano pesquisa os índices da base de dados por meio de uma pesquisa binária e tem como ponto forte a velocidade de resposta, além do baixo custo computacional para processamento (CAMARGO, 2007). Uma grande desvantagem do modelo booleano é a incapacidade em ordenar os documentos recuperados da busca. Logo, o modelo não seria adequado aos modernos sistemas de texto integral, como os mecanismos de busca da web, nos quais o ranqueamento dos documentos é essencial, em função do grande volume de documentos que geralmente é recuperado neste tipo de busca (FERNEDA, 2003).

2.2.4.3 Modelo Vetorial

No modelo vetorial, um registro informacional é representado por um vetor em que cada elemento representa o peso, ou a relevância, do respectivo termo de indexação para o documento. Segundo Ferneda (2003, p. 28) cada vetor descreve a posição do documento em um espaço multidimensional: “cada termo de indexação representa uma dimensão ou eixo”. Assim, cada elemento do vetor é normalizado e ranqueado, assumindo valores indicativos entre zero e um. Os valores mais aproximados de 1 indicam termos com maior importância para a descrição do registro informacional.

Ainda, segundo o mesmo autor, diferentemente do modelo booleano, o modelo vetorial utiliza pesos tanto para os termos de indexação, quanto para os termos da expressão de busca. Esta característica permite o cálculo de um valor numérico que representa a relevância de cada documento em relação à busca (FERNEDA, 2003).

O maior benefício do modelo vetorial é a definição de um modelo conceitual, componente essencial em qualquer teoria científica. Desse modelo surgiu o projeto SMART – *System for the Manipulation and Retrieval of Text* (SALTON, 1971). O Sistema SMART foi uma implementação bem-sucedida do modelo vetorial onde cada documento que se busca é representado por um vetor ordenado pela importância e relevância dos termos da descrição do documento. Este padrão ainda é utilizado como referência para implementação de sistemas de Recuperação da Informação e pesquisas, principalmente no meio acadêmico.

2.2.4.4 Modelo Probabilístico e Modelo Fuzzy

O termo probabilidade deriva do Latim *probare* (provar ou testar). Em essência, existe um conjunto de regras matemáticas para manipular a probabilidade e outras regras para quantificar a incerteza, como a teoria de Dempster-Shafer e a lógica difusa, “fuzzy logic” (GRINSTEAD e SNELL 1996). Em estatística, a teoria das probabilidades estuda os experimentos aleatórios que, repetidos em condições idênticas, podem apresentar resultados diferentes e imprevisíveis. Quando observa-se a face superior de um dado, ou quando verifica-se o naipe de uma carta retirada de um baralho apenas se pode estimar a possibilidade ou a chance de um evento advir.

O intuito da lógica fuzzy é capturar e operar com a diversidade, a incerteza e as verdades parciais dos fenômenos da natureza de uma forma sistemática e rigorosa (SHAW; SIMÕES, 1999). Em teoria, os conjuntos existentes no mundo real não possuem limites precisos. Já um conjunto fuzzy é um agrupamento indefinido de elementos, no qual a transição de cada elemento de não-membro para membro do conjunto é gradual. Esse grau de imprecisão de um elemento pode ser visto como uma “medida de possibilidade”, ou seja, a “possibilidade” de que um elemento seja membro do conjunto. Assim, um modelo Fuzzy de RI trará os registros informacionais que possivelmente terão a informação que o usuário do SRI almeja.

A proposta de modelo probabilístico de Robertson e Jones (1976), posteriormente conhecido como *Binary Indepence Retrieval*, busca demonstrar a

recuperação de informação sob um enfoque meramente probabilístico, ou seja, dada uma expressão de busca qualquer fornecida pelo usuário do SRI, o modelo deve recuperar os documentos com a maior probabilidade possível de conter a informação desejada.

Nesse contexto, em uma expressão de busca, pode-se dividir a base de dados explorada em quatro subconjuntos distintos: O conjunto dos documentos recuperados; o conjunto dos documentos relevantes; o conjunto dos documentos relevantes que foram recuperados; e o conjunto dos documentos não relevantes e não recuperados. Conclui-se, então, que o conjunto dos documentos relevantes e recuperados tem a maior probabilidade de conter a informação desejada pelo usuário do sistema de Recuperação da Informação.

2.2.5 Modelos Dinâmicos

No processo de recuperação de informação, os modelos quantitativos estabelecem uma restrita representação dos documentos e ativos informacionais; é a representação gerada pela associação de termos de indexação e respectivos pesos aos documentos da base de dados. O problema é que estes modelos são limitados, impositivos e unilaterais, e não possibilitam a intervenção do usuário na representação dos registros informacionais (FERNEDA, 2003).

Os modelos dinâmicos trazem novas possibilidades e favorecem a geração de novas relações entre os documentos para uma RI mais eficiente. Os principais modelos dinâmicos são: Sistemas especialistas, redes neurais e algoritmos genéticos. À medida que o usuário aprofunda-se na busca e no processo de recuperação, pode redefinir dinamicamente as expressões, condições e caminhos que o sistema percorrerá dentro da base de dados para resultar no melhor conjunto de documentos.

Os processos de análise de dados em formato não estruturado são, seguramente, atividades mais complexas, ao compararmos com a análise de dados estruturados, justamente pela “não estruturação” dos dados (MORAIS; AMBRÓSIO 2007). Isto se dá principalmente no que se refere aos atributos técnicos, estruturais e negociais desses dados pela falta de metadados.

Conforme Beppler et al. (2005), a descoberta de conhecimento em textos, *(KDT – Knowledge Discovery from Text),* contempla as técnicas e ferramentas inteligentes e automáticas que apoiam a análise de grandes volumes de dados com o objetivo de "minerar" o conhecimento útil, beneficiando qualquer domínio que utilize textos não estruturados. Os sistemas de KDT processam um considerável volume informações e produzem uma grande quantidade de "padrões", que nem sempre serão úteis ao usuário (MORAIS; AMBRÓSIO 2007). Justifica-se, então, a aplicação de métodos qualitativos, como por exemplo a AC, para complementar a análise. Nesse contexto, KDT, mineração de textos e Recuperação de Informação, são altamente dependentes do processamento de linguagem natural e linguística computacional.

Nesta pesquisa, será realizado o processamento de toda a base documental do acervo do Prêmio Professor Samuel Benchimol. Este processamento será feito em um sistema de mineração de textos, para que se possa analisar e identificar os padrões e analisar os dados de forma quantitativa e qualitativa.

Ao utilizar os recursos de mineração de textos, não serão realizadas buscas, mas sim, análises dos documentos. Contudo, não se espera como resultado o conhecimento por si. É primordial que o resultado ainda seja analisado e contextualizado, o que possibilitará posterior descoberta de conhecimento (MOURA, 2004). O detalhamento dos processos de mineração ocorrerá segundo passos descritos na seção Metodologia.

2.3 BIBLIOMETRIA

Na contemporaneidade, a explosão cientifico-tecnológica trouxe consigo uma grande quantidade de avanços e inovações que demandaram novas estratégicas para a avaliação e determinação dos desenvolvimentos alcançados em cada área de conhecimento. As últimas décadas foram marcadas pela expansão das diversas formas de medição e avaliação da ciência e dos fluxos informacionais, entre elas: Bibliometria, Cienciometria, Informetria e Webometria, onde cada uma destas áreas

possui suas especificidades. Esta autora define ainda, em termos genéricos, as possibilidades de aplicação dessas áreas:

> – identificar as tendências e o crescimento do conhecimento em uma área; – identificar as revistas do núcleo de uma disciplina; – mensurar a cobertura das revistas secundárias; – identificar os usuários de uma disciplina; – prever as tendências de publicação; – estudar a dispersão e a obsolescência da literatura científica; – prever a produtividade de autores individuais, organizações e países; – medir o grau e padrões de colaboração entre autores; – analisar os processos de citação e co-citação; – determinar o desempenho dos sistemas de Recuperação da Informação; – avaliar os aspectos estatísticos da linguagem, das palavras e das frases; – avaliar a circulação e uso de documentos em um centro de documentação; – medir o crescimento de determinadas áreas e o surgimento de novos temas (VANTI, 2002, p. 152).

O uso de técnicas bibliométricas contribui de forma decisiva, em épocas de recursos escassos, quando um bibliotecário deve resolver que títulos ou publicações periódicas podem ou não ser suprimidas de uma biblioteca. Indicadores de uso são obtidos, assim, para definir uma lista de publicações periódicas prioritárias e para prever a demanda futura. É fundamental ter como detectar a utilização real dos títulos que constam em uma biblioteca, possibilitando determinar a obsolescência das coleções.

Nesse contexto, a Bibliometria, desenvolve-se inicialmente a partir da construção de leis relativas ao comportamento da literatura, também chamadas, leis bibliométricas empíricas (TAGUE-SUTCLIFFE, 1992), cronologicamente: i) Método de medição da produtividade de cientistas de Lotka (1926); ii) A lei de dispersão do conhecimento científico de Bradford (1934); iii) O modelo de distribuição e frequência de palavras num texto de Zipf (1949).

A base dessas leis é a aplicação de técnicas estatísticas e modelos matemáticos para produzir inter-relacionamentos e descrições de aspectos da literatura e de textos processados ou analisados de forma predominantemente quantitativa. Nesse contexto, a Bibliometria caracterizava-se como uma evolução do antigo conceito de “bibliografia estatística” de Hulme, cunhado em 1923, sendo o termo “Bibliometria” criado por Otlet em 1934 no documento “Traité de Documentation”. Todavia, o conceito materializou-se após a década de 1970, quando Pritchard explicou as diferenças e discutiu a questão em seu trabalho “bibliografia estatística ou bibliometria?” (VANTI, 2002, p. 153). Segundo os autores Nicholas e Ritchie (1978, p. 38), o debate essencial na diferenciação entre a tradicional

Bibliografia e a Bibliometria é que a última utiliza-se mais de métodos quantitativos do que de processos discursivos. A Bibliometria vale-se de métodos quantitativos na busca por uma avaliação objetiva da produção científica. Torna-se clara a percepção de Araújo (2006) da importância dessa área para a análise da distribuição e dos números de autores, trabalhos, temas, semantemas, países, periódicos, revistas entre outros, possibilitando a categorização, a classificação e a análise de produtividade, entre diversas possibilidades de uso dos dados quantitativos gerados.

Segundo Figueiredo (1977), a Bibliometria, em essência, possui duas ramificações básicas: a primeira, focada na análise da produção científica, procura aplicações práticas imediatas para bibliotecas (incremento de coleções, gerenciamento de serviços bibliotecários); a segunda, que foca o melhoramento do controle bibliográfico (entender o volume e as distinções dos acervos, preparar predições para o aumento entre outros), é a ramificação mais óbvia a que se refere a aplicação atual da Bibliometria.

2.3.1 Leis Clássicas da Bibliometria

As leis bibliométricas são proposições baseadas em métodos matemáticos e estatísticos para análise e construção de indicadores quantitativos relativos à dinâmica da informação e da produção científica. As leis tratam das relações entre a informação, autores e conteúdos das mensagens, na tentativa de quantificar os processos de comunicação escrita. As três leis clássicas empíricas da Bibliometria são:

2.3.1.1 Lei de Lotka

A Lei bibliométrica de Lotka surgiu como resultado de um estudo sobre a produção de cientistas, para fazer a volumetria das ocorrências de autores no *Chemical Abstracts*, no início do século XX (ARAUJO, 2006). Alfred J. Lotka mapeou a frequência de publicação de autores em um determinado campo e afirmou que o número de autores “a” que fazem “n” contribuições é de cerca de $1/n^a$ das pessoas que fazem uma única contribuição, onde “n/aproximadamente 2 (dois)”.

Reescrevendo, o número de autores que publicam certo número de artigos possui uma relação fixa com o número de autores que publicam um único artigo.

À medida que o número de artigos publicados aumenta, os autores que produzem muitas publicações tornam-se menos frequentes. Apesar de a própria Lei abranger muitas disciplinas, as proporções reais envolvidas (como uma função de "a") são muito singulares para cada área do conhecimento. A lei Geral diz:

$$X^n = C \text{ ou } Y = C/X^n$$

Onde “X” é o número de publicações, “Y” a frequência relativa dos autores com publicações “X”, e “n” e “C” são constantes em função do domínio específico em que “n/aproximadamente 2”. Tem-se, então, que uma porção da produção literária científica é oriunda de um seleto grupo de autores, e um alto volume de simples autores igualam-se, em trabalhos científicos, a um pequeno grupo de autores muito produtivos. Com o estabelecimento dessa “Lei”, a partir da década de 1930, muitas pesquisas foram realizadas no sentido da investigação da produtividade dos autores em múltiplas áreas do conhecimento. Urbizagástegui Alvarado (2002) mapeou até dezembro do ano 2000 mais de 200 pesquisas, entre teses, dissertações, artigos, capítulos de livros, comunicações a congressos, e monografias que tinham sido produzidas no sentido de confirmar, refutar, criticar, replicar ou reformular essa lei bibliométrica.

2.3.1.2 Lei de Bradford

A lei empírica bibliométrica de Samuel C. Bradford é aplicada aos conjuntos de periódicos científicos. Surgiu com o objetivo de investigar a extensão na qual artigos que tratam de um determinado contexto ou assunto científico específico ocorriam em periódicos científicos específicos de outras áreas do conhecimento, comparando a distribuição dos artigos no que se refere às variáveis de proximidade ou de afastamento (ARAÚJO, 2006). Foi então que, em 1934, após uma série de estudos, Bradford formulou a lei da dispersão em três zonas.

Bradford, por meio de sua lei, estima os retornos decrescentes de forma exponencial ao alargar a busca de referências em revistas científicas. Em muitas disciplinas esse padrão é chamado de distribuição de Pareto. Como exemplo prático, se um pesquisador tem cinco principais revistas científicas para o seu campo de pesquisa e, por hipótese, em um mês há doze artigos de interesse nessas revistas e esse pesquisador ainda necessita encontrar mais uma dúzia de artigos correlatos, o pesquisador teria que ir a um adicional de dez novos periódicos. Logo, o multiplicador de Bradford "BM" – *Bradford's Multiplier* –, do pesquisador é igual a 2 (ou seja, 5/10).

Nessa esteira, para o pesquisador localizar cada nova dúzia de artigos de seu interesse, ele terá que dobrar o número de revistas pesquisadas, abrangendo um número cada vez maior de revistas. Para selecionar doze artigos ele olharia apenas em cinco revistas, mas para selecionar sessenta artigos, por exemplo, ele não teria que olhar vinte e cinco revistas e sim oitenta revistas, e assim sucessivamente, a maioria dos pesquisadores rapidamente percebe que não há sentido em olhar mais longe. A proposta de Bradford sugere que:

$$BM = 2$$

Com o avanço da pesquisa, essa lei também foi objeto de reformulação e aperfeiçoamento. Em 1948, Vickery propôs que o número de zonas não precisa ser três, mas qualquer número. Outros aperfeiçoamentos foram feitos por outros autores no sentido de corresponder à realidade específica de cada literatura científica, tais como Leimkuhler, em 1967, Fairthorne, em 1969, e Goffman e Warren em também em 1969 (DONOHUE, 1973, p. 17 apud ARAÚJO, 2006).

2.3.1.3 Lei de Zipf

George Kingsley Zipf propôs uma lei bibliométrica empírica baseada em estatística e matemática. Refere-se ao fato de que muitos tipos de dados estudados em ciências físicas e sociais podem ser aproximados com uma distribuição "Zipfian". A terceira das leis, formulada em 1949, mostra o relacionamento entre palavras num determinado texto de tamanho suficientemente grande e uma ordem de série das palavras. A lei é mais facilmente observada por meio da representação gráfica dos

dados em um gráfico “log-log”, com os eixos sendo log (ordem de classificação) e log (frequência).

Por exemplo, a palavra “a” (como descrito acima) que aparece em x = log (1), y = log (69,971). Também é possível traçar classificação recíproca contra a frequência ou a frequência recíproca ou intervalo entre palavras. Os dados estão em conformidade com a lei de Zipf, na medida em que o enredo é linear. Composição: “N” é o número de elementos; “k” a classificação; “s” o valor do expoente que caracteriza a distribuição.

A lei de Zipf, em seguida, prevê que, de uma população de “N” elementos, a frequência de elementos de posto k, f (k; s, N) é:

$$f\,(k; s, N) = \frac{1/k^s}{\Sigma^N n=1\ (1/n^s)}\ .$$

Zipf analisou a obra *Ulisses, de James Joyce,* e encontrou uma conexão entre o número de palavras diferentes e a constância de seu uso, concluindo que existe uma simetria fundamental na seleção e uso de palavras (ARAÚJO, 2006). Reescrevendo, um pequeno conjunto de palavras é utilizado com muita frequência e um grande conjunto de palavras é logrado com baixa recorrência. No texto de *Ulisses* ele percebeu que a palavra mais utilizada aparecia 2653 vezes, que a centésima palavra mais utilizada ocorria 256 vezes, e ainda que a ducentésima palavra mais frequente ocorria 133 vezes. Zipf observou, então, que a disposição de uma palavra ponderada pela sua presença era igual a uma constante de $\pm$ 26500. A equação para esse relacionamento é:

$$R \times F = K$$

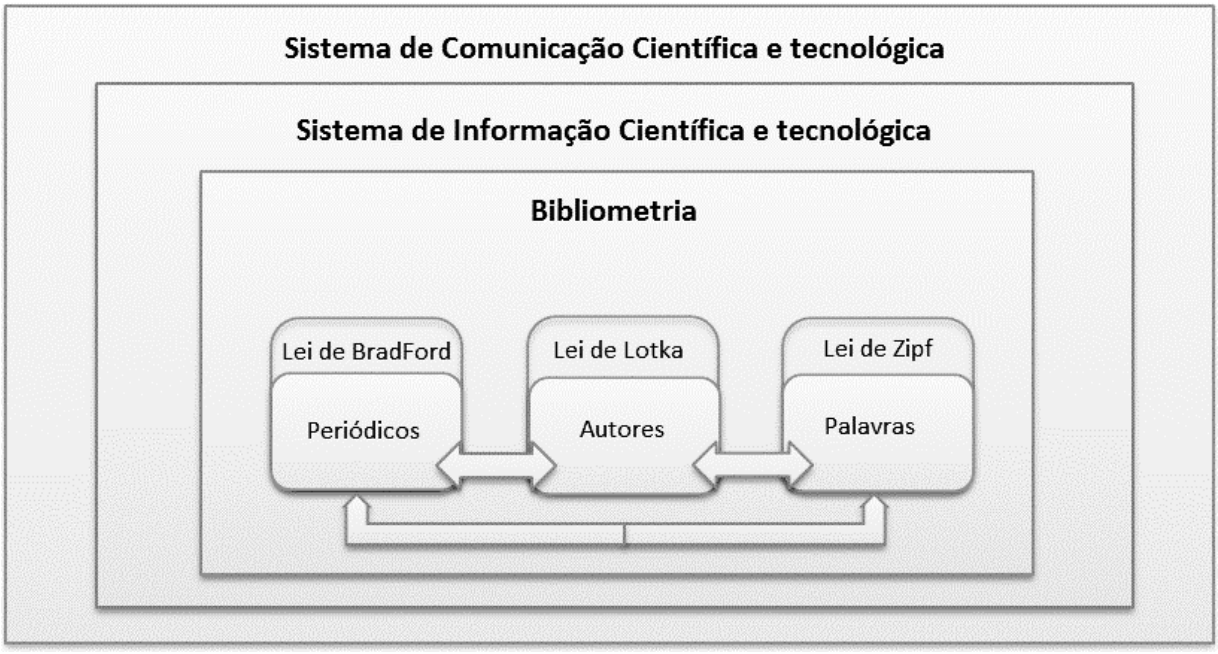
Onde “R” é a posição da palavra, “F” é a sua frequência e “K” é a constante. A partir deste ponto, Zipf estabeleceu o princípio do “menor esforço”. A ideia é que existe uma inconsciente economia do uso de palavras, a tendência de usar o mínimo significa que elas não vão se espalhar, justo o oposto, uma mesma palavra vai ser usada com muita frequência. Esta teoria também foi bastante revisitada. Kendall, foi um dos que propôs um paralelo entre Zipf, Bradford e Brookes, criador da distribuição Bradford/Zipf. Brookes argumenta que um grande número de “fontes” contribui com “itens” casuais para um “campo” determinado. Outros autores que acrescentaram

contribuições à Lei de Zipf: Booth, Donohue e Mandelbrot (RAO, 1986, p. 181, apud ARAÚJO, 2006).

As leis bibliométricas em seus ensaios e métodos vêm sendo aprimoradas e aplicadas em diferentes contextos, principalmente no core de software que realizam processamento e análise automatizada de textos. Nessa esteira, Guedes e Borschiver (2009) defendem que os estudos que utilizam a frequência de ocorrência de palavras como ferramenta de representação temática da informação têm evoluído com o propósito de facilitar a criação de algoritmos que possam contribuir para automatização, "[...] em parte ou no todo, da indexação temática da informação".

A figura a seguir apresenta uma representação espacial entre as três principais leis bibliométricas nas suas áreas de estudo, inserindo-as em um sistema de informação científico-tecnológica e este em um super-sistema de comunicação científico-tecnológica.

Figura 1 – Relação das leis clássicas da Bibliometria



Fonte: Adaptado de Guedes e Borschiver (2009).

2.4 ANÁLISE DE CONTEÚDO (AC)

Nos Estados Unidos da América, o desenvolvimento da AC como metodologia iniciou-se durante a Segunda Guerra Mundial, quando o governo estadunidense patrocinou projetos de Harold Lasswell, visando melhorar a avaliação e a análise da propaganda inimiga. Assim, “Os recursos alocados para a pesquisa e avanços metodológicos no contexto dos problemas sob investigação contribuíram de forma significativa para a emergência da metodologia na Análise de Conteúdo” (ROSSI, SERRALVO, JOÃO 2014). A Análise de Conteúdo é utilizada de forma quantitativa, com dados de texto codificados em grupos explícitos e, em seguida, descritos com o uso de modelos estatísticos, uma abordagem por vezes conhecida como “análise quantitativa de dados qualitativos” (MORGAN, 1993).

A Análise de Conteúdo correlaciona símbolos com seus significados em duas abordagens: Qualitativas e quantitativas. Na abordagem qualitativa, é utilizada a inferência, buscando uma interpretação aprofundada. Já na abordagem quantitativa, a AC estuda as medidas de análise de frequência de termos e conceitos.

Para Olabuenaga e Ispizúa (1989), a Análise de Conteúdo é uma técnica que possibilita a leitura e a interpretação de conteúdo de toda classe de documentos que, analisados adequadamente, possibilitam chegar ao conhecimento de aspectos e fenômenos da vida social de outro modo inacessíveis. Atualmente, as possibilidades de uso da AC como método de análise qualitativa têm sido reconhecidas, o que provocou um aumento de suas aplicações e democratização do seu uso (NANDY e SARVELA, 1997). A Análise de Conteúdo tem sido aplicada de forma qualitativa, como um método de pesquisa para o entendimento subjetivo do conteúdo de dados e textos, por meio de processos de classificação sistemática de codificação e da identificação de assuntos ou padrões (HSIEH E SHANON, 2005).

A Análise de Conteúdo (AC), de acordo com a tradição, é uma técnica predominantemente quantitativa – conteúdo manifesto. Todavia, há também a qualitativa, que vem crescendo a passos largos (GRANEHEIM, LUNDMAN, 2003). A vertente qualitativa lida com aspectos de relações e envolvem a interpretação de significações ocultas do texto – conteúdo latente –, (DOWNE-WAMBOLDT, et al., 1992).

Nas duas abordagens, latente e manifesto, a AC lida com interpretação. Todavia, a interpretação varia em profundidade e nível de abstração. Neste contexto, a primeira decisão quanto à AC é se esta será aplicada quantitativamente ou qualitativamente ou ainda, se utilizará de ambas as abordagens (BABBIE, 1998).

Antes de adentrar às características da AC, faz-se necessária a diferenciação entre Análise de Conteúdo e Análise de Discurso (AD). Ambas apresentam metodologias bastante diferenciadas e objetivos distintos. AD pretende compreender e refletir acerca dos discursos, para além daquilo que é inerente ao texto. Por exemplo, o tom da voz e a comunicação paralinguística (VILELA JUNIOR, 2015). Na Análise de Conteúdo, o componente de estudo é o registro em si, pertencente a um texto, a um arquivo, a uma fala, a um áudio ou a um vídeo. Assim, pode-se afirmar que a Análise de Conteúdo está contida na Análise do Discurso. Contudo, o contrário não é verdadeiro.

Laville e Dionne (1999, p. 214) afirmam que o princípio da AC consiste em desmontar a estrutura e os elementos de um material informacional bruto, previamente organizado, empreender um minucioso estudo de seu conteúdo, das palavras e frases que o compõem, procurar-lhes o sentido, captar-lhes as intenções, comparar, avaliar, descartar o acessório, reconhecer o essencial, e selecioná-lo em torno das ideias principais, esclarecer suas diferentes características e extrair sua significação.

Chizotti (2006, p. 115) explica que a AC construiu um conjunto de procedimentos e técnicas de extrair o sentido de um texto por meio de unidades elementares que compõem produtos documentários: palavras-chave, léxicos, termos específicos, categorias, temas e semantemas, procurando identificar a frequência ou constância dessas unidades para, então, fazer inferências e extrair os significados inscritos no texto a partir de indicadores objetivos.

Esse autor aponta ainda que a AC parte do pressuposto de que o léxico, um vocábulo que é uma unidade discreta do texto, constitui uma síntese condensada da realidade. Assim, a frequência de seu uso pode revelar a concepção de seu emissor, os seus valores, opções, preferências. Conclui que é possível, também, fazer uma leitura do contexto e das circunstâncias em que a mensagem foi feita, autorizando uma leitura subjacente ao texto, daquilo que está além do que é manifesto, do que é

preterido: as omissões, as ignorâncias consentidas, as preferências seletivas por palavras, os termos ambíguos, ou seja, os significados subjacentes que o texto contém (CHIZOTTI, 2006, p.117).

No que diz respeito à interpretação de textos, Campos (2004) afirma que “produzir inferências sobre o texto é a razão de ser da Análise de Conteúdo”. Significa não somente produzir suposições subliminares a respeito das mensagens, todavia, em embasá-las com pressupostos e bases teóricas de diversas visões de mundo e com as situações concretas que vivem os produtores ou receptores das mensagens, situações que podem ser visualizadas observando o contexto histórico e social de produção e da recepção das mensagens (CAMPOS, 2004, p. 613).

Para Amado (2000, p. 59), entretanto, é possível assegurar um equilíbrio entre a Análise de Conteúdo quantitativa (com cálculos de frequência, percentagens, correlações, análise fatorial, em função das distribuições da amostra e das hipóteses levantadas) e a análise qualitativa (com a descrição das características, independentemente de sua frequência relativa no texto), ou tender para um ou outro lado, conforme as exigências e os objetivos da investigação. Nessa direção, esse autor analisa as ideias de Krippendorf (1990, apud AMADO, 2000), para mostrar como a Análise de Conteúdo “resulta numa verdadeira construção realizada pelo analista”, Leville e Dionne (1999) deixaram pistas a respeito do uso da tecnologia da informação em Análise de Conteúdo, quando afirmaram que “existem software que permitem recuperar e enumerar automaticamente a ocorrência de palavras ou expressões, [mas] os dados assim obtidos permanecem, todavia, superficiais, pois não consideram nem o contexto nem mesmo o sentido exato que uma palavra ou expressão pode ter” (LEVILLE e DIONNE, 1999, p. 217).

Conway (2006), mais recentemente, abordou o uso do computador e enumerou diversos software que auxiliam na Análise de Conteúdo quantitativa, principalmente em análises governamentais: propaganda, eleições, partidos políticos. Porém, Moraes (1999) já apontava para essa tendência, ao afirmar que a metodologia quantitativa “está atingindo novas e mais desafiadoras possibilidades na medida em que se integra cada vez mais na exploração qualitativa de mensagens e informações” (MORAES, 1999, p. 7).

Segundo Ander-Egg (1974), a Análise de Conteúdo possui três fases principais: i) Estabelecer a unidade de análise – que se refere ao elemento básico de análise, relativo às palavras chave ou às proposições acerca dos assuntos abordados; ii) Determinar as categorias de análises – que se refere à segregação e classificação dos conjuntos de dados, categorizando a matéria que trata da identificação dos temas abordados na comunicação; e iii) Selecionar uma amostra do material de análise – que trata dos critérios escolhidos para definição da amostra que será analisada.

Segundo Campos (2004), a AC é um conjunto de técnicas para comunicação e, assim sendo, torna possível estabelecer seus domínios. Infere-se, então, que o método de Análise de Conteúdo é balizado por duas vertentes: De um lado, a linguística tradicional e, do outro, o território da interpretação da significância das palavras, a hermenêutica. No centro, estão os métodos lógico-semânticos, pois os processos de classificação seguem uma "lógica", parâmetros moderadamente delimitados, possibilitando que os analistas possam utilizar estas definições, problemas normais da Lógica.

Quadro 1 - Método de Análise de Conteúdo

Linguística	Métodos lógicos, estéticos e formais	Métodos lógicos semânticos	Métodos semânticos e semânticos naturais	**Hermenêutica**

Fonte: Adaptado de Campos (2004).

Um texto contém muitos significados e, conforme afirmam Olabuenaga e Ispizúa (1989, p. 185): "i) O sentido que o autor pretende expressar pode coincidir com o sentido percebido pelo seu leitor; ii) O sentido do texto poderá ser diferente de acordo com cada leitor; iii) Um mesmo autor poderá emitir uma mensagem, sendo que diferentes leitores poderão captá-la com sentidos diferentes; e iv) Um texto pode expressar um sentido do qual o próprio autor não esteja consciente." A Análise de Conteúdo foi fundamental para auxiliar a interpretação dos valores informacionais

contidos no acervo documental do Prêmio Professor Benchimol, conforme será demonstrado na seção de metodologia deste trabalho.

A Análise de Conteúdo possibilita múltiplas formas para processar e analisar os conteúdos. Para esta pesquisa em questão, serão utilizadas as indicações metodológicas de Bardin (1977). Na pesquisa quantitativa, o que serve de informação é a frequência com que surgem certas características do conteúdo, enquanto na pesquisa qualitativa, o que é considerado é a presença ou a ausência de uma dada característica do conteúdo, ou de um conjunto de características em um determinado fragmento da mensagem (BARDIN, 1977, p. 21). A mesma autora define Análise de Conteúdo como

> conjunto de técnicas de análise das comunicações visando obter, por procedimentos sistemáticos e objetivos de descrição do conteúdo das mensagens, indicadores (quantitativos ou não) que permitam a inferência de conhecimentos relativos às condições de produção/recepção (variáveis inferidas) destas mensagens (BARDIN, 1977, p. 42).

Uma importante aplicação da Análise de Conteúdo na Ciência da Informação está na Bibliometria, pelo seu objeto e pelos seus métodos, uma vez que esta encarrega-se da medição estatística da frequência de palavras em um texto. Para melhor entendimento dessa intersecção, pode-se citar a Lei de Zipf, lei bibliométrica também conhecida como a Lei do Mínimo Esforço, que, como já foi abordado, estima a frequência de ocorrência de palavras em um texto, supondo que um pequeno grupo de palavras ocorra muitas vezes e um grande número de palavras ocorra com reduzida frequência. As palavras são ordenadas em um ranking, de forma decrescente, de acordo com sua frequência no texto.

Dessa forma, conforme Guedes e Borschiver (2005, p. 9-10), além de ser utilizada como ferramenta estatística em diferentes áreas do conhecimento, pode-se aplicar a Lei de Zipf para, em um texto, também identificar os distintos estilos de autores e, ainda, sua sobrecidade ou temacidade, ou seja, sobre que assunto trata um determinado texto.

Percebe-se que a Análise de Conteúdo tem em seu cerne a inferência, a dedução lógica, que é o procedimento que permite ao analista captar, em um tipo de documento, os vestígios que permitirão descobrir “a manifestação de estados, dados e fenômenos, no que diz respeito tanto à procedência da mensagem (a situação ou o

meio em que se encontra o emissor) quanto ao seu destinatário" (BARDIN, 1995, p. 39).

A noção de inferência é especialmente importante na Análise de Conteúdo. De acordo com White e Marsh (2006, p. 27), o pesquisador utiliza constructos analíticos, ou regras de inferência, para mover-se do texto para as respostas às questões da pesquisa. Assim, o texto e o contexto são logicamente independentes, e o pesquisador pode tirar conclusões tanto do texto quanto do contexto.

Também com relação aos conteúdos manifestos (explícitos) do texto, é deles que se deve partir (tal como se manifesta) e não falar "através deles", como também é importante que os resultados da AC devem refletir os objetivos da pesquisa e terem como apoio indícios manifestos no conteúdo das comunicações (CAMPOS, 2004, p. 613).

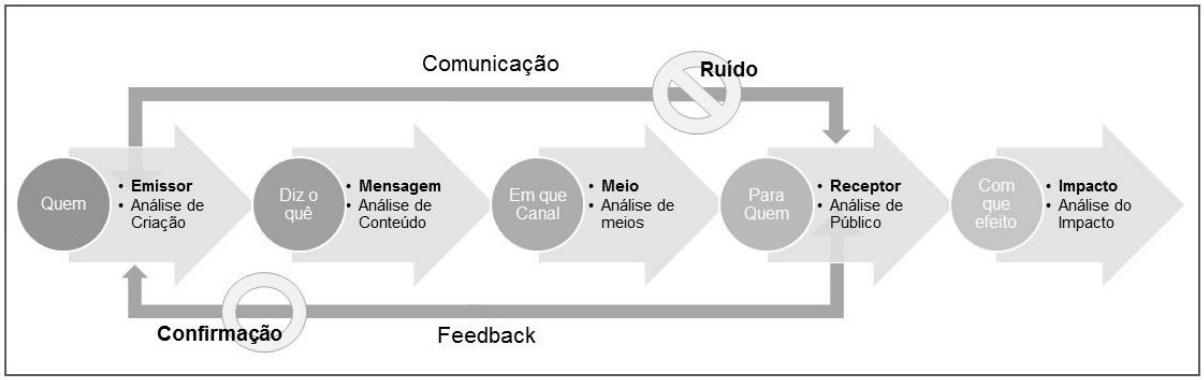
Na Análise de Conteúdo

> a análise qualitativa não rejeita toda e qualquer forma de quantificação. Somente os índices é que são retidos de maneira não frequencial, podendo o analista recorrer a testes quantitativos: por exemplo, a aparição de índices similares em discursos semelhantes. Em conclusão, pode-se dizer o que caracteriza a análise qualitativa é o fato de a inferência -sempre que é realizada -ser fundada na presença do índice (tema, palavra, personagem, etc.), e não sobre a frequência da sua aparição, em cada comunicação individual (BARDIN, 1977).

Moraes (1999, p. 15) destaca que inferir refere-se à pesquisa quantitativa, em que os achados de um estudo, geralmente feitos a partir de uma amostra, são passíveis de generalização para a população da qual a amostra provém, são dados estatísticos, enquanto a interpretação está mais associada à pesquisa qualitativa, à busca de compreensão. A Análise de Conteúdo faz essa busca não só sobre os conteúdos manifestos pelos autores, mas também sobre os conteúdos latentes, aqueles ocultados consciente ou inconscientemente pelos autores.

A Figura 2 abaixo descreve, de acordo com Lasswell, as respostas que se buscam para cada pergunta realizada:

Figura 2 – Modelo de comunicação de Lasswell



Fonte: Elaborado pelo Autor (2016).

A Análise do Conteúdo evoluiu ao longo dos anos, em princípio orientada pelo paradigma positivista, valorizando a objetividade e a quantificação, em que o método era utilizado, por exemplo, nas décadas de 1930 e 1940, para estudos de vocabulário, quantidades de palavras mais utilizadas, opinião pública. Já Funkhouse (1960 apud TRAQUINA, 2005, p. 34), fez uma pesquisa na década de 60 para encontrar os termos-chave da opinião pública (CUNHA, 1983, p. 251), aplicando AC como mecanismo para lidar com o crescente volume de conteúdo disperso e desestruturado.

Minayo (1996, p. 202-203), abordando as tendências históricas das técnicas de Análise de Conteúdo, afirma que todas as metodologias visam ultrapassar o senso comum e o subjetivismo na interpretação e buscam uma interpretação crítica dos documentos, textos literários, biografias, entrevistas ou observação. Para isso, a AC relaciona significantes com significados, em abordagens qualitativas e quantitativas. Na primeira abordagem, usa-se a inferência para uma interpretação mais profunda, enquanto na segunda buscam-se medidas de análise de frequência.

2.4.1 Mineração de textos e descoberta de conhecimento

Os SRI evoluíram continuamente, desde as primeiras iniciativas de simples consultas a tabelas e índices de bancos de dados, até os modernos software de Análise Semântica de Conteúdo e mineração de textos. A Mineração de Textos, que também pode ser conhecida como Descoberta de Conhecimento em Textos

(*Knowledge Discovered in Texts* – KDT), refere-se à extração de informação útil (conhecimento) em suportes de informação textual não estruturados.

Os processos de descoberta de conhecimento em textos são estruturados em duas fases: a primeira etapa consiste no tratamento do texto para convertê-lo à uma estruturada formal; a segunda trata da aplicação efetiva da mineração para a descoberta do conhecimento (CORRÊA, 2003). Dessa maneira, tendo como alvo as bases de informação textual, a mineração de textos disponibiliza ferramentas e técnicas para a extração de padrões ou tendências de grandes volumes de dados em linguagem natural.

A descoberta de conhecimentos junta ferramentas e técnicas inteligentes e automatizadas de análise de grandes volumes de para minerar conhecimento, sendo suporte aos processos de tomadas de decisões e formulação de estratégias organizacionais. Os desafios da KDT, no entanto, estão nos bancos de dados textuais que, de maneira geral, são desestruturados, não sendo possível o uso de técnicas tradicionais de gerenciamento de bancos de dados, assim, métodos específicos para tratamento de textos devem ser utilizados para obter-se conhecimentos não explícitos.

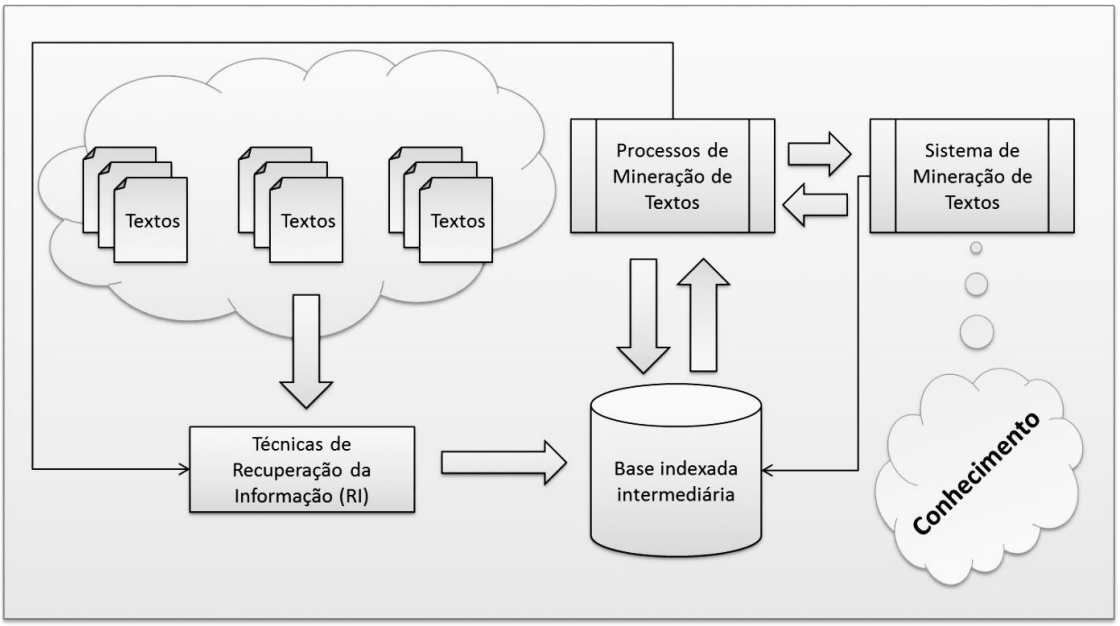
Os processos de mineração de textos envolvem o tratamento de informação textual, de bases de dados estruturadas, semiestruturadas ou não estruturadas, extraindo índices numéricos significativos oriundos do texto e tornando-os acessíveis aos sistemas de descoberta de conhecimento. A mineração de textos consiste na extração de informações acerca de tendências ou padrões em grandes volumes de documentos textuais em que uma amostra significativa de informações é avaliada em textos contidos em bases textuais e em fontes de informação (POLANCO; FRANÇOIS, 2000, apud ARAÚJO JÚNIOR; TARAPANOFF, 2006).

A mineração de textos, segundo Morais e Ambrósio (2007), é um subprocesso de descoberta de conhecimento que, por sua vez, é subprocesso da Recuperação da Informação, que utiliza técnicas de análise e extração de dados a partir de textos.

Os processos de mineração de textos estão diretamente relacionados à recuperação da informação, em que esta ocupa-se das técnicas e ferramentas para buscar documentos, focalizando em dados relacionados a algum tópico específico,

enquanto aquela usa técnicas de Recuperação da Informação em parte do seu fluxo. A Figura 3 a seguir exemplifica esta dinâmica:

Figura 3 – Técnicas de recuperação da informação na mineração de textos



Fonte: Elaborado pelo autor (2016).

Nesse contexto, no projeto de mineração de textos torna-se possível a análise de palavras, conjuntos de palavras ou mesmo o processamento e análise de documentos inteiros entre si, através das similaridades ou das relações semânticas com outras variáveis.

2.4.1.1 Técnicas de mineração de textos

Após os processos para estruturação dos textos, dados e metadados, técnicas de mineração são aplicadas sobre as bases de dados indexadas geradas, com etapas já conhecidas e utilizadas na mineração de dados (processos de Data Mining).

2.4.1.1.1 Regras de associação

A extração de regras de associação é uma técnica de Data mining que também pode ser aplicada à mineração de textos. Ela gera regras do tipo "Se X, Então Y" tendo

por base um determinado banco de dados de transações, onde X e Y são conjuntos de itens que coexistem em várias transações (SANTOS, 2002).

O propósito de um algoritmo que gera regras de associação está associado com o grande volume de aplicações possíveis para as regras geradas (BARION, 2008). Essas regras podem tratar desde questões acerca dos hábitos de consumo até análise de sentimento (positivo ou negativo) de um determinado tema nas mídias sociais, objetivando maximizar as vendas de certos produtos ou tomada de decisão em estratégica (PICHILIANI, 2008):

O algoritmo de regras de associação indica a correlação de termos e, a partir do conhecimento do fato, uma ação pode ser tomada. Pode-se dizer que as regras de associação mostram: X => Y, onde X e Y são conjuntos que podem conter um ou mais termos em um conjunto total (CT) de transações. Esta técnica é largamente empregada na mineração de textos para identificar associações existentes entre termos, classes e categorias de documentos.

Nesse prisma, destaca-se o algoritmo APriori, frequentemente utilizado para detectar associações relevantes entre termos ou itens de dados. O algoritmo APriori quando é aplicado em algum texto encontra conjuntos frequentes de palavras nos documentos. As regras utilizadas são do tipo X => Y, onde X é um conjunto de palavras e Y é uma categoria.

2.4.1.1.2 Sumarização e Clusterização

A sumarização seleciona as informações mais relevantes de um texto, tornando a descrição mais condensada, todavia, mantendo o sentido e a mesma informação. É uma técnica muito utilizada em mineração de textos para localizar termos ou expressões mais relevantes nos documentos. O foco da sumarização está na produção de listas de sentenças dos vários documentos de origem de forma a resumir o conteúdo destes arquivos, reduzindo o tamanho e o volume, contudo preservando o sentido informacional.

Em contraponto, as técnicas de clusterização consistem no agrupamento de conjuntos de dados considerados similares, em clusters ou grupos, gerando matrizes de similaridade. Esta técnica possibilita, por exemplo, a construção de tesauros. O uso das técnicas de clusterização na mineração de texto torna-se importante uma vez que se pode extrair a hierarquia de textos em linguagem natural, agrupando os termos adjacentes ou de suas relações sintáticas, o que permite determinar os termos que carregam um considerável potencial descritivo, podendo inferir a semântica, hierarquia e os conceitos correlacionados aos termos (MOURA, 2004).

2.4.1.1.3 Classificação e Categorização

Os processos de classificação ou categorização consistem em identificar termos mais relevantes em um determinado conjunto de documento bem como suas associações com base em um algoritmo pré-existente. Por meio da análise de todos os exemplos de documentos, este algoritmo entende as regras e padrões e os armazena em uma base de conhecimentos.

Em etapa posterior, os documentos a serem classificados passam por um categorizador, que se utiliza dos padrões regras e anteriormente identificadas e inseridas na base de conhecimento, classificando e categorizando cada um dos documentos, determinando a qual classe eles pertencem (CORRÊA, 2003).

Do ponto de vista tecnológico, a mineração de textos baseia-se na aplicação de algoritmos computacionais que processam textos e identificam relações entre as informações para externalizar conhecimento útil e implícito, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, uma vez que a informação contida em dados desestruturados não pode ser obtida de forma direta.

2.4.1.1.4 Algoritmo Naive Bayes

O algoritmo selecionado e aplicado na ferramenta de mineração de texto nessa pesquisa foi o Naive Bayes. Trata-se de um classificador probabilístico baseado na aplicação do teorema de Bayes que, seguramente, é um dos mais utilizados em

software de mineração de textos. Segundo Rennie et al (2003), é frequentemente utilizado por ser rápido e fácil de implementar, facilitando de maneira considerável o processo de classificação dos documentos processados. O teorema de Bayes expressa-se pela seguinte fórmula:

$$P(A|B) = \frac{P(A) \text{ x } P(B|A)}{\sum P(A) \text{ x } P(B|A)}$$

Partindo do pressuposto que “A” representa um evento que ocorreu previamente e assumindo que “B” é um evento que depende de “A”, para que se possa calcular a probabilidade do evento “B” ocorrer uma vez que o evento A já ocorreu, o teorema indica que se deve contar o número de casos em que “A” e “B” ocorreram juntos e, posteriormente, dividir pelo número de vezes em que “A” ocorreu sozinho. Para Gomes (2013), o Naive Bayes é considerado um dos mais eficazes, eficientes e efetivos em questões relacionadas a processamento e precisão na classificação de novas amostras.

Segundo SANTOS et al. (2015), a classificação de documentos é também um dos campos da mineração de textos. A classificação, também chamada de categorização, é o processo de atribuição automatizada de categorias pré-determinadas a um documento, dado seu tema. De maneira geral, um classificador ou categorizador é uma função f: do documento $\{d_1,...d_n\}$ que aponta um documento para a categoria ou classe na qual está contido.

A mineração de textos ou Descoberta de Conhecimento em Textos (KDT), visa encontrar padrões e tendências em um conjunto de documentos, realizar classificação de documentos, ou ainda comparar documentos. A Figura 4 apresenta as duas ramificações principais da descoberta de conhecimento:

Figura 4: Tipos de descoberta de conhecimento

Fonte: Morais e Ambrósio (2007), adaptado.

A KDT utiliza-se de abordagens já aplicadas nas áreas de recuperação de informação e descoberta de conhecimento em Banco de Dados. Segundo Barion (2008), mais de 80% das informações de uma organização estão armazenadas em formato textual. Nesse prisma, as técnicas de mineração de textos podem agregar valor aos processos de busca e recuperação da Informação.

Em mineração de textos, os processos de descoberta de conhecimento basicamente estão segmentados em dois ramos: i) Dados não estruturados; e ii) Descoberta de conhecimentos em dados estruturados (WIVES, 2002). A primeira – dados estruturados – refere-se às informações e dados contidos nos Sistemas Gerenciadores de Bancos de Dados (SGBDs) das organizações.

O enfoque desta pesquisa está na descoberta de conhecimento em bases de dados não estruturadas - Descoberta de Conhecimento em Textos (*Knowledge Discovery from Text – KDT*). Esta escolha justifica-se, pois praticamente 100% das informações acerca do desenvolvimento sustentável da Região Amazônica, contidas no acervo documental do Prêmio Professor Samuel Benchimol, estão em diversos formatos eletrônicos de texto. Os aspectos práticos da aplicação de mineração de textos no âmbito desta pesquisa serão tratados na seção de metodologia.

A Descoberta de Conhecimento em Textos ou mineração de textos diferencia-se dos mecanismos de busca. Na busca convencional, os usuários detêm um conhecimento prévio do objeto da sua pesquisa, ou seja, eles já sabem o que desejam encontrar. A mineração de textos auxilia o usuário na descoberta conhecimento e

informações anteriormente desconhecidas, trazendo ao usuário percepções que ele não teria sem a ajuda de um computador. Essas ferramentas criam novos horizontes e auxiliam os usuários permitindo que esses especifiquem novos critérios de busca e recuperação da informação.

Como a maior parte das informações estão armazenadas em forma de textual, as técnicas de mineração de textos mostram-se muito importantes para facilitar os processos de recuperação do conhecimento implícito presente nas bases textuais não só intradocumental, bem como interdocumental.

3 UM RETRATO DA AMAZÔNIA

Os indígenas da região chamavam de *amassunu* o ruído intenso das águas que cortavam aquela floresta. O barulho ressoava fortemente como que anunciando a quem o ouvisse a dispersão e a força desafiadora das águas, bem como, a sua centralidade nesse lugar. O *amassunu* dos índios foi traduzido como *Amazonas* pelos colonizadores espanhóis; o rio grandioso por onde percorreram em busca de novas terras.

Tempos depois, Amazonas passou a nomear o estado localizado no coração da floresta dita então *Amazônica*, um estado serpenteado pela grande floresta e por águas abundantes, cristalinas e escuras, em que a rede hidrográfica e a cobertura vegetal criaram um complexo meio ambiente que lhe confere tempo e compasso próprios em detrimento às dinâmicas naturais de fluxos humanos na região (SOUSA, 2008). Se o *amassumu* indígena descrevia a magnitude e o poder das águas amazônicas, o Amazonas estado herdou, em termos geográficos e em riquezas naturais, a grandiosidade do vocábulo que lhe batizou.

Situado na região Norte do país, maior estado brasileiro, ocupando uma área de aproximadamente 1.577.987km^2, o que corresponde a 18,42% do território nacional (SOUSA, 2008, p. 75). O estado do Amazonas abriga uma população de mais de três milhões de habitantes, porém, desigualmente distribuída neste espaço. Os demais estados que compõem a região amazônica têm características semelhantes.

Detentora de inúmeras riquezas naturais, a região sustentou-se por muito tempo sob uma base produtiva exclusivamente extrativista com ciclos de produtos extrativos, determinados pela demanda do mercado externo. Benchimol (2001), utilizando conceitos da geoastronomia e da astrofísica, afirmou que a região amazônica passa por um *Zênite* ecológico (apogeu) ao mesmo tempo em que vive um Nadir (perigeu) econômico-social. A constatação poética de Samuel Benchimol retrata o ciclo mais recente de desenvolvimento da região, com uma pseudo valorização ecológica, com poucas ações voltadas ao desenvolvimento econômico-social.

Objeto de interesse internacional e influenciada pelo contexto geopolítico, a região já passou por inúmeros ciclos de desenvolvimento, com altos e baixos. Dois deles merecem destaque: as drogas do sertão e o ciclo da borracha.

O primeiro se caracterizou pela exploração e comércio de especiarias e durou até meados do século XIX. Já a exploração da borracha dos seringais nativos atingiu seu apogeu no início do século XX, consolidando uma base econômica cuja riqueza esteve circunscrita a alguns poucos, como os “senhores da borracha”. Seu poderio foi de tal magnitude que edificaram belíssimos edifícios em estilo europeu, na tentativa de erguer uma capital à altura do império da borracha. Assim é que em Manaus foi nominada como a "Paris dos Trópicos" (BECKER, 2005).

Segundo Nazareth *et al.* (2011), é comum na região a alternância de períodos expansivos e recessivos ao longo da história econômica. O marco inicial de um novo período de desenvolvimento econômico para a região foi proporcionado pela Constituição de 1946, na qual foi determinado que 3% da renda tributária da União seriam destinadas à valorização da Amazônia por um período de vinte anos.

Para Moura e Moreira (2010, p. 216-221), foi a partir da expansão de formas de acumulação e de investimentos públicos que se procedeu a ocupação das fronteiras amazônicas, em um movimento de expansão que privilegiou determinados espaços, atraindo para lá pessoas de várias regiões do país, ao mesmo tempo em que contribuiu ou determinou a estagnação das atividades econômicas do interior.

A partir de 1950 a região amazônica começou gradativamente a retomar o crescimento por meio de incentivos do Governo Federal. Esse processo culminou com a criação da Zona Franca de Manaus (ZFM), em 1967, em um movimento de fomento à industrialização, com o objetivo de irradiar o desenvolvimento a vastas porções interiorizadas da Amazônia Ocidental. Desde então, a ZFM, com seu livre comércio e incentivos fiscais, tem sido a principal propulsora da economia amazonense e fator de atração de migrantes, seja da área rural estagnada economicamente ou mesmo de áreas urbanas e outros países. Mais recentemente, os projetos minerais, o garimpo e as reservas minerais, além da extração de madeira, são as atividades que mobilizam investimentos na região.

Lasmar *et al.,* (2010), avaliando positivamente a intervenção estatal, usam dados do Instituto Brasileiro de Geografia e Estatística – IBGE para salientar que a atividade industrial permitiu à economia do estado do Amazonas alcançar, em 2005, o 15º posto entre os PIB estaduais, ao tempo em que sua capital, que concentra a indústria incentivada no estado, atingiu significativa 7ª posição dentre todos os municípios do Brasil no mesmo ano.

Vinculado ao crescimento econômico, ocorreu um intenso incremento populacional. De acordo com os censos demográficos, só a cidade de Manaus passou de 171.343 habitantes em 1960 para 1.802.525 habitantes em 2010, grande parte deste crescimento como fruto do intenso processo migratório (LASMAR *et al*., 2010).

A situação apontada anteriormente repetiu-se em outros centros urbanos amazonenses. Entretanto, como a política de investimentos estatais concentrou-se principalmente em Manaus, as demais cidades não conseguiram criar condições para atender e receber a considerável população migrante que para elas dirigiram-se, alterando assim toda a estrutura socioeconômica e ambiental do estado.

Outro dado importante em termos migratórios, é que o estado tem recebido um fluxo intenso de imigrantes de outros países da América Latina. Esta constatação não chega a surpreender, pois além dos limites com diversos estados brasileiros – Acre, Rondônia, Roraima, Mato Grosso e Pará – o Amazonas possui uma extensa fronteira de 3.600 km com a Venezuela, Colômbia e Peru. Nesse sentido, questiona-se: Como monitorar, fiscalizar e garantir a soberania e segurança de um espaço tão vasto e cobiçado internacionalmente?

3.1 O CENÁRIO DO ACOLHIMENTO

Conforme Mattos (2012), do ponto de vista geopolítico, a Amazônia trata-se de um dos poucos subsistemas mundiais ainda quase inexplorados pelo homem e que, se bem utilizado pelo Brasil, é possível tornar-se em indutor do desenvolvimento sustentável do País, com suporte no aproveitamento das riquezas da região e na preservação ambiental.

Abandonar grandes mitos – esta é a proposição racional para a compreensão de uma região marcada por grandes transformações ao longo da história. Para Ianni (1994) a Amazônia é um Novo Mundo, do outro mundo das maravilhas. A realidade parece ser outra, quando verifica-se que os fatos contradizem a imaginação. Ocorre que a Amazônia tornou-se o emblema de algum lugar, uma identidade.

Conforme resultados apresentados no Seminário de Segurança da Amazônia, realizado no período de 11 a 15 de agosto de 2010, na cidade de Manaus (AM), organizado pela Secretaria de Assuntos Estratégicos em parceria com o Comando do Exército, por meio do Estado-Maior do Exército e do Comando Militar da Amazônia, a Amazônia brasileira é atualmente prioridade nacional, de acordo com a Estratégia Nacional de Defesa;

> A Amazônia brasileira abrange uma área de 5,2 milhões de Km2, com densidade populacional de 3,2 hab/km2, 1/3 das florestas tropicais da Terra, maior diversidade biológica do planeta e maior bacia de água doce do mundo. Essa região é detentora de exuberante fauna e flora. Suas riquezas estão praticamente intocados e minuciosos levantamentos indicam que abriga uma das mais extraordinárias províncias minerais do planeta. Tudo isso deixa evidenciado que a Amazônia é já há muito tempo, área estratégica de alto interesse para os brasileiros. Impõe-se a urgente necessidade de integrá-la ao ambiente nacional e articulá-la com os nossos vizinhos, também depositários desse patrimônio. Este é o motivo principal da prioridade nacional hoje emprestada à nossa Amazônia. Para ela orienta-se o destino manifesto do Brasil (BRASIL, 2010).

No entanto, ainda no referido Seminário, segundo o mapa Fronteiras do Brasil apresentado pela Coordenação de Operações Especiais de Fronteira (COESF), para operacionalizar a segurança nacional nas fronteiras do Brasil com outros países, percebe-se que para defender a essa extensa faixa de fronteira faz-se necessário que governo brasileiro disponibilize alguns bilhões de dólares para custear a criação de infraestruturas específicas, para a fiscalização e implementação de ações que venham de fato ter o sentido de “proteção”. Entretanto a defesa da Amazônia brasileira, em especifico, torna-se mais difícil em virtude de sua posição geográfica e ambiental, pelo vazio demográfico e por sua precária infraestrutura, levando o Brasil a articular com países vizinhos por meio de cooperações multilaterais.

A população e seu estilo de vida, o potencial medicinal e extrativista disponível, a imensa malha hidrográfica que atravessa por enormes e densas florestas primárias. A extraordinária biodiversidade, muito ainda por descobrir, revelar e patentear. Tudo

isso e tudo o mais de desconhecido que compõem o emblemático espaço territorial amazônico continuam sendo motivo de intensa reflexão e inquietação. A Amazônia não é o pulmão do mundo, mas é cobiçada insistentemente por países que visam a hegemonia econômica e desenvolvimentista.

Segundo Bertha Becker (2005), geógrafa e pesquisadora, “a Amazônia já é uma floresta urbanizada”, o que para os especialistas é motivo de polêmica. Floresta urbanizada porque registrou a maior taxa de crescimento urbano do país nas últimas três décadas. O que reforça a análise de Bertha é o censo de 2010 do IBGE, sinalizando que quase 85% da população na região Norte do país vive em centros urbanos com uma das piores distribuições de renda do País. Para a pesquisadora,

> a falta de infraestrutura e de serviços condena o reconhecimento desses núcleos como pequenas cidades, apontados ainda como aglomerados inchados meio rurais e meio urbanos (BECKER, 2005).

As demandas sobre a Amazônia brasileira estabelecem urgências e prioridades, tanto no campo da segurança, da política, do econômico, social e ambiental. Sobre esses motes recaem influências e pressões de toda ordem, “algumas até questionando a soberania do País sobre a região.

Assim, segundo a SAE (BRASIL, 2010), a Política de Defesa Nacional (PDN) define a segurança da Amazônia como a condição que permite ao País a preservação da soberania e da integridade territorial, a efetivação dos seus interesses nacionais, livre de pressões e ameaças de qualquer natureza, e a garantia aos cidadãos do exercício dos direitos e deveres constitucionais. Destaca que a segurança de um país é afetada pelo grau de estabilidade da região na qual ele está inserido e que é desejável que ocorram: o consenso; a harmonia política; e a convergência de ações entre os países vizinhos, visando lograr a redução da criminalidade transnacional, na busca de melhores condições para o desenvolvimento econômico e social que tornarão a região mais coesa e mais forte.

3.2 A TRÍPLICE FRONTEIRA

A referência e os pontos de convergência na tríplice-fronteira são as cidades de Santa Rosa, no Peru; Tabatinga, no Brasil e Letícia, na Colômbia. Todas são

cidades de pequeno porte perdidas nos confins dos três países, tendo como um único cenário que as envolvem o barrento rio Solimões. No entanto, como bem aponta (OLIVEIRA, 2006, p. 186), a cidade de Tabatinga, distante 1.105 km de Manaus em linha reta e 1.607 por via fluvial, é emblemática como ponto de maior movimentação migratória na região amazônica, concentrando uma porcentagem significativa de migrantes colombianos e peruanos e apresentando-se também como porta de entrada no território brasileiro.

É relativamente comum na cidade o fluxo desenfreado e contínuo de pessoas e o elevado contingente de indivíduos vivendo em situação irregular ou na clandestinidade, comumente explorados em sua força de trabalho. Outra situação por ali vivenciada é a pressão crescente sobre os já precários serviços de saúde, educação e segurança. Em essência, a cidade é uma arena multifacetada de tensões implícitas e explícitas, em que o controle sobre o ir e vir de pessoas e objetos é constante, ainda “[...] considerando a vastidão da selva amazônica, é humanamente impossível manter um controle 100% eficaz nessas condições de traslado permanente” (OLIVEIRA, 2006, p. 186).

Como lugar de confluências, Manaus e Tabatinga mesclam diferentes culturas, mercados e estratos sociais. É relativamente comum o fluxo desenfreado e contínuo de pessoas e o elevado contingente de indivíduos vivendo em situação irregular ou na clandestinidade, comumente explorados em sua força de trabalho. Situação também vivenciada é a pressão e demanda crescente sobre os já precários serviços de saúde, educação e segurança nestes municípios.

A localização geográfica fronteiriça e as implicações advindas desta singularidade, qual seja, a necessidade de proteção física do território brasileiro em um complexo contexto geográfico e geopolítico. Como esperado para o contexto descrito, as políticas públicas estão longe de conseguir responder às demandas. Entre o que se preconiza como direito do ser humano e a realidade vivenciada em Tabatinga existe uma larga e tortuosa fronteira, que só não é maior que a esperança dos migrantes haitianos que ali chegam diariamente.

A realidade é que não há verbas específicas dos municípios brasileiros para lidar com migrantes e com esse volume de pessoas. Os migrantes peruanos e

colombianos têm a cara dos índios e caboclos brasileiros e frequentemente são do mesmo grupo étnico, confundem-se à movimentação de etnias, índios e caboclos na cidade, os haitianos, não.

Uma das características dessa parte da Amazônia é o intenso tráfego fluvial entre os rios brasileiros, colombianos e peruanos, em que as cidades são embutidas nessa geografia entre os rios e a grande floresta, resultando num admirável mosaico complexo e delicado, envolvendo a todos num só espaço de tempo. A 'fronteira' é de uma violência silenciosa: o narcotráfico movimenta pequenos povoados, penetra em terras indígenas e alicia indivíduos desempregados que circulam entre os pequenos municípios do entorno de Tabatinga.

3.3 O DESMATAMENTO E A EMISSÃO DE CARBONO

As florestas tropicais podem ser consideradas enormes depósitos de carbono – 200 bilhões de toneladas (IPCC, 2000) –, e precisam continuar intactas para conservar o aquecimento global sob controle (IPCC, 2007). Entretanto, as emissões de gases de efeito estufa (GEE) provenientes do desmatamento seguem elevadas, piorando o aquecimento global. Durante a década de 1990, a derrubada das florestas gerou uma emissão de carbono (na forma de CO2, um potente GEE) para atmosfera da ordem de 800 milhões a 2.2 bilhões de toneladas de carbono por ano (tC/ano), o equivalente a 10-35% da emissão global (HOUGHTON, 2005; ACHARD et al., 2002; DeFRIES et al. 2002, IPCC, 2007).

Somente na Amazônia brasileira, o desmatamento liberou durante a última década 200 milhões TC/ano ou 3% do total global (HOUGHTON, 2005). Os prejuízos para a biodiversidade (SOARES FILHO et al., 2006) e para o sistema hidrológico mantido pela floresta (SALATI e VOSE, 1984) foram incalculáveis. Entre os anos 2000 e 2008, a emissão média proveniente de desmatamento foi de 220 milhões tC/ano. Isto representa aproximadamente 55% das emissões totais do Brasil, um valor superior se comparado àquela por queima de combustíveis fósseis (100 Milhões tC/ano; ano de referência, 2008; EIA, 2009).

O volume das emissões brasileiras pode, contudo, ter sido ainda maior (o dobro) se incluirmos as emissões resultantes dos incêndios florestais amazônicos, um montante que, por sinal, não foi incluído no primeiro Inventário de Emissões Brasileiras, o relatório que cada país deve emitir junto a Convenção-Quadro das Nações Unidas sobre Mudança do Clima (UNFCCC, sigla em inglês). Por consequência, a emissão de carbono por desmatamento e incêndios florestais na Amazônia brasileira poderá anular, nos próximos anos, mais da metade dos esforços de redução de emissões realizados pelos países desenvolvidos através do Protocolo de Quioto (SANTILLI et al. 2005, MOUTINHO & SCHWARTZMAN 2005).

De acordo com Moutinho (2016), a taxa histórica de desmatamento amazônico foi da ordem de 20 mil km² durante as décadas de 1980 e 1990 (18.165 km² durante 1990) com um pico em 1995 de 29.059 km². A constatação desses dados provocou uma grande preocupação internacional em que o Brasil ocupou um papel de protagonismo negativo, que começou a mudar após a Conferência das Nações Unidas sobre o Meio Ambiente e o Desenvolvimento, Rio +10 em 1992, também conhecida como Eco-92.

Samuel Benchimol, um visionário, propôs a criação de um imposto ambiental internacional que seria pago pelos países desenvolvidos, que já desmataram suas florestas, para que os países amazônicos pudessem desenvolver-se sem a necessidade de devastar a floresta “[...] não se reconhece que a contrapartida e o ônus devem recair sobre aqueles países beneficiados, que devem assumir as suas responsabilidades e obrigações de contribuintes de um necessário imposto internacional ambiental, que deve ser criado e exigido pelos países amazônicos pelo suprimento de tais benefícios e serviços” (BENCHIMOL, 2010).

O desmatamento tropical é decorrente da interação de inúmeros fatores que variam ao longo de dois eixos: um geográfico e outro temporal. É, portanto, um fenômeno complexo. Contudo, as causas do desmatamento parecem ser aparentemente as mesmas nas diferentes regiões tropicais do planeta. Resumidamente, as causas podem ser diretas e indiretas. As diretas estão ligadas à conversão de áreas florestais para agricultura ou criação de gado, mineração, exploração madeireira e incêndios florestais. Já as indiretas são os subsídios para o agronegócio, política inadequada de investimentos em infraestrutura, problemas

fundiários, ausência de governança e fiscalização adequada por parte do governo, demanda por produtos florestais (madeira e outros produtos florestais) e mercado, preço, favorável a produtos cultivados em áreas antes ocupados por florestas (grãos e carne, por exemplo).

Para medir a situação da Amazônia hoje, sob as perspectivas de transição para um modelo sustentável de desenvolvimento, é necessário reconhecer uma realidade básica que nem sempre é percebida na visão comum do problema, inclusive em escala internacional. A análise do acervo documental do Prêmio Professor Samuel Benchimol poderá ratificar as questões-chave ao desenvolvimento da Amazônia sob uma ótica contemporânea.

3.4 PRÊMIO PROFESSOR SAMUEL BENCHIMOL

A justificativa para a realização deste trabalho apoia-se no legado deixado pelo Professor Samuel Isaac Benchimol (1924-2002), o qual era internacionalmente reconhecido como uma das maiores autoridades no tema Amazônia. Com mais de uma centena de publicações científicas e livros, deixou um legado intelectual extraordinário. Por sua contribuição acadêmica e social, foi homenageado pelo MDIC com a criação do “Prêmio Professor Samuel Benchimol”.

3.4.1 A Biografia

Samuel Isaac Benchimol, nascido no dia 13 de julho de 1923, em Manaus, foi um escritor (com cerca de 110 trabalhos publicados, nomeado à Academia Amazonense de Letras), professor (Emérito da Universidade do Amazonas, onde lecionou por mais de 50 anos), pesquisador (catedrático da cadeira de Introdução à Amazônia na Universidade do Amazonas), líder comunitário (presidente do Comitê Israelita do Amazonas) e empresário (fundador do grupo Bemol/Fogás).

Em 1942 Samuel Benchimol fez vestibular e ingressou na Faculdade de Direito do Amazonas, na qual formou-se Bacharel. Trabalhou inicialmente como despachante de bagagens e passageiros na Panair do Brasil, onde obteve o contato que o estimulou a escrever o trabalho "O Cearense na Amazônia". Baseado no sucesso deste trabalho recebeu bolsa para fazer Mestrado em Sociologia na Miami University, nos EUA, em 1946/1947, de onde retornou para seguir sua carreira na Amazônia.

No conjunto de sua vasta produção intelectual, merecem destaque: Amazônia: Um Pouco-Antes e Além-Depois, O Pacto Amazônico e a Amazônia Brasileira, Amazônia: Formação Social e Cultural, Eretz Amazônia: Os Judeus na Amazônia e Zênite Ecológico e Nadir Econômico-Social. Também destacou-se pela compilação e análise de estatísticas socioeconômicas locais, especialmente no que diz respeito à arrecadação de impostos e comercio exterior na região amazônica. Em toda a sua obra, sempre defendeu a necessidade de que o desenvolvimento sustentável da Amazônia deve respeitar quatro parâmetros e paradigmas fundamentais: ser economicamente viável, ecologicamente adequado, politicamente equilibrado e socialmente justo.

Como professor, Samuel lecionou por quase cinquenta anos na Universidade Federal do Amazonas em cursos de economia e direito. Seu interesse no estudo dos mais diversos aspectos da Amazônia culminaram na sua escolha para criar uma matéria de Introdução à Amazônia. Dada a sua dedicação ao estudo e à educação, Samuel foi nomeado à Academia Amazonense de Letras. O prédio principal da Universidade Estadual do Amazonas e a Escola Estadual Samuel Benchimol, no bairro Nova Cidade, são dedicados ao seu nome. Após seu falecimento, foi criado o Prêmio Amazônia Professor Samuel Benchimol, patrocinado pelo Ministério do Desenvolvimento, Indústria e Comércio e pela FIEAM.

Como empreendedor, Samuel Benchimol foi, juntamente com seus irmãos Israel e Saul, fundador do grupo Bemol/Fogás, em 1942. As empresas do grupo estão em diversas atividades, incluindo distribuição de gás de cozinha, lojas de departamento, centros comerciais virtuais e exportação de produtos naturais da Amazônia, como bálsamo de copaíba e óleo de pau-rosa. A exportação de produtos naturais foi iniciada por um dos fundadores, Samuel Benchimol e hoje é liderada por Ilko Minev, diretor das empresas Bemol e Fogás e vice-presidente da Bemol.

3.4.2 A comenda

Segundo Oscar Soto Lorenzo Fernandez (2012), o Prêmio que leva o nome de Samuel Benchimol é mais do que o reconhecimento do notável pesquisador e empresário, é um indicador do verdadeiro caminho da Amazônia e do Brasil. É um instrumento que revela a transformação do mundo e a superação dos limites da

realidade, em termos de pesquisas teóricas e aplicadas de desenvolvimento tecnológico e de geração de riqueza.

A morte do Professor Samuel Isaac Benchimol, no ano de 2002, foi uma grande perda para a sociedade brasileira, em especial ao povo da Região Amazônica. Considerado uma das maiores autoridades internacionais no tema “Amazônia”, Benchimol deixou uma lacuna que dificilmente poderá ser preenchida.

No mesmo ano em que o Professor Samuel Benchimol faleceu, o Presidente Luiz Inácio Lula da Silva foi eleito para seu primeiro mandato, com início em 1º de Janeiro do ano seguinte, 2003. Nesta época, o Brasil vivia uma onda de otimismo e o povo tinha esperança que com a chegada de um operário ao poder, inúmeros desafios e problemas do país pudessem ser enfrentados.

A Região Amazônica naquele momento sofria com graves problemas ambientais, obstáculos logísticos, desigualdades sociais, baixo nível de renda, infraestrutura de serviços públicos precária, baixa densidade demográfica e grande extensão territorial, êxodo rural, exclusão social e um grande número de estrangeiros (empresários, pesquisadores, religiosos e cientistas) atuando de forma ilegal na região.

Segundo Alvares e Ferreira (2014), os problemas da região também estavam graves no campo da Ciência e Tecnologia, haja vista que no ano de 2002 a região representava apenas “2% dos grupos de pesquisa cadastrados no Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), 1,7% de pesquisadores, 1,5% de doutores e 1,9% de cursos de pós-graduação credenciados”. Havia um grande espaço para estímulo à pesquisa e ao desenvolvimento da temática da região amazônica.

Nesse contexto, o Ministério do Desenvolvimento, Indústria e Comércio Exterior (MDIC), entendendo o momento do país e a necessidade latente de estímulo à produção científica, tecnológica e empresarial da região, com apoio da Confederação Nacional da Indústria (CNI), instituiu no ano de 2003 o prêmio que levou o nome do Professor Samuel Benchimol.

O principal objetivo do prêmio era “conclamar a inteligência brasileira e estrangeira a pensar a Amazônia na abrangência das dimensões ambiental,

econômica, tecnológica e social, estimulando a criação e a implementação de projetos para o desenvolvimento sustentável da região amazônica (ALVARES, FERREIRA, 2014)". Visava-se o aumento da produtividade das empresas; maneiras mais rentáveis de exploração dos recursos naturais com menores impactos ambientais, além de elevar os índices de qualidade de vida da população local.

Segundo o seu regulamento (2015), o Prêmio Professor Samuel Benchimol foi concebido, organizado e dividido com premiações em quatro categorias distintas: ambiental, econômico-tecnológica, social e personalidade amazônica, sempre selecionando e agraciando aquelas propostas mais inovadoras, bem como reconhecendo a pessoa que se destacou por alguma ação em prol do desenvolvimento sustentável da região amazônica:

i. Categoria Projetos de Natureza Ambiental, cujo objetivo é mostrar como o meio ambiente pode ser utilizado de forma racional e responsável, permitindo a regeneração contínua dos recursos naturais. As propostas podem perseguir objetivos mais ambiciosos, tais como recuperação de áreas degradadas, preservação de espécies, desenvolvimento de tecnologias mais limpas, reintrodução de espécies nativas, sensoriamento de recursos naturais, dentre outras. A avaliação das candidaturas observa a visão geral da problemática e das soluções propostas pelo projeto, indicando se há efetivo desenvolvimento sustentável e os benefícios a serem gerados;

ii. Categoria Projetos de Natureza Econômico-Tecnológica, cujo objetivo é incentivar a realização de projetos que beneficiem a economia regional e as estruturas produtivas da Amazônia durante ou após a sua execução. Deverão, por exemplo, propiciar o aumento perceptível e verificável no PIB regional, na balança comercial, na arrecadação de impostos, na geração de emprego e renda, na qualidade e produtividade de produtos da região ou ainda na ampliação da oferta e da diversidade de produtos e serviços gerados e consumidos na Amazônia;

iii. Categoria Projetos de Natureza Social, cujo objetivo é selecionar projetos que tenham impacto positivo no tecido social e melhorarem as

condições e a qualidade de vida da população amazônica. Deverão, por exemplo, ampliar e garantir o acesso de todos à saúde, à educação, à habitação, ao entretenimento e à cultura; melhorar a distribuição de renda, diminuindo as diferenças sociais; ou ainda, combater a discriminação, a miséria e a fome.

iv. Categoria Personalidade Amazônica: tem como objetivo agraciar empresários, executivos e gestores de políticas públicas, acadêmicos, bem como qualquer personalidade do meio amazônico, que se destacaram em ações em prol do desenvolvimento sustentável da região amazônica.

3.5 O PRÊMIO BANCO DA AMAZÔNIA DE EMPREENDEDORISMO CONSCIENTE

Em 2009, os Prêmios Professor Samuel Benchimol e Banco da Amazônia de Empreendedorismo Consciente foram unificados com regulamento único. Foi criada uma parceria entre o MDIC e o Banco da Amazônia S/A e contam com o apoio da Confederação Nacional da Indústria (CNI), do Serviço Brasileiro de Apoio à Micro e Pequena Empresa (SEBRAE), entre outros.

Segundo o site oficial do Prêmio Banco da Amazônia de Empreendedorismo Consciente, a comenda contempla três naturezas de premiação: a primeira busca a identificação de projetos com abordagem integrada, mas com o potencial de transformação da realidade socioeconômica; a segunda, com as iniciativas de suporte ao desenvolvimento regional, tendo como compromisso estimular o desenvolvimento de projetos inovadores na Amazônia Legal e a terceira para o reconhecimento de empresas que contribuem para o desenvolvimento sustentável da região.

Este trabalho não contemplou as inscrições do Prêmio Banco da Amazônia de Empreendedorismo Consciente, bem como desconsiderou as inscrições na Categoria “Personalidade Amazônica”.

Samuel Benchimol deixou as seguintes palavras: Gostaria de deixar como lembrança uma singela e profunda mensagem: o mundo amazônico não poderá ficar

isolado ou alheio ao desenvolvimento brasileiro e internacional, porém ele terá que se autossustentar em quatro parâmetros e paradigmas fundamentais:

“O mundo amazônico deve ser economicamente viável,
ecologicamente adequado,
politicamente equilibrado,
e socialmente justo”.
Samuel Benchimol (1924-2002).

4 METODOLOGIA

Esta pesquisa seguiu duas linhas metodológicas para o completo atingimento dos objetivos e questão declarados. A primeira é predominantemente quantitativa, marcada pela Bibliometria e mineração de textos e a segunda com enfoque qualitativo, a Análise de Conteúdo sob a ótica de Bardin (1977).

O método para AC proposto por Bardin (1977) é caracterizado por um conjunto de instrumentos metodológicos que se aplicam a discursos altamente diversificados. O desenvolvimento desses instrumentos de análise de comunicações é acompanhar, passo a passo, o crescimento quantitativo e as diversas formas qualitativas das pesquisas empíricas, apoiadas em uma das técnicas conhecida como Análise de Conteúdos (VILELA JUNIOR, 2015).

Pogré (2006) argumenta que é importante estabelecer uma matriz de tipificação, porque essa é uma ferramenta que auxilia a pesquisa, tornando-a mais rápida e eficiente. Todavia, não existem fórmulas seguras para elaborar uma boa matriz, mas sim instruções e orientações básicas que podem servir de auxílio para a construção. É preciso que o pesquisador evite a criação de categorias que sejam sobrepostas, redundantes ou muito longas.

Na evolução do conceito de AC evidencia-se a necessidade de compreender o contexto para se compreender o texto. Embora os dados estejam expressos no texto, o contexto precisa ser reconstruído pelo pesquisador. Não existem limites lógicos para delimitar o contexto da análise. Isto vai depender do pesquisador, da disciplina, e dos objetivos propostos para a investigação (MORAES, 1999, p. 9).

A metodologia, segundo Payne e Payne (2004), indica conceitos e suposições filosóficas que justificam o uso de métodos específicos. Segundo Kothari (2004), metodologia é a ciência que estuda como a pesquisa é realizada cientificamente. A seguir são apresentados a metodologia e os procedimentos metodológicos da pesquisa.

O Prêmio Professor Samuel Benchimol já existe há mais de uma década e, ao longo deste período, foram submetidas centenas de pesquisas, projetos e estudos, conforme demonstrado na seção 4 desta pesquisa – Resultados e Análises.

O quadro síntese a seguir apresenta uma descrição resumida dos procedimentos metodológicos aplicados para a realização desta pesquisa:

Quadro 2 – Quadro síntese da metodologia

ELEMENTO	DESCRIÇÃO
Modelo conceitual-teórico	**Conceitos:** Mineração de dados; Descoberta de Conhecimento; Análise de Conteúdo; Bibliometria; Recuperação da Informação; Propostas de desenvolvimento sustentável da Amazônia.
	Objetivo: Identificar, classificar e analisar as propostas submetidas ao Prêmio Professor Benchimol, durante os anos de 2004 a 2015, por meio da mineração de textos, para definição das questões-chave de desenvolvimento da Região Amazônica sob a ótica desse prêmio.
	Foco: Propostas submetidas ao prêmio.
	Corpus documental: Conjunto de 1856 trabalhos, pesquisas, teses e dissertações do acervo do prêmio.
Objeto de estudo	Perfil das propostas de desenvolvimento da Amazônia do Prêmio Professor Samuel Benchimol.
Dupla natureza da pesquisa	**Aplicada**. Foi aplicado o algoritmo bayesiano (Naive Bayes) com a ferramenta de mineração de textos Tropes, além de Análise de Conteúdo sob a ótica de Laurence Bardin. **Exploratória.** Nunca houve um estudo que avaliasse o acervo do Prêmio Professor Samuel Benchimol, gerando dados predominantemente quantitativos, por meio desta pesquisa, disponibilizados à sociedade com vistas a subsidiar a discussão do desenvolvimento da Região Amazônica.
Abordagem metodológica	**Mista.** Predominantemente quantitativa (Bibliometria), porém, também houve Análise de Conteúdo AC, o que caracterizou um Método Misto.
Horizonte temporal	**Longitudinal**. Estudo da variação das propostas de desenvolvimento sustentável da Amazônia ao longo de doze anos, 2004-2015.
Métodos	**Análise de Conteúdo**. (BARDIN, 1977). **Bibliometria.** (Zipf, Lotka e Bradford).

ELEMENTO	DESCRIÇÃO
Técnicas	**Text Minning:** Aplicação de software de mineração de texto para realização de Bibliometria. **Análise de Conteúdo digital:** Aplicação de software de "Text-Analisys" para realização de AC.
	Análise estatística: Probabilística e de correlação.
	Ferramentas computacionais de apoio: Software de Data Analisys, Data Minning e Text Minning da empresa Semantic Knowledge, em especial o software TROPES. **Revisão de literatura e pesquisas bibliográficas:** Para o desenvolvimento deste trabalho foram realizadas pesquisas bibliográficas e efetuadas buscas a partir das diversas bibliografias empregadas às discussões a respeito do tema. Conforme orientam Cervo e Bervian (1983).

Fonte: Elaborado pelo autor, 2015.

A aplicação prática deu-se pelo processamento eletrônico dos textos – em software de mineração de dados – de forma exploratória, uma vez que pesquisa desta natureza ainda não havia sido realizada com o acervo em questão.

Para responder às questões e ao problema da pesquisa, foram utilizados concomitantemente os métodos de Análise de Conteúdo de Laurence Bardin e Bibliometria com a mineração de textos. Seguindo a metodologia de Bardin, foram desenvolvidas as três etapas cronológicas a seguir:

i) **Pré-Análise**: Análise flutuante; Escolha dos documentos; Preparação do material; Referenciação de índices e Elaboração de Indicadores;

ii) **Exploração do Material**: Esta foi a etapa mais longa e cansativa. Foi a efetivação das decisões tomadas na pré-análise. O momento em que os dados brutos foram transformados de forma organizada e agregados em unidades, as quais permitiram uma descrição das características pertinentes do conteúdo;

iii) **Tratamento dos Resultados, Inferência e Interpretação**: Esta última etapa consistiu no tratamento estatístico simples dos resultados, permitindo a elaboração de tabelas que condensaram e destacaram as informações fornecidas para análise.

Para aplicação dos software de Text Minning foram adotados os seguintes processos:

i) **Coleta das informações:** Consistiu em buscar todo o corpus de pesquisas e trabalhos apresentados desde a primeira edição do Prêmio Professor Samuel Benchimol (2004), independentemente dos formatos dos arquivos (Papel, PDF, DOC, JPEG e outros);

ii) **Padronização dos formatos de arquivos:** Consistiu na conversão de todos os formatos de arquivos para um mesmo padrão (TXT);

iii) **Criação do repositório:** Consistiu na montagem de um repositório indexado de todos os arquivos padronizados (banco de dados textual), para posterior análise;

iv) **Limpeza dos dados:** Consistiu na exclusão das *Stopwords* (palavras com baixo conteúdo semântico, exemplo: “da”, “que”, “em”, “se” etc.). Essas palavras foram desprezadas pelo motor de mineração de dados para potencializar a assertividade das análises. A lista completa de *Stopwords* utilizadas está no Anexo I desta pesquisa;

v) **Aplicação dos modelos de análise e mineração de textos:** Foram utilizados os software da Semantic Knowledge para mineração de texto com o algoritmo Naive Bayes para identificação de padrões, classificação automática de textos, agrupamento por semelhança e prospecção de informações em texto de línguas naturais.

Durante os pré-testes da mineração de textos, realizados para a fase de qualificação dessa pesquisa, foram processadas e mineradas amostras dos arquivos textuais para verificar a aderência da metodologia e das ferramentas computacionais aos objetivos geral e específicos. Percebeu-se que o software recuperava um volume muito grande de termos, categorias, correlações lógicas, cenários semânticos, relações semânticas e episódios prováveis, que tornava inviável e humanamente impossível realizar análise qualitativa, inferência e interpretação de todos os caminhos possíveis.

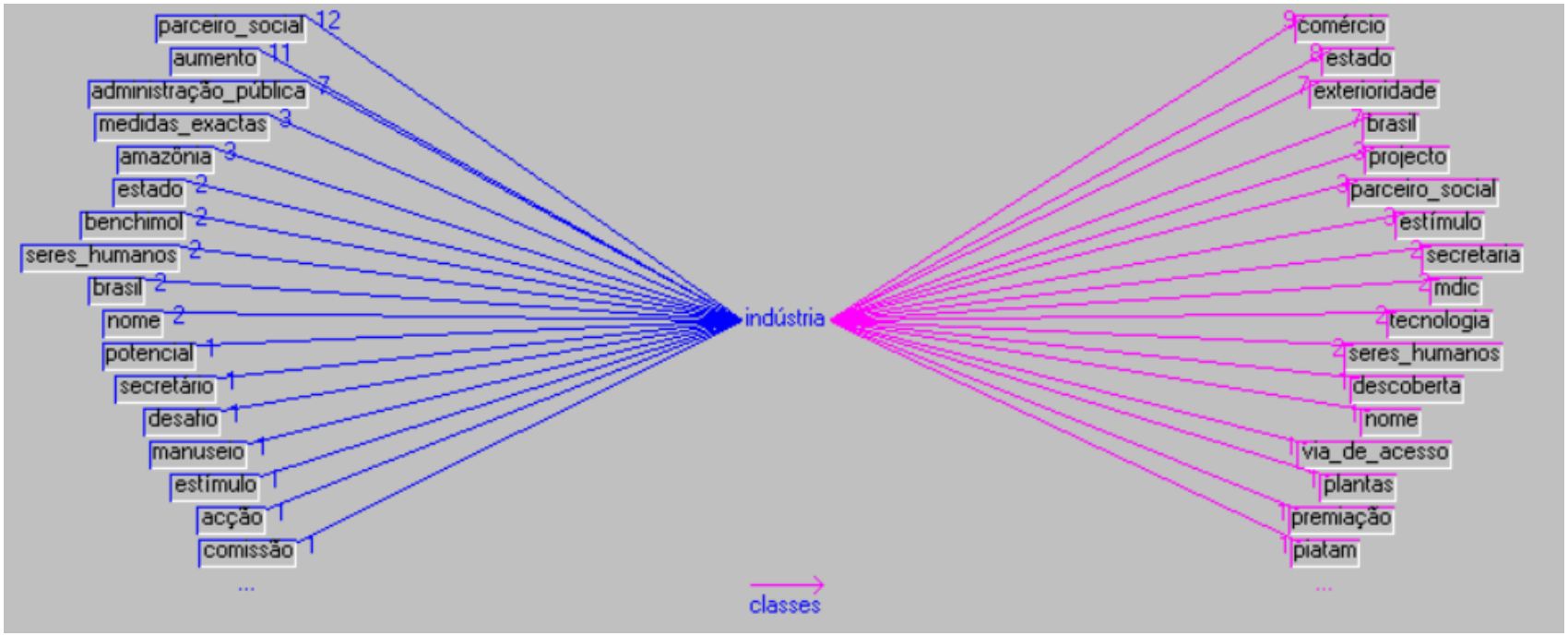
Ao processar os arquivos, centenas de relações semânticas foram identificadas automaticamente pelos software, em diversas naturezas: Ambiental, cultural,

econômica, social, cientifica, étnica, demográfica, política, local, regional, tecnológica, empresarial etc. Todavia, apesar de a base de dados permitir a análise multidimensional de cada uma dessas naturezas, foram selecionadas para análise qualitativa apenas três dimensões: Social, Ambiental e Econômico-tecnológica, o que deixa um amplo campo de trabalho para pesquisas futuras.

A plataforma Zoom desenvolvida pela empresa Acetic Semantick Knowledge em especial o software Tropes, disponibiliza várias interfaces gráficas para consolidar, compilar e apresentar os dados minerados. Os principais são o gráfico em Estrela, o gráfico de relação entre Atores, o gráfico de Esferas e o gráfico de Episódios:

i. **Gráfico em Estrela**: O gráfico em estrela mostra as relações entre referências, ou entre uma categoria de palavras e referências. Os números que aparecem no gráfico indicam a quantidade de relações (frequência de ocorrência) que existe entre as referências. Este tipo de gráfico permite analisar o contexto de uma referência ou categoria. As referências são orientadas: As referências apresentadas à esquerda da classe central são as que vêm antes, as que são apresentadas à direita são as que vêm depois.

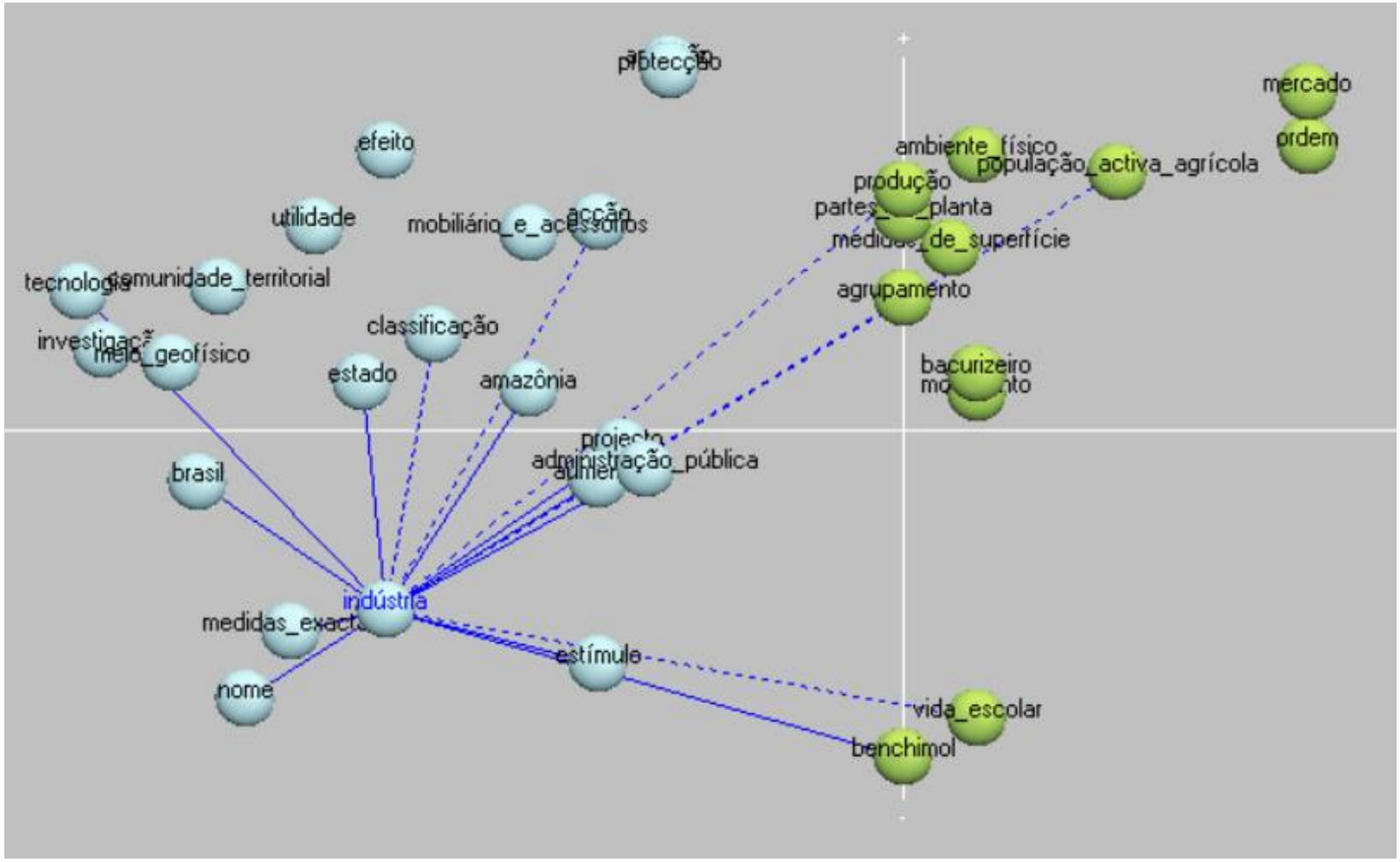
Figura 5 - Exemplo de Gráfico em Estrela



Fonte: Tela do software Tropes (2016).

ii. **Gráfico de relação entre Atores**: Este gráfico representa a concentração de relações entre atores. Ele permite fazer uma comparação visual do peso das relações entre as principais referências. O eixo X (horizontal) indica a taxa sujeito/objeto (da esquerda à direita). O eixo Y (vertical) indica a concentração de relações para cada referência mostrada. Os traços indicam as relações entre a variável selecionada e as outras referências ilustradas. Um traço em picotado indica uma relação pouco frequente. Apenas as referências que apresentam um grande número de relações são representadas no gráfico.

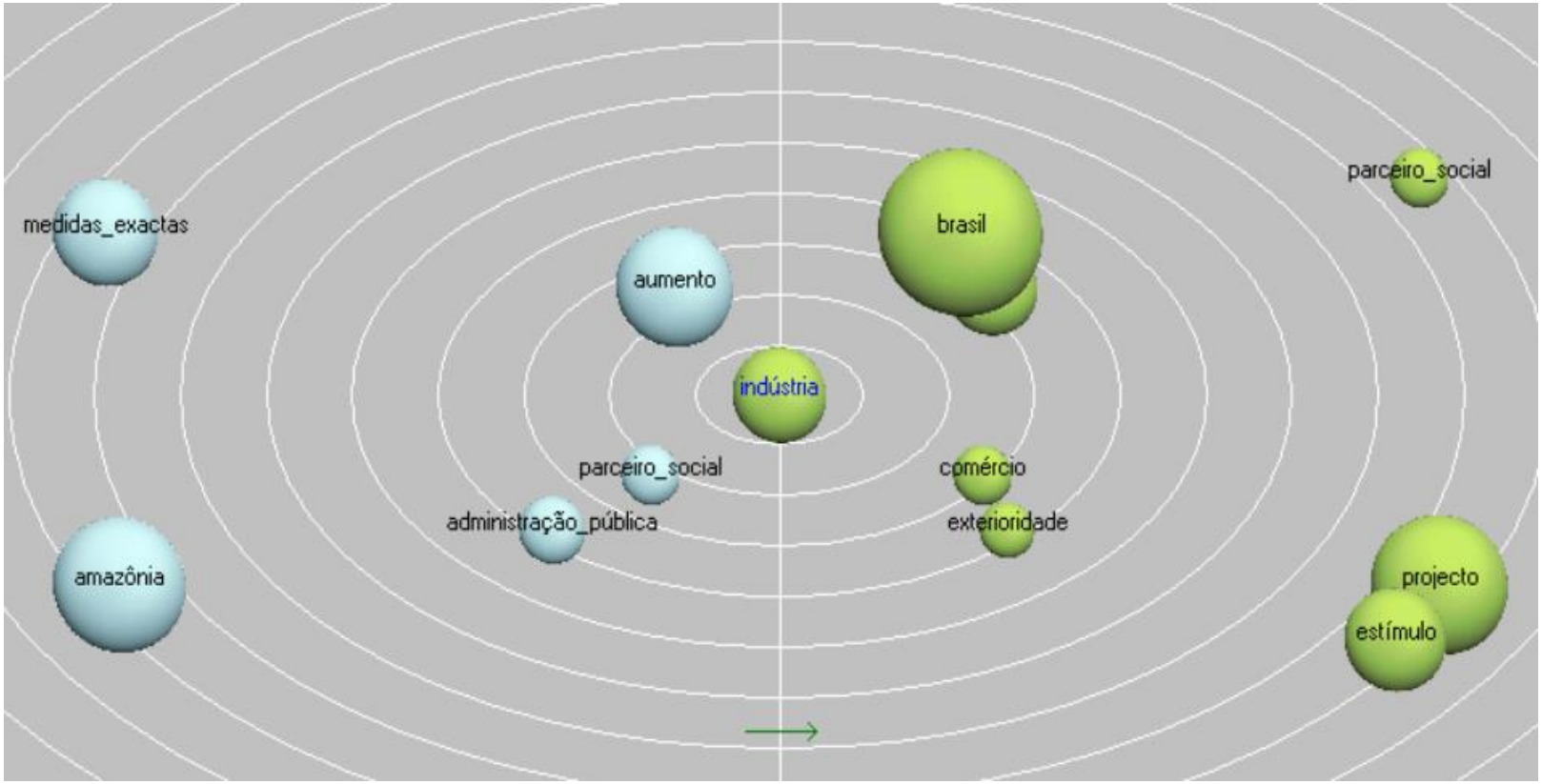
Figura 6 - Exemplo de Gráfico de relação entre Atores



Fonte: Tela do software Tropes (2016).

iii. **Gráfico de Esferas**: Neste gráfico cada referência está representada por uma esfera cuja superfície é proporcional ao número de palavras contidas. A distância entre a classe central e as outras referências é proporcional ao número de relações que as ligam. Em outras palavras, quando duas referências estão próximas elas têm muitas relações em comum, e quando estão distantes elas têm poucas relações em comum. Este tipo de gráfico permite analisar o contexto de uma referência ou categoria. As referências são orientadas: Aquelas apresentadas à esquerda da classe central são as que vêm antes, as que são apresentadas à direita são as que vêm depois.

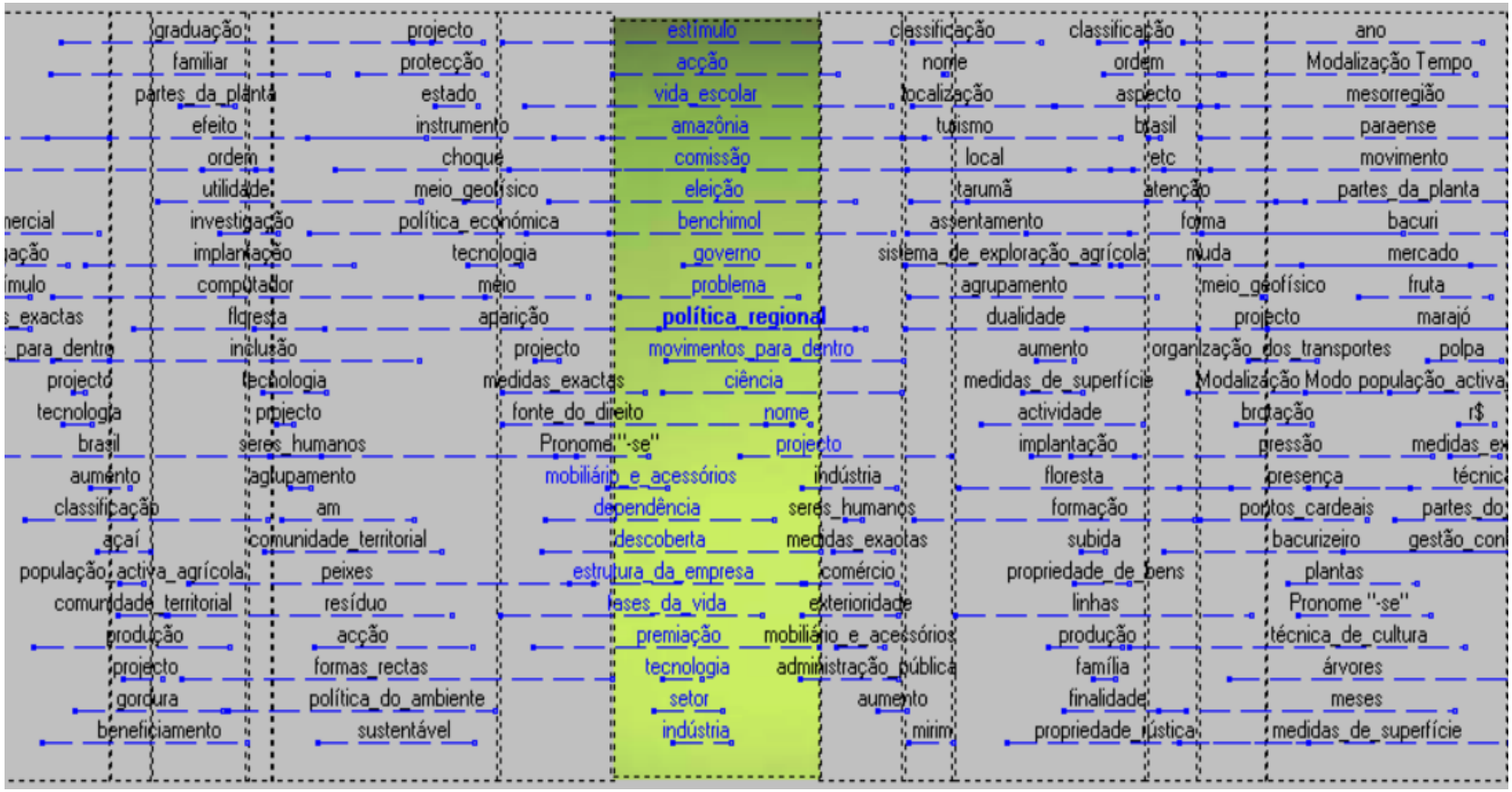
Figura 7 - Exemplo de Gráfico de Esferas



Fonte: Tela do software Tropes (2016).

iv. **Gráfico de episódios:** Neste gráfico cada rajada é mostrada como uma linha horizontal picotada que indica a sua extensão (comprimento da rajada) e a sua posição a partir do início do texto. Tal como em um gráfico de repartições, a ordem cronológica é representada no eixo horizontal do início (à esquerda) até o fim do texto (à direita). As rajadas são apresentadas de cima para baixo e da esquerda à direita, em função da sua ordem de chegada. Quando a parte baixa do episódio estiver ativa, o software tenta uma disposição em zigue-zague, para mostrar o máximo de informação possível. Os episódios são apresentados no mesmo gráfico, da esquerda para a direita, na ordem cronológica, sob a forma de grandes retângulos picotados.

Figura 8 - Exemplo de Gráfico de Episódios

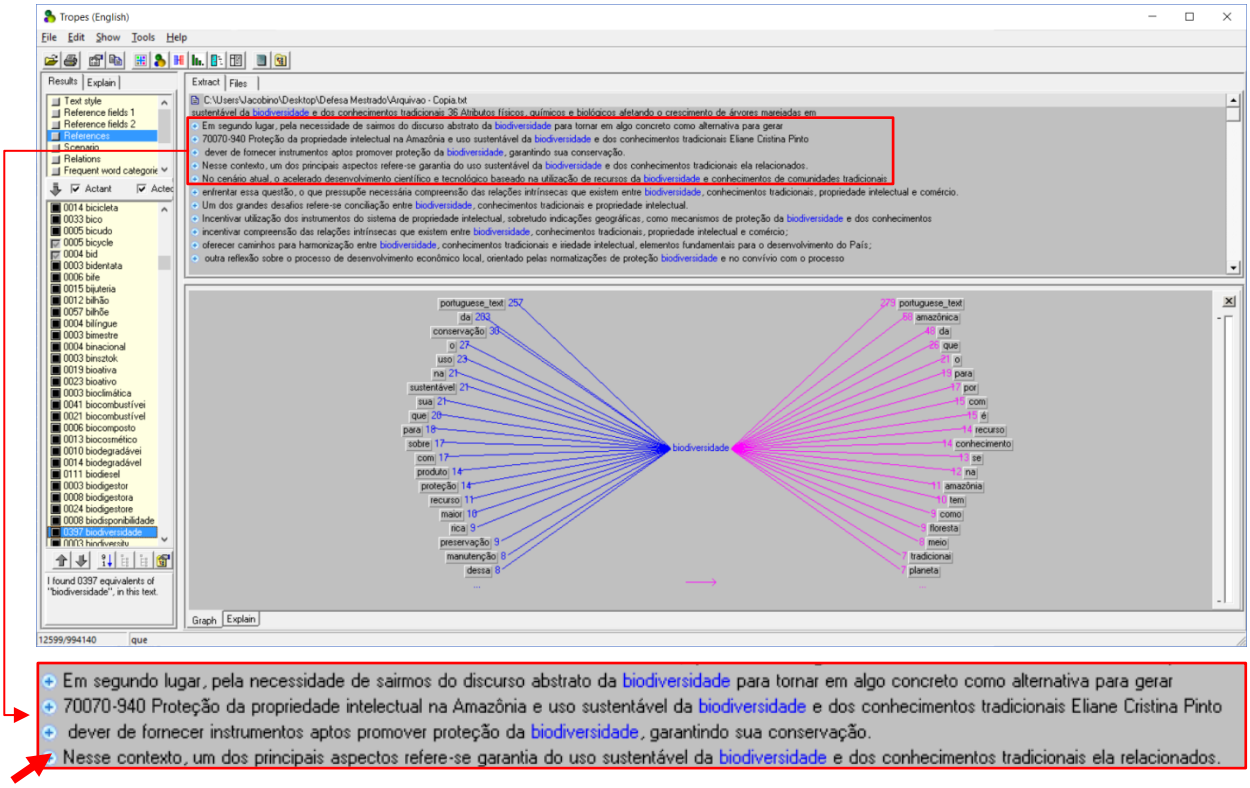


Fonte: Tela do software Tropes (2016).

Algumas dessas interfaces gráficas são pouco amigáveis, tornando difícil a leitura, compreensão e análise dos resultados apresentados por parte dos usuários. Assim, para facilitar a interpretação dos resultados foram escolhidos pelo autor o gráfico em Estrela e Gráfico de Esferas. Os resultados obtidos no sistema de mineração de textos e demonstrados automaticamente nos gráficos de estrela foram transformados em tabelas para facilitar a leitura e compreensão, todavia foram criados apêndices a este trabalho com a fonte original dos dados.

O Software permite que com apenas um clique possa-se acessar as fontes de informação que constituíram a relevância das correlações semânticas. Para exemplificar, com um simples acionamento da classe “biodiversidade”, o pesquisador pode acessar rapidamente as fontes originais de informação dos diversos documentos e arquivos de textos processados e minerados. A figura 9, a seguir, evidencia essa funcionalidade:

Figura 9 – Tela do software Tropes explorando as fontes de informação



Fonte: Extraído do software Tropes (2016) – Adaptada pelo autor.

O ambiente computacional legado por essa pesquisa possibilita que as análises possam ser feitas sob inúmeros prismas.

O termo “portuguese_text”, que aparecerá nos gráficos, é um agrupamento de palavras do idioma português (da, do, ou, em, algum, porquê, aquele, porém e outras) que o software considerou com baixo potencial semântico em relação à categoria principal que está sendo analisada. Compete destacar também que não faz parte do escopo dessa pesquisa adentrar a profundidade de cada questão oriunda do processo de descoberta de conhecimento apoiado por software. A ideia é apontar as relações semânticas com maior relevância e deixá-las evidentes para posterior investigação e estudos pelos especialistas no assunto.

Em paralelo ao trabalho de mineração de textos, foi realizada revisão de literatura tradicional. Esta revisão não se limitou às publicações dos autores participantes do prêmio. Cervo e Bervian (1983, p.55) defendem que a pesquisa bibliográfica é “aquela que explica um problema a partir de referenciais teóricos publicados em documentos. [...] Busca-se conhecer e analisar as contribuições culturais ou científicas do passado existentes sobre um determinado assunto, tema ou problema”. Por outro lado, Gil (1999) aponta que a pesquisa bibliográfica é feita mediante material já elaborado, sobretudo livros e artigos científicos. Publicações do próprio Professor Samuel Benchimol, entre outros autores, fizeram parte da revisão de literatura do presente trabalho.

Nessa perspectiva, o presente trabalho aplicou AC e mineração de textos a toda base de trabalhos e pesquisas acadêmicas do Prêmio Professor Benchimol com vistas à descoberta de questões relevantes relativas ao desenvolvimento sustentável da Amazônia. A ideia central foi olhar para todo o corpus de trabalhos apresentados, selecionados ou não, premiados ou não.

Espera-se que as análises realizadas tragam contribuições e possam apoiar a sociedade na definição de políticas para o desenvolvimento sustentável da Amazônia.

5 RESULTADOS E ANÁLISES

Esta seção apresenta o tratamento dos resultados, análises quantitativas, análises qualitativas, interpretações e inferências obtidas pela análise do corpus da pesquisa. As análises foram completadas pela extração das relações semânticas obtidas pelo processamento do acervo documental pelos software de mineração de textos.

5.1 ANÁLISE QUANTITATIVA DO CORPUS

Entre os anos de 2004 e 2015 foram recebidas um total de 1856 (um mil oitocentas e cinquenta e seis) propostas. No ano de 2009 foi lançado conjuntamente à Comenda que leva o nome de Samuel Benchimol o Prêmio Banco da Amazônia de Empreendedorismo Consciente, contudo, essas inscrições não fizeram parte do escopo dessa pesquisa. Seiscentas e vinte e nove propostas foram apresentadas na categoria “Ambiental”, seiscentas e três propostas na categoria “Econômico-Tecnológico” e seiscentas e dezenove propostas na categoria Social. O Quadro 3 apresenta a distribuição das propostas ao longo dos anos:

Quadro 3 – Número de candidaturas apresentadas por categoria[1]

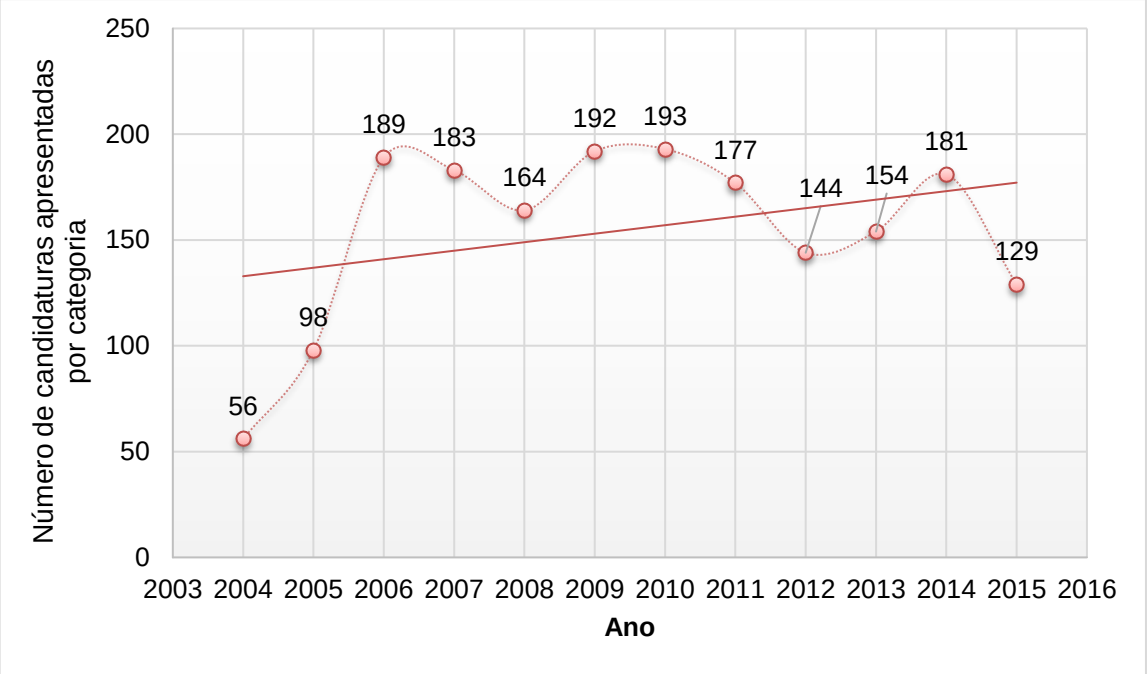
Ano	Cidade	Ambiental	Econômico-Tecnológica	Social	*Total*
2004	Manaus	13	16	18	**52**
2005	Belém	28	41	29	**98**
2006	Boa Vista	60	68	61	**189**
2007	Rio Branco	76	56	51	**183**
2008	Palmas	54	48	62	**164**
2009	Porto Velho	64	49	79	**192**
2010	Manaus	78	63	52	**193**
2011	Macapá	56	55	66	**177**
2012	Belém	44	47	53	**144**
2013	Boa Vista	48	61	45	**154**
2014	Rio Branco	62	53	66	**181**
2015	Porto Velho	46	46	37	**129**
	TOTAL	**629**	**603**	**619**	**1856**

Fonte: Elaborado pelo autor (2016).

[1] Essa pesquisa não considerou as inscrições do Prêmio Banco da Amazônia de Empreendedorismo Consciente, bem como deixou de considerar as inscrições na Categoria “Personalidade Amazônica”.

O Quadro 3 torna possível inferir que: Houve uma grande aceleração no número de candidaturas nos primeiros três anos; um aumento gradativo do número de propostas, com pequenas quedas, chegando ao topo no ano de 2010, com cento e noventa e três candidaturas; e, então, percebe-se uma tendência de diminuição no número de propostas nos últimos cinco anos. O Gráfico 1 ilustra a percepção da evolução das candidaturas ao longo dos anos:

Gráfico 1 – Evolução das Candidaturas



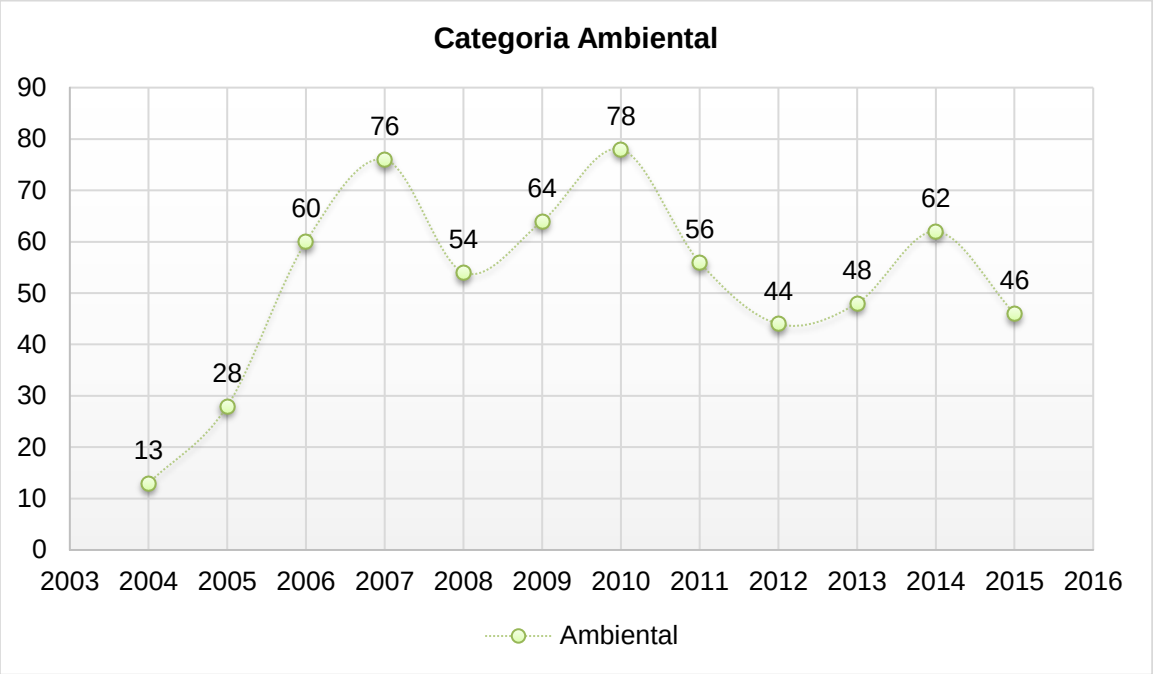
Fonte: Elaborado pelo autor (2016).

Pode-se inferir pelo gráfico que, mesmo com uma linha de tendência ascendente, o número de propostas caiu consideravelmente nos últimos anos. Em apenas cinco anos, foram três pontos abaixo da linha de tendência, o pior deles no ano de 2015, com apenas 129 candidaturas, uma queda de 33,16% em relação ao ano de 2010.

Para Alvares e Ferreira (2014), essa trajetória descendente está relacionada ao cenário da pesquisa científica e tecnológica no Brasil dos últimos anos, marcado por longos períodos de greve nas universidades federais. Os autores levantam a hipótese de que “estando o calendário acadêmico fortemente comprimido para não haver perdas de semestre nos cursos de graduação e pós-graduação, a motivação para participar da premiação foi comprometida” (ALVARES, FERREIRA, 2014).

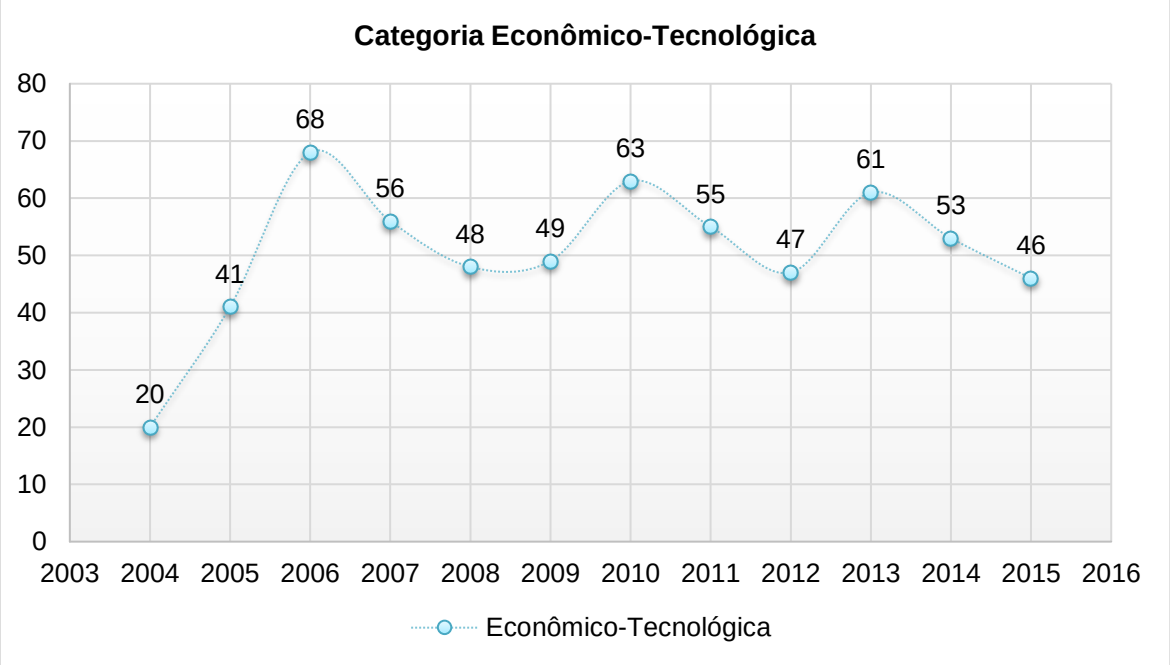
Os Gráficos 2, 3 e 4 a seguir demonstram a evolução das propostas submetidas ao Prêmio Professor Samuel Benchimol segmentadas por categoria, enquanto o Gráfico 5 demonstra a distribuição geral.

Gráfico 2 - Evolução das propostas da Categoria Ambiental



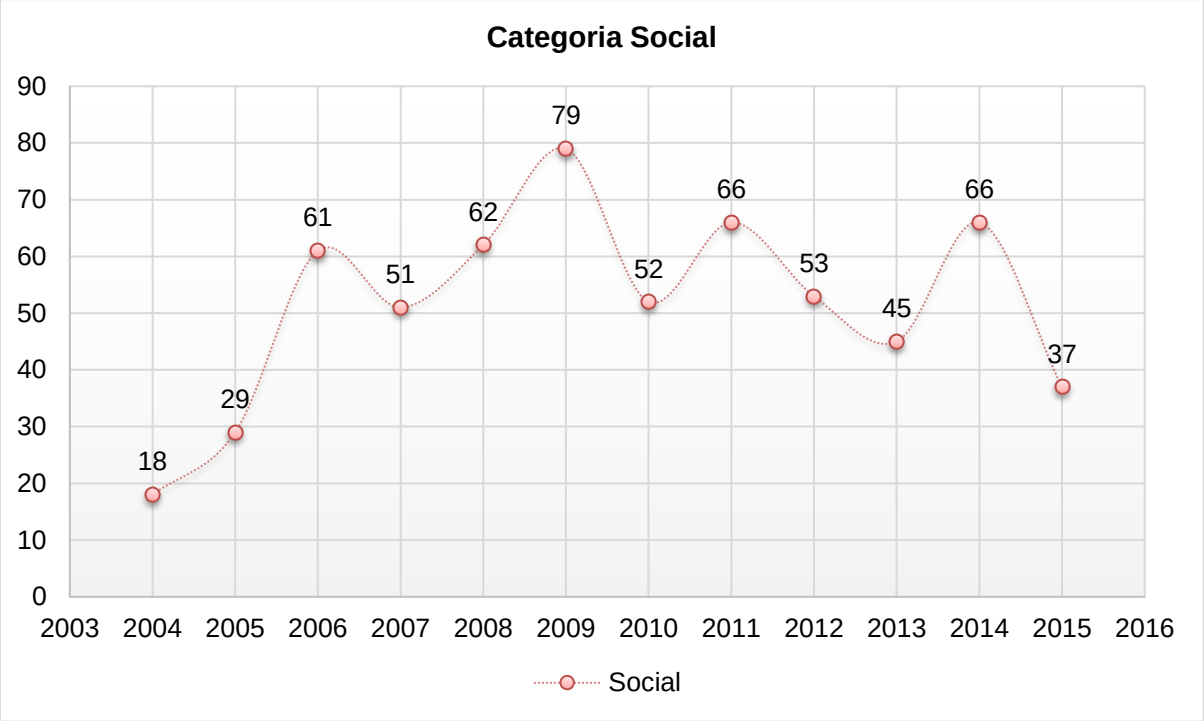
Fonte: Elaborado pelo autor (2016).

Gráfico 3 - Evolução das propostas da Categoria Econômico-Tecnológica



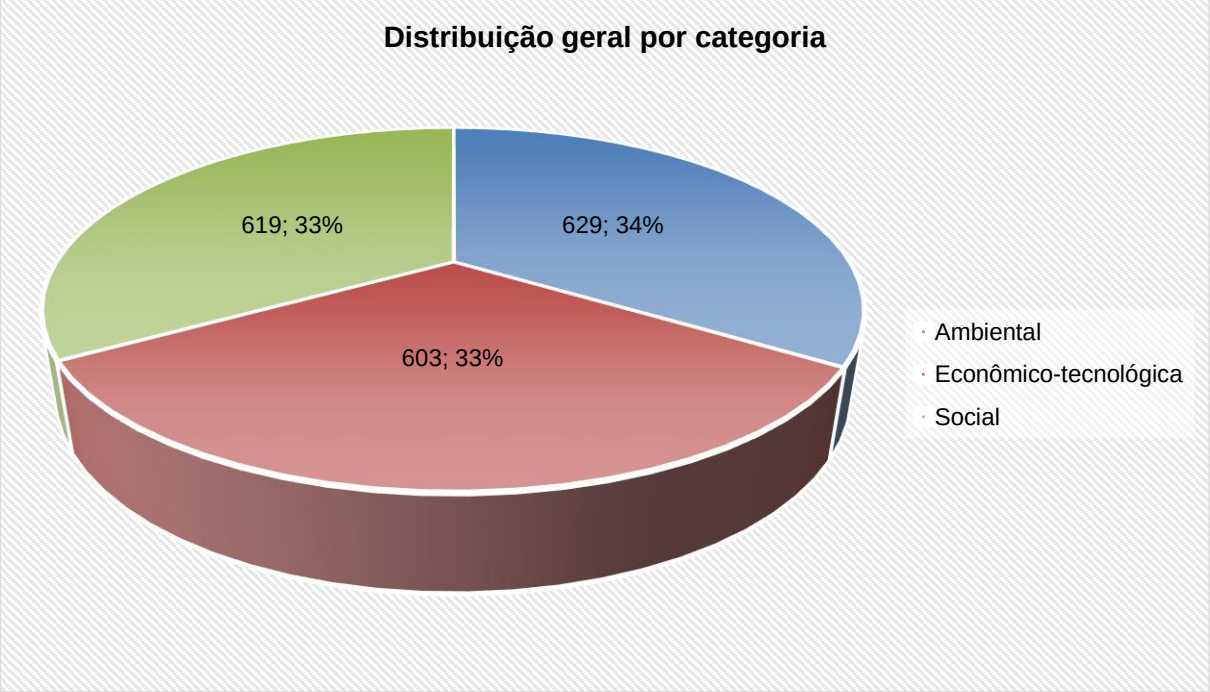
Fonte: Elaborado pelo autor (2016).

Gráfico 4 - Evolução das propostas da Categoria Social



Fonte: Elaborado pelo autor (2016).

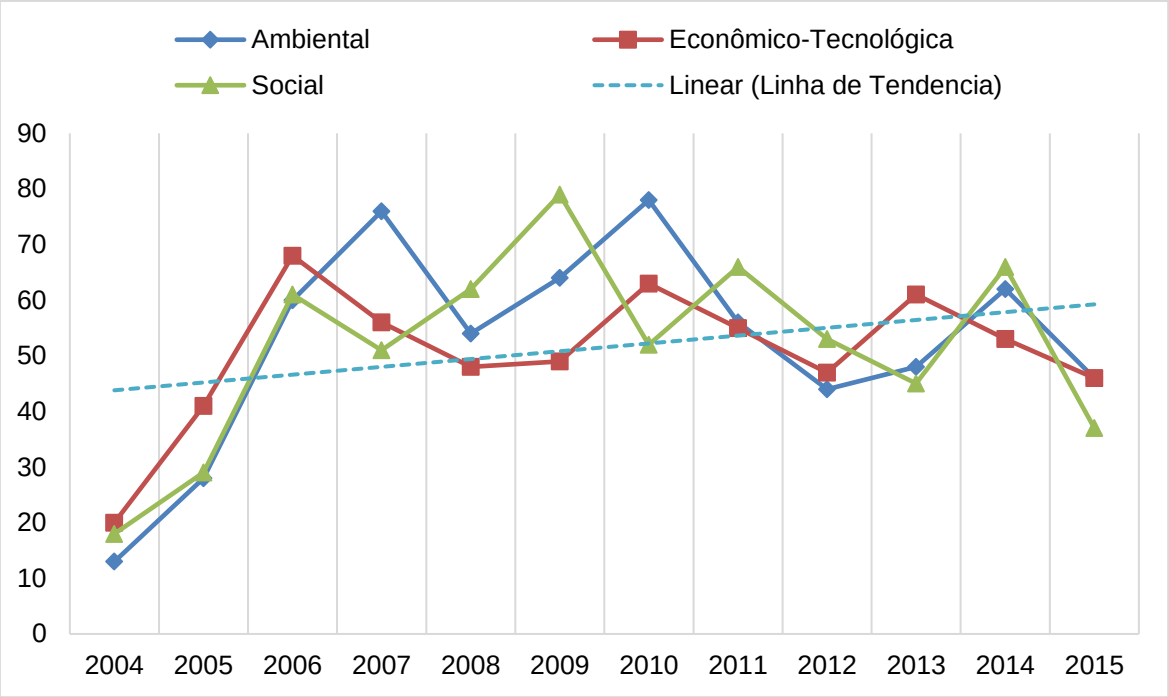
Gráfico 5 - Distribuição geral por categoria



Fonte: Elaborado pelo autor (2016).

A Categoria Ambiental representa aproximadamente 34% do total das candidaturas, enquanto as Categorias Econômico-Tecnológica e Social representam aproximadamente 33% cada, mostrando um equilíbrio na distribuição média total. Todavia, quando se analisa a distribuição ano-a-ano, percebe-se que existem variações bem mais acentuadas. O Gráfico 6 a seguir evidencia essa constatação:

Gráfico 6 – Análise Comparativa da Evolução das Propostas



Fonte: Elaborado pelo autor (2016).

Em que pese haver distribuição média muito parecida para os quantitativos de candidaturas apresentadas, quando são considerados os números totais acumulados, podemos observar pelo gráfico que a cada ano a proporção varia consideravelmente. No ano de 2009, por exemplo, enquanto a Categoria “Social” recebeu setenta propostas, a Categoria “Econômico-tecnológica” alcançou apenas quarenta e nove candidaturas.

Os números apresentados e demonstrados no quadro e gráficos, representam a grandeza do acervo documental do prêmio. São quase duas mil propostas apresentadas nestes onze anos, formando uma valiosa base de conhecimento sobre as propostas e questões-chave da Amazônia.

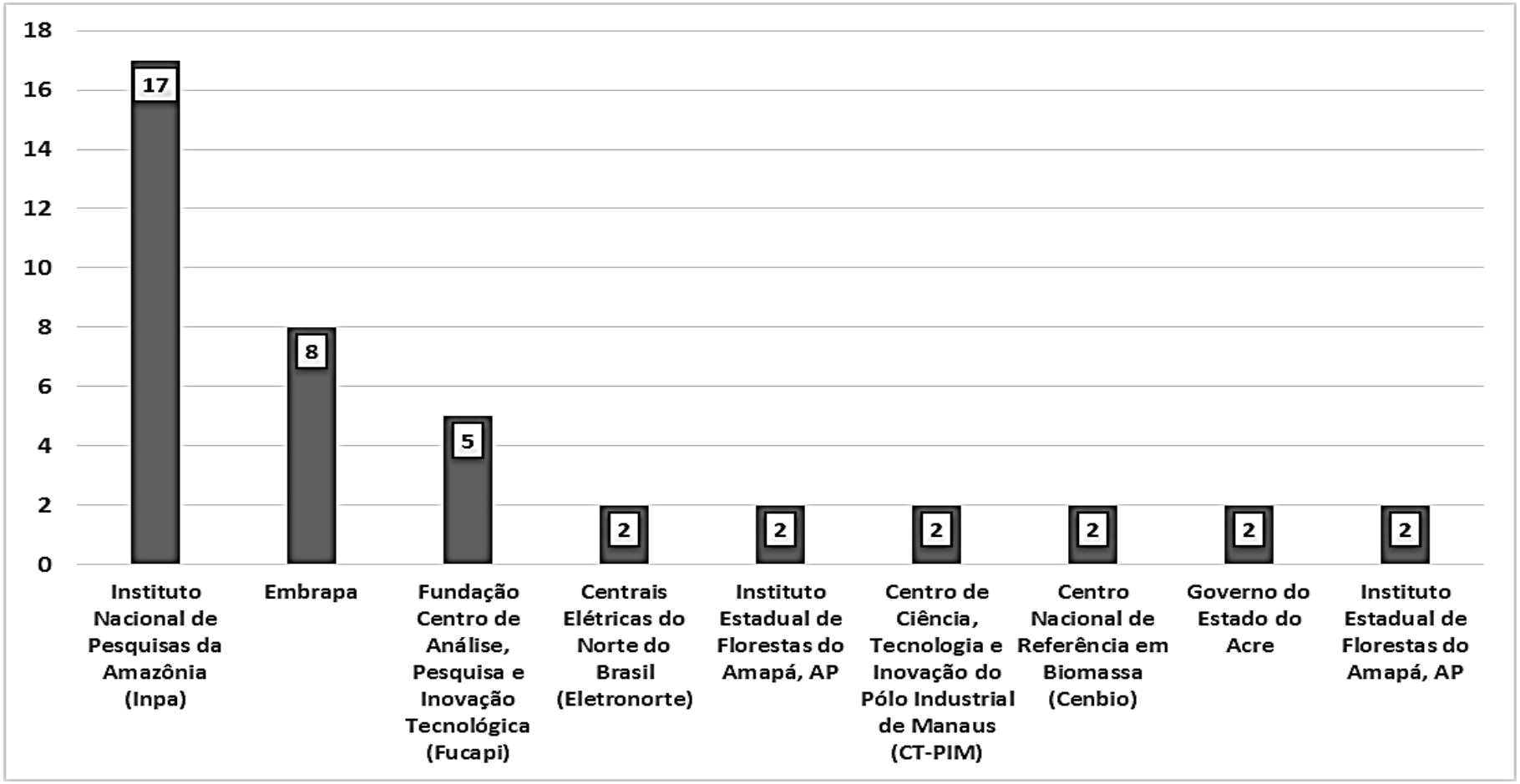
5.1.1 Os participantes, instituições autores e vencedores

Desde a sua primeira edição em 2004 o Prêmio Professor Samuel Benchimol recebeu propostas, projetos e pesquisas científicas de inúmeros autores, instituições, institutos de pesquisa, universidades, empresas, além de candidaturas avulsas de pessoas não vinculadas a uma instituição em particular. Percebeu-se, porém, que é mantida a correlação de dispersão proposta por Bradford (1934), em que um seleto grupo de pesquisadores respondem por um elevado número de pesquisas na temática específica, enquanto a maioria dos pesquisadores apresentam apenas uma pesquisa.

Pesquisadores como: João Tito Borges; Alfredo Kingo Oyama Homma; Suani Teixeira Coelho; Raimundo Nonato Lemos da Silva; Marilene Gomes de Sá Ribeiro; Ruy Alexandre de Sá Ribeiro; Jadir de Souza Rocha; Cynthia Lins Falcone Pontes; Tereza Maria Farias Bessa; Vania Maria Oliveira da Câmara Lima; e Idelfonso Generoso da Silva apresentaram inúmeras propostas ao longo dos anos. Esses pesquisadores foram destacados não apenas pelo volume da produção científica, propostas e projetos, mas pela qualidade do trabalho. Todos foram reconhecidos e receberam a comenda em primeiro lugar em mais de uma edição do Prêmio.

O fenômeno identificado por Bradford (1934) também pode ser observado em relação às instituições, institutos de pesquisa, empresas e universidades. O gráfico 7, a seguir, traz as instituições agraciadas nas categorias ambiental, econômico-tecnológica e social, tanto no primeiro quanto nos segundo e terceiro lugares. Cabe destacar os resultados obtidos pelo Instituto Nacional de Pesquisas da Amazônia (INPA), dezessete vezes premiado.

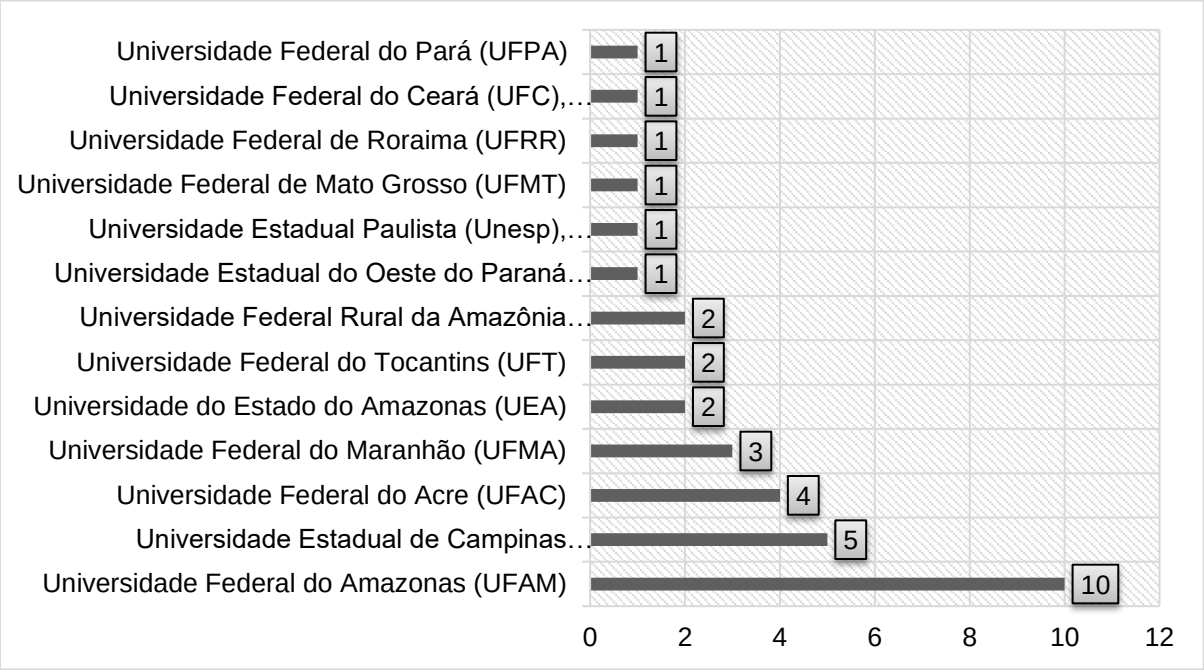
Gráfico 7 – Instituições com maior número de premiações



Fonte: Elaborado pelo autor (2016).

Seguindo o caminho antes percorrido por Alvares e Ferreira (2014), foi realizado um recorte específico para a análise da participação das universidades. O gráfico 08 apresenta a distribuição e ranqueamento das mais agraciadas:

Gráfico 8 – Universidades mais agraciadas



Fonte: Elaborado pelo autor (2016).

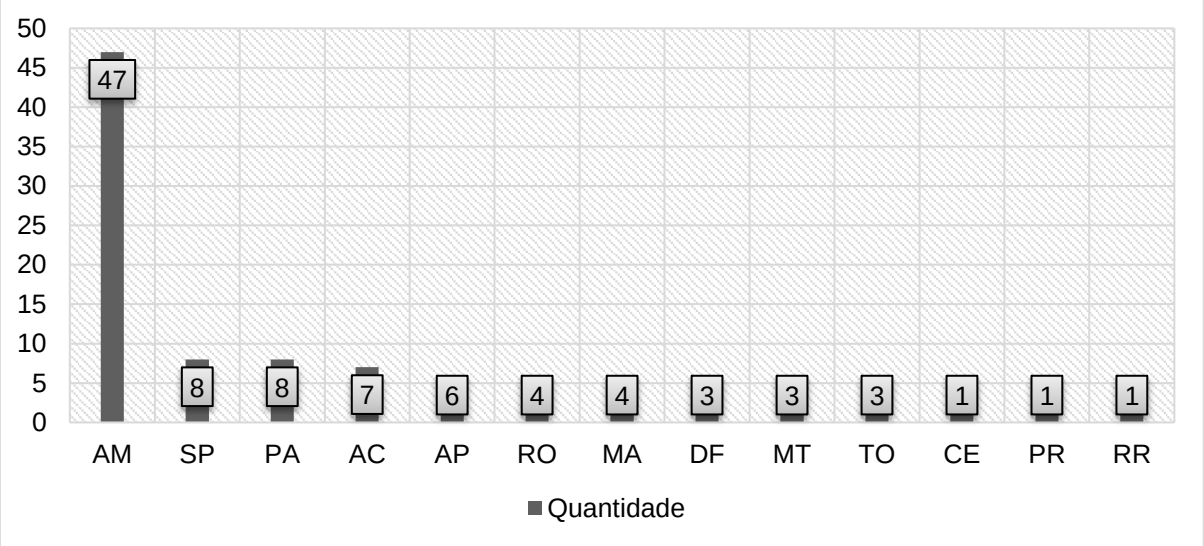
Cabe realçar as posições de destaque: da Universidade Federal do Amazonas, com dez propostas vencedoras; da Universidade Estadual de Campinas (Unicamp), com cinco propostas vencedoras; e da Universidade Federal do Acre (UFAC), com quatro propostas vencedoras. As universidades juntas arrematararam o primeiro lugar no Prêmio Professor Samuel Benchimol por 34 vezes.

Para Alvares e Ferreira (2014), a maciça participação das instituições de pesquisa nos primeiros lugares “é uma afirmação da missão dessas instituições”. O conjunto da obra demonstra que o Prêmio Professor Samuel Benchimol tem conseguido alcançar os seus objetivos. Percebe-se o claro envolvimento dos institutos de pesquisa, das universidades, da Embrapa, de empresas, de agentes públicos e privados voltados à temática do desenvolvimento ambiental, econômico-tecnológico e social da região amazônica.

5.1.2 Participação dos estados

A grande maioria das candidaturas apresentadas e premiadas vem dos estados da Região Amazônica. O estado do Amazonas, com quarenta e sete propostas vencedoras, representa sozinho aproximadamente 50% do total de premiações. Contudo, outros estados como São Paulo, Ceará, Paraná e o Distrito Federal tiveram propostas vencedoras. Um grande destaque deve ser dado ao estado de São Paulo, que figura na segunda posição na quantidade de propostas vencedoras, com oito premiações, à frente de outros estados da região Norte, como Pará e Acre.

Gráfico 9 – Participação dos Estados



Fonte: Elaborado pelo autor (2016).

Por meio do gráfico podem-se extrair claramente dois fatos: i) Uma grande disparidade entre o número de candidaturas e vitórias oriundas do estado do Amazonas, logo, uma grande oportunidade para crescimento das pesquisas nos outros estados; e ii) E também que todos os estados pertencentes à Amazônia Legal (Acre, Amapá, Amazonas, Pará, Rondônia, Roraima, Tocantins, Parte do Mato Grosso e do Maranhão), tiveram candidaturas agraciadas ao longo das várias edições do prêmio.

Ao analisar de forma quantitativa as propostas apresentadas ao Prêmio Professor Samuel Benchimol, de um total de um mil oitocentas e cinquenta e seis propostas, tem-se a média de cento e cinquenta e cinco candidaturas por ano. Pode-

se deduzir que os objetivos, relativos ao estímulo das pesquisas, ideias e projetos inovadores, têm sido alcançados.

Há uma tendência de estabilização no número de candidaturas apresentadas por ano (cento e oitenta), com aproximadamente cinquenta e cinco candidaturas por categoria. Além dos estados da região amazônica, destacaram-se outras unidades da federação vencedores da premiação (Ceará, Distrito Federal, Paraná e São Paulo). Todos esses elementos somados mostram a maturidade do prêmio, bem como consolidam-no como forte instrumento na geração de inovação, no que tange ao desenvolvimento sustentável da região amazônica.

5.2 ANÁLISE QUALITATIVA DO CORPUS

Conforme previsto na metodologia desta pesquisa e devido ao grande volume de dados do corpus estudado, foram utilizadas ferramentas computacionais para apoio à análise semântica de conteúdo. A mineração de textos envolveu processamento de toda a base textual das propostas apresentadas ao Prêmio Professor Samuel Benchimol.

5.2.1 Análises da Natureza Ambiental

A análise da natureza “ambiental” mostra as principais associações e relações semânticas vinculadas a este tema, quer sejam problemas, quer sejam oportunidades. As questões foram apresentadas, identificadas, ordenadas, classificadas e analisadas levando em consideração o número e a relevância das ocorrências detectadas automaticamente pelos software ou, excepcionalmente, quando o pesquisador julgou necessário acrescentar alguma inferência.

A natureza “ambiental” apresentou um conjunto total de 1.766 (um mil setecentos e sessenta e seis) correlações semânticas. Cabe dar ênfase ao termo “educação”, que apresentou um conjunto de 295 (duzentas e noventa e cinco) correlações de predecessão (lado esquerdo do gráfico, em azul) com o termo

“ambiental”. De tal modo, pode-se inferir que as propostas, projetos e pesquisas relacionadas à “educação ambiental” apareceram com grande frequência no corpus pesquisado. Contudo, explorando o lado direito do gráfico, com cinquenta e seis relações de sucessão à categoria “ambiental”, destaca-se o termo “comunidade” (lado direito do gráfico, em rosa). Um cenário possível e passível de maior investigação seria analisar as propostas de ações voltadas à educação ambiental nas comunidades locais, seguindo a trilha apontada pelo software: educação/ ambiental/ comunidade.

Tabela 1 – Correlações semânticas da categoria "ambiental"

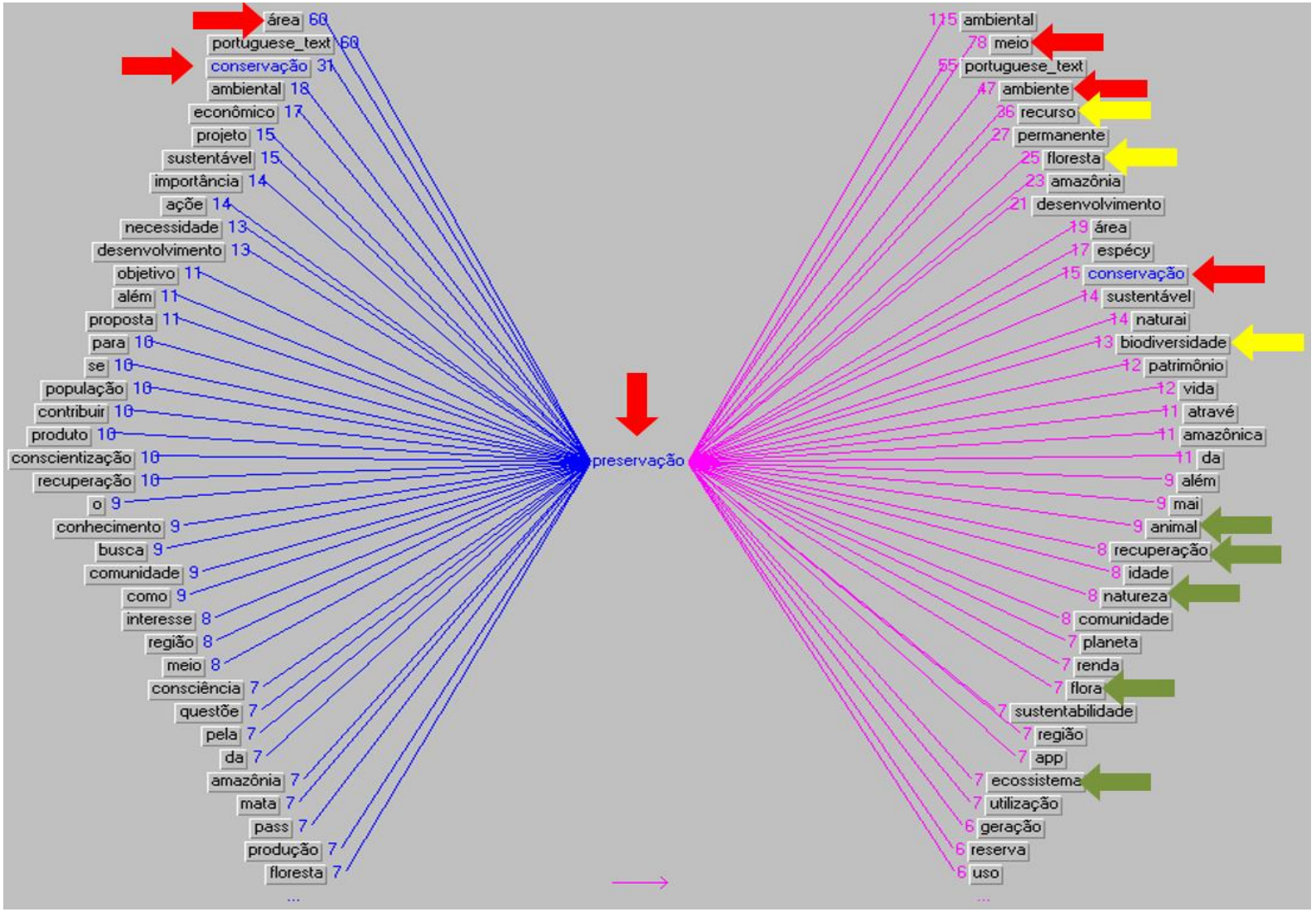
Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
educação	295				56	**comunidade**
preservação	115				39	**desenvolvimento**
gestão	102				41	**econômico**
impacto	94				31	**escola**
área	88		ambiental		29	**sustentável**
conservação	62				22	**indígena**
sustentabilidade	55				20	**sistema**
degradação	40				19	**resíduo**
meio	34				18	**processo**
natureza	27				18	**preservação**

Fonte: Extraído do software Tropes (2016) – Adaptado pelo autor – Apêndice II.

A Tabela 1 apresenta uma visão das principais categorias identificadas automaticamente pelo Software de Text Mining que se conectam à categoria “ambiental”. O sistema evidenciou um conjunto considerável de possíveis linhas de estudo, ou seja, outras relações podem ser examinadas, como por exemplo, a educação ambiental na escola, que apresentou trinta e uma relações de sucessão à cadeia educação / ambiental. Em que pese a correlação “educação ambiental na escola” não estar entre as melhores ranqueadas pelo software, saltou aos olhos desse autor. Acredita-se que um especialista no tema poderia explorar as principais questões associadas à “educação ambiental” e aprofundar o entendimento destas.

Com 115 (cento e quinze) correlações semânticas de predecessão à categoria “ambiental”, o termo “preservação”, do mesmo modo, mereceu atenção. Ao centralizar a categoria preservação, o gráfico em estrela mudou para a seguinte configuração:

Gráfico 10 – Correlações semânticas da categoria preservação



Fonte: Tela do software Tropes (2016).

O Gráfico 10, acima, e a Tabela 2, abaixo, expõem as principais categorias identificadas automaticamente pelo que se conectam à categoria “preservação”. Analisando o lado esquerdo do gráfico, tem-se como principal correlação semântica à categoria central a classe “área”, com sessenta correlações, seguida por “conservação”, com trinta e uma ocorrências, possibilitando que algumas inferências pudessem ser feitas a partir desse ponto: i) Conservação do “meio” + “ambiente”; ii) Conservação do “recurso” natural; iii) Conservação “permanente” da “floresta”; iv) Conservação da “biodiversidade”; v) Conservação da vida “animal”, fauna; vi) Conservação da “flora”; vii) Conservação do “ecossistema”; entre outras.

Tabela 2 - Correlações semânticas da categoria "preservação"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
			→			
área	60		preservação		115	ambiental
conservação	31				78	**meio**
ambiental	18				47	**ambiente**
sustentável	15				36	**recurso**
desenvolvimento	13				25	**floresta**
população	10				15	**conservação**
conscientização	10				13	**biodiversidade**
recuperação	10				9	**animal**
comunidade	9				8	**recuperação**
mata	7				8	**natureza**

Fonte: Extraído do software tropes (2016) - Adaptado pelo autor.

Observe-se que ainda no âmbito da categoria “ambiental” pode-se observar que os problemas relativos à “conservação” e “preservação” do meio ambiente e seus “recursos” naturais foram muito citados nas pesquisas, projetos, propostas e trabalhos apresentados ao Prêmio Professor Samuel Benchimol.

A Tabela 3, a seguir, apresenta uma análise relativa aos problemas com a conservação do meio ambiente. A trilha do software foi a seguinte: ambiente/ preservação/ conservação/ recuperação/ área + degradada.

Tabela 3 - Correlações semânticas da categoria "recuperação"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
alternativa	28				115	**área**
área	17				75	**degradada**
degradada	12				25	**mata**
construção	9				16	**ciliar**
nativa	9		recuperação		10	**pastagem**
preservação	8				10	**igarapé**
atividade	7				8	**nascente**
econômica	6				7	**recurso**
programa	5				6	**solo**
produção	5				6	**reserva**

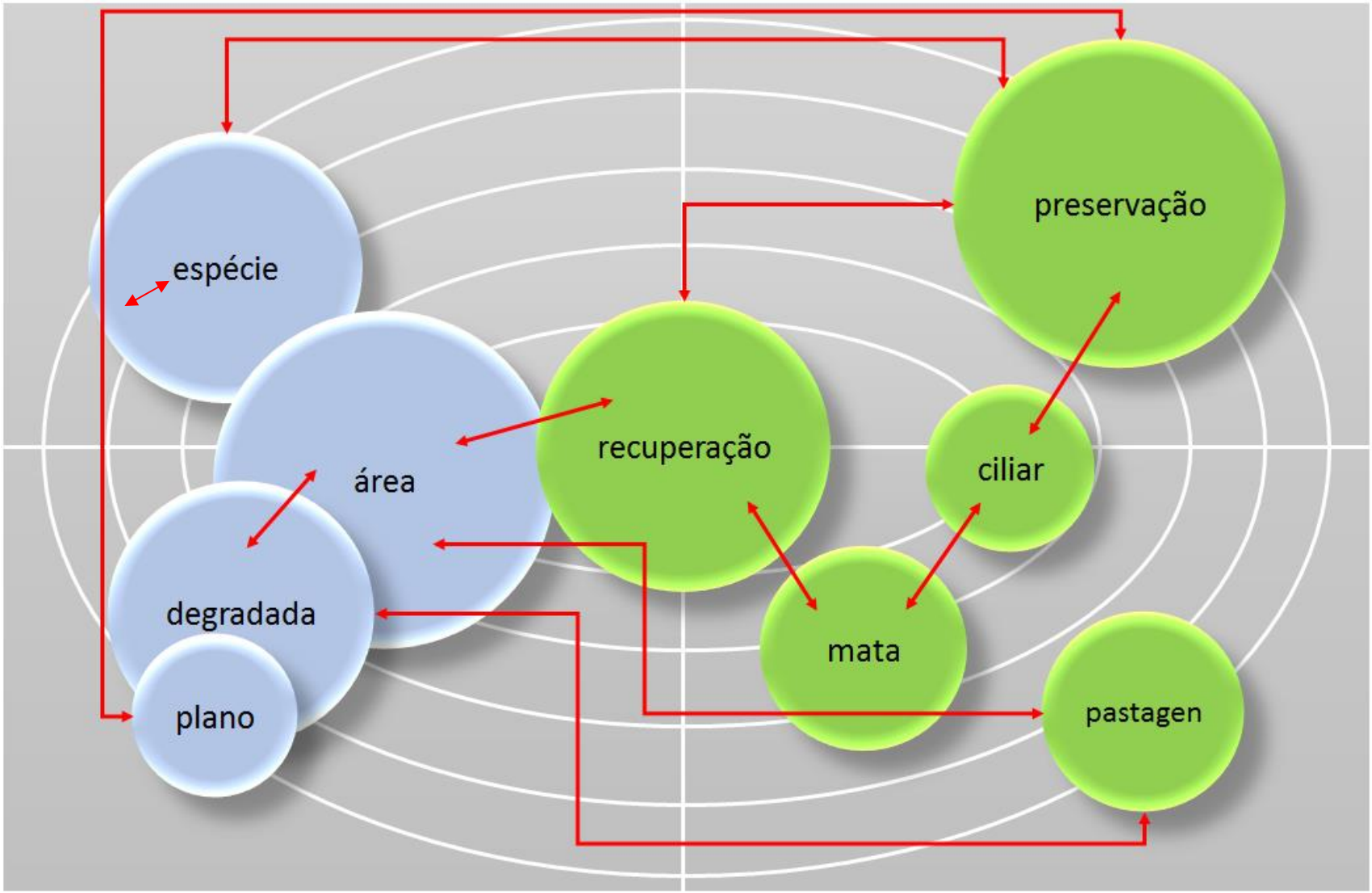
Fonte: Extraído do software Tropes (2016) - Adaptado pelo autor - Apêndice III.

O lado esquerdo da Tabela 3 exclama a relevância das “alternativas” para recuperar as “áreas” “degradadas”. Já o lado oposto, aponta as principais classes que sucedem a categoria “recuperação”. Nesse prisma, pode-se inferir a urgência e a importância das questões relativas à recuperação da(s) “mata(s) + ciliar(es)” no contexto das áreas degradadas.

A trilha do software que possibilita essa inferência é: alternativa/ recuperação/ área/ degradada/ mata/ ciliar(es). Destacada pelo software, uma causa provável da degradação dessas áreas é o aumento do desmatamento para inserção de novas áreas de pastagens. O software correlacionou vários eventos à degradação da mata ciliar com a classe “pastagen”.

O Gráfico de esferas número 11, a seguir, possibilita uma melhor visualização das correlações semânticas entre as categorias em questão:

Gráfico 11 – correlações semânticas da “recuperação” de áreas degradadas



Fonte: Adaptado da tela do software Tropes (2016).

Como pode-se observar pelo Gráfico 11, a classe “pastagen” aparece com muita relevância, o que permite inferir a relação do termo com a degradação das matas ciliares. Todavia, faz-se necessária uma análise mais aprofundada das reais relações de causa e efeito para confirmar ou refutar esta tese.

Até aqui, explorando a natureza “Ambiental”, tem-se como uma questão-chave apontada pelo software de mineração de textos a recuperação das áreas degradadas pelo desmatamento e aumento das áreas de pastagens, deixando o problema evidenciado. Contudo, pode-se inferir alguma proposta provável para solução deste problema? A resposta é: Sim.

Ao centralizar a classe “degradada” pela trilha ambiente/ preservação/ conservação/ recuperação/ degradada, salta aos olhos o termo “reflorestamento”, com nove ocorrências de predecessão e quatro ocorrências de sucessão ao termo “degradada”. A Tabela 4 abaixo torna visível a correlação pela seguinte trilha: recuperação/ área/ degradada/ reflorestamento.

Tabela 4 - Correlações semânticas da categoria "degradada"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
área	296				20	**amazônia**
recuperação	75				12	**recuperação**
alternativa	17				11	**mata**
reflorestamento	9				8	**atividade**
construção	8		degradada		7	**desenvolvimento**
desenvolvimento	5				7	**nativa**
regeneração	5				6	**mesorregião**
plantio	5				6	**produção**
parcialmente	5				5	**preservação**
sustentável	3				4	**reflorestamento**

Fonte: Extraído do software Tropes (2016) - Adaptado pelo Autor - Apêndice IV.

Por meio da análise das correlações semânticas da categoria “preservação” pode-se fazer inferências acerca das principais questões, problemas e possíveis soluções da região Amazônica sob a ótica do acervo documental do Prêmio Professor Samuel Benchimol. As inferências realizadas não consideraram a opinião de especialistas em meio ambiente, foram apenas constatações lógicas, obtidas pelo autor desta pesquisa, por meio da observação do número de ocorrência das classes e categorias de palavras geradas pelo software de mineração de textos.

Em síntese, os principais desafios encontrados para a categoria “Ambiental”, no âmbito do acervo do Prêmio Professor Samuel Benchimol, identificados por meio da mineração de textos, estão relacionados à implantação de alternativas para: i) Educação ambiental nas comunidades locais e na escola, para que o conhecimento traga benefícios no curto, médio e longo prazos, gerando uma cultura de valorização do meio ambiente e de seus recursos; ii) Preservação e conservação do ecossistema e da biodiversidade de forma a garantir o equilíbrio ecológico da região e subsistência das espécies de fauna e flora; e iii) Recuperação das áreas degradadas, principalmente a recuperação das matas ciliares, tendo como principais alternativas os projetos de reflorestamento.

5.2.2 Análises da Natureza Econômico-Tecnológica

A análise da natureza “Econômico-tecnológica” apresenta as principais associações e relações semânticas vinculadas ao desenvolvimento econômico, tecnológico e do empreendedorismo na região amazônica.

O software de mineração de textos identificou um conjunto de 4584 (quatro mil quinhentas e oitenta e quatro) correlações semânticas para o termo “amazônia”, sendo que, a classe “empreendedorismo”, foi a melhor ranqueada na rede de sucessão, apresentou 1017 (um mil e dezessete) correspondências lógicas com a classe central conforme pode-se observar na Tabela 5, a seguir:

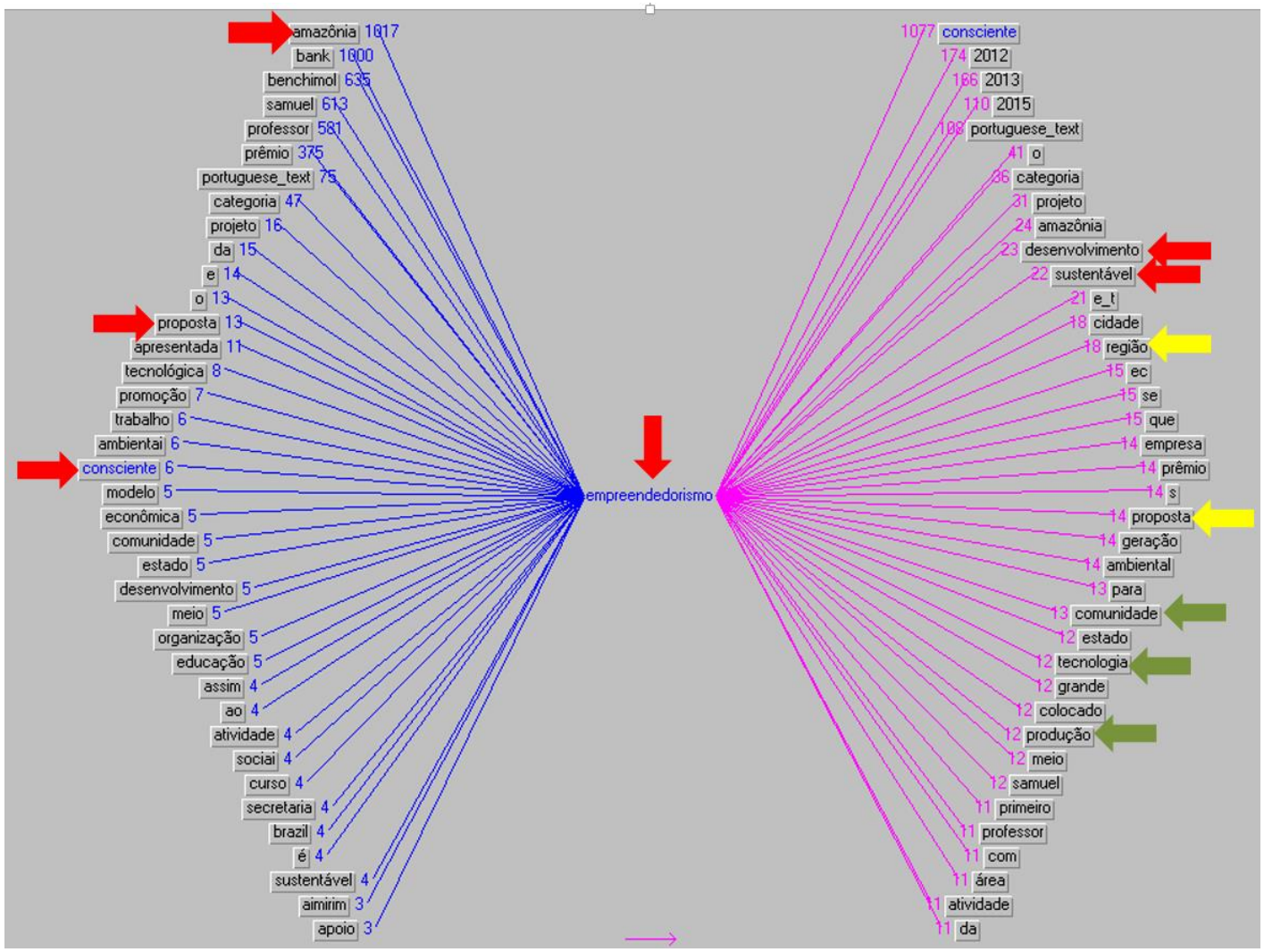
Tabela 5 - Correlações semânticas da categoria "amazônia"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
benchimol	662				1017	**empreendedorismo**
pesquisa	213				981	**consciente**
desenvolvimento	205				107	projeto
sustentável	162				81	desenvolvimento
comunidade	82		amazônia		77	**proposta**
empresa	58				59	**sustentável**
tecnologia	51				46	**produção**
ambiental	49				37	**tecnologia**
recurso	42				36	**indústria**
ribeirinha	37				36	**floresta**

Fonte: Extraído do software Tropes (2016) - Adaptado pelo autor - Apêndice V.

A Tabela 5 apresenta inúmeras trilhas para investigação e descoberta de conhecimentos relacionados ao desenvolvimento da Amazônia. Entre as correlações mais evidentes, destaca-se a trilha: desenvolvimento/ sustentável/ amazônia/ empreendedorismo/ consciente/ tecnologia. Centenas de relações lógicas aparecem ligando essas classes, o que abre um grande caminho para a investigação, o estudo e a exploração dessa trilha como uma das questões-chave para a natureza “econômico-tecnológica”. Centralizando a categoria de sucessão (empreendedorismo) com a maior quantidade de ocorrências, obtém-se a seguinte configuração:

Gráfico 12 – correlações semânticas da classe “empreendedorismo”



Fonte: Tela do software Tropes (2016).

As principais relações semânticas da classe central identificadas pelo software foram: i) “consciente”, com um mil setenta e sete correlações; ii) “desenvolvimento + sustentável” com vinte e três e vinte e duas correlações, respectivamente; iii) “proposta”, com quatorze ocorrências; iv) “comunidade”, com treze ocorrências; e v) “tecnologia” e “produção” com doze ocorrências cada.

Tabela 6 - Correlações semânticas da classe "empreendedorismo"

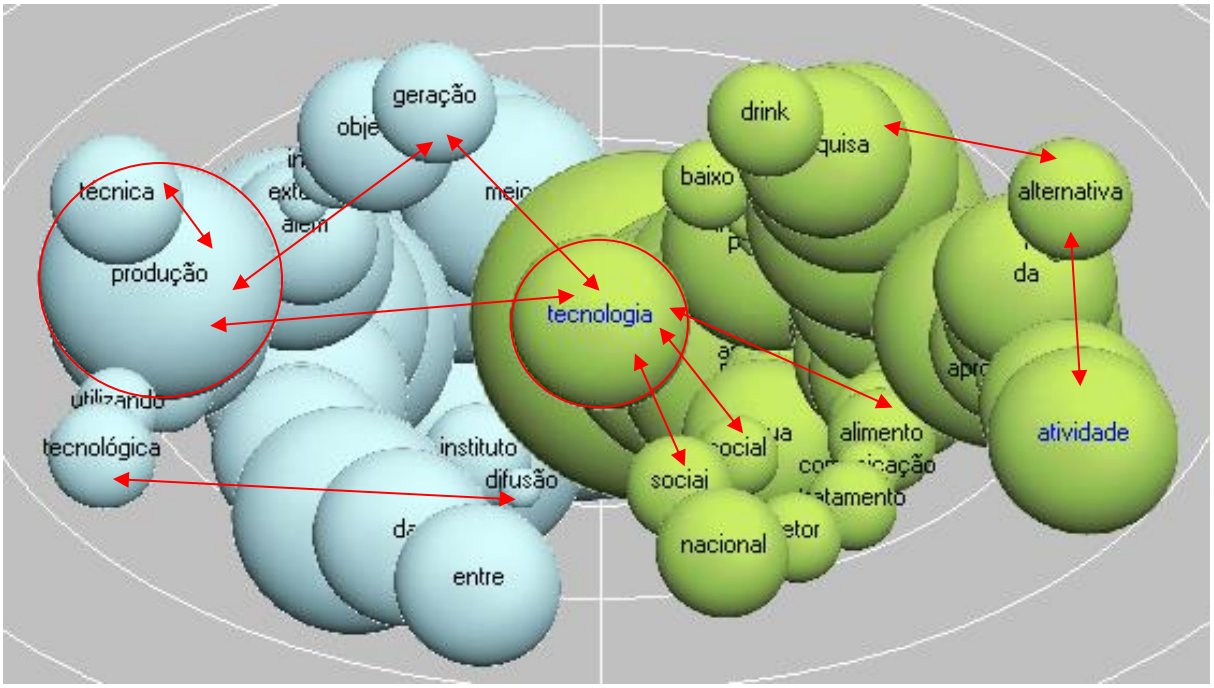
Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
			→			
amazônia	1017		empreende-dorismo		1077	consciente
benchimol	635				36	categoria
proposta	13				23	**desenvolvimento**
tecnológica	8				22	**sustentável**
trabalho	6				18	**região**
consciente	6				14	**proposta**
comunidade	5				13	**comunidade**
desenvolvimento	5				12	**tecnologia**
atividade	4				12	**produção**
sustentável	4				11	**atividade**

Fonte: Extraído do software Tropes (2016) - Adaptado pelo autor.

Pode-se inferir, pelas trilhas apresentadas pelo software, em especial: amazônia/ empreendedorismo/ tecnologia/ produção, que as principais propostas para desenvolvimento sustentável da região amazônica, sob a ótica do Prêmio Professor Samuel Benchimol, perpassam o empreendedorismo consciente para desenvolvimento da região com o envolvimento da comunidade local e o uso de tecnologias sustentáveis para aumento da produção.

Nesse ponto, por escolha do autor dessa pesquisa, buscou-se entender melhor a classe “tecnologia”, suas correlações semânticas e principais associações lógicas indicadas pelo software de mineração de textos. Ao centralizar o termo “tecnologia”, na visualização do gráfico de esferas, obteve-se a seguinte configuração:

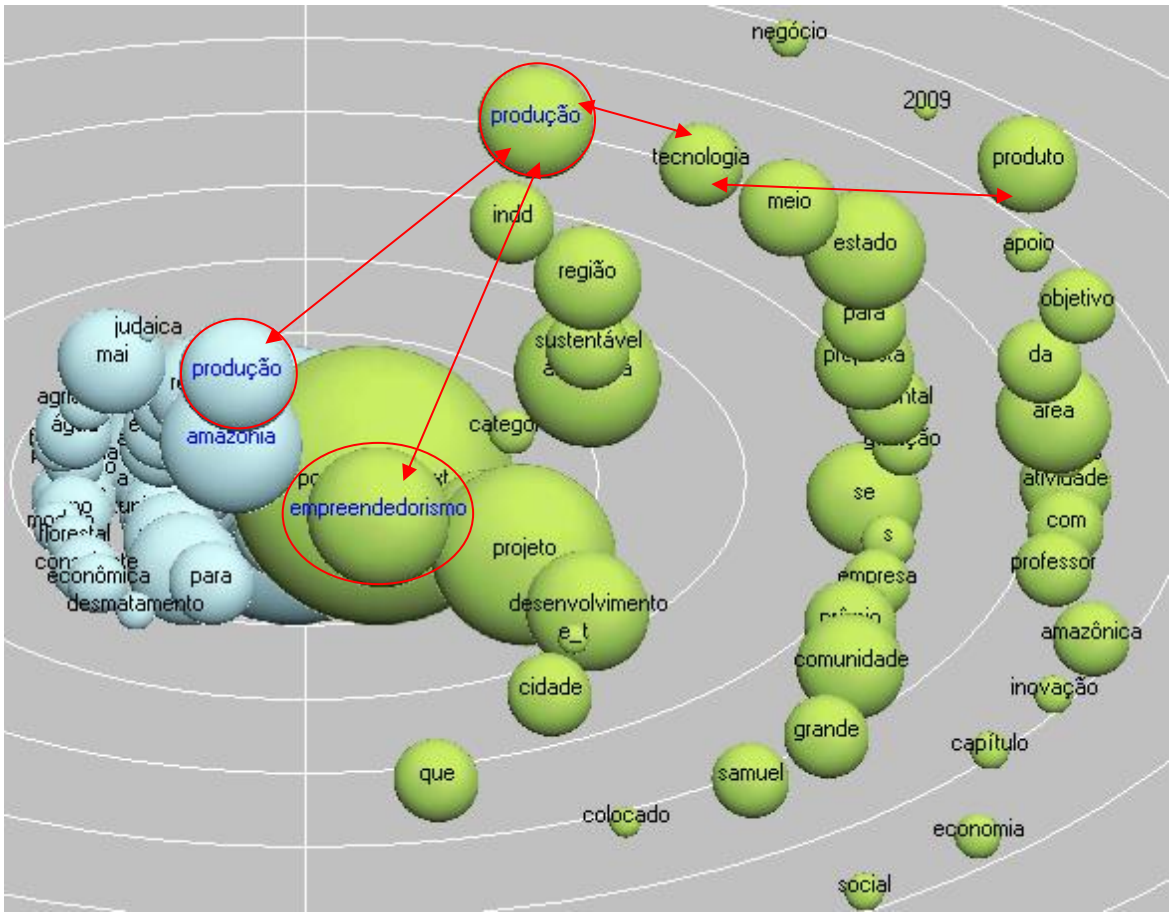
Gráfico 13 – correlações semânticas da classe “tecnologia”



Fonte: Tela do software Tropes (2016).

O termo “tecnologia” guardou relação com diversas classes e em destaque, pelo tamanho da esfera, com a classe “produção”. O termo produção também aparece em relação de predecessão e sucessão à classe “empreendedorismo”, conforme Gráfico 14, a seguir:

Gráfico 14 – correlações semânticas em esferas da classe “empreendedorismo”



Fonte: Tela do software Tropes (2016).

Analisando em maior profundidade a categoria “produção”, chega-se a um nível de granularidade de informação interessante, em que se podem extrair algumas constatações. A Tabela 7, a seguir, torna mais evidente:

Tabela 7 - Correlações semânticas em esferas da classe "produção"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
sistema	118				117	**muda**
desenvolvimento	83				106	**produto**
custo	76				101	**óleo**
atividade	68				95	**alimento**
aumento	56		produção		85	**comercialização**
produto	41				68	**agrícola**
implantação	41				62	**fruto**
resíduo	38				53	**planta**
proposta	34				48	**peixe**
indústria	29				48	**energia**

Fonte: Extraído do software Tropes (2016) - Adaptado pelo autor - Apêndice VI.

Inúmeras trilhas podem ser estudadas e investigadas a partir da análise das relações semânticas intrínsecas à categoria “produção”. Os resultados da mineração de textos deixam evidentes as inúmeras alternativas para produção e geração de riqueza disponíveis e viáveis para a Amazônia. Pode-se destacar, entre outras: i) A produção de mudas, em especial de madeiras de alto valor comercial e plantas ornamentais; ii) A produção de óleos e extratos vegetais para as áreas farmacêuticas, cosméticas e fins medicinais; iii) A produção e comercialização de alimentos, produtos orgânicos e derivados agrícolas; e iv) Produção de energia limpa a partir de fontes renováveis como biodiesel e etanol, além do potencial hídrico.

O software de mineração de textos utilizado apresentou uma limitação na quantidade de informação que pode disponibilizar na tela. Foram exibidas apenas as classes que apresentaram mais de trinta correlações semânticas à categoria central (produção). Todavia, várias outras propostas interessantes foram percebidas, como por exemplo: i) A produção madeireira sustentável com práticas de manejo; ii) O extrativismo de frutos e sementes; iii) As atividades de ecoturismo; além iv) Da atividade de industrialização, comercialização e exportação da produção local. Essa visão pode ser ampliada pela Tabela 8, que se segue:

Tabela 8 - Correlações semânticas da classe "indústria"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
federação	148				164	estado
ministério	111				50	**cosmético**
nacional	40				30	**madeireira**
confederação	30				29	**farmacêutica**
produção	26		indústria		20	tecnologia
produto	22				15	**alimentícia**
implantação	15				14	**cosmética**
insumo	14				13	**química**
interesse	12				11	**turismo**
comércio	10				11	**siderúrgica**

Fonte: Extraído do software Tropes (2016) - Adaptado pelo Autor - Apêndice VII.

A indústria tem destaque nas alternativas para o desenvolvimento econômico-tecnológico da região amazônica. Dentre os principais ramos da indústria, identificados automaticamente pelo software de mineração de textos pode-se destacar: i) A indústria “cosmética”, com cinquenta relações semânticas; ii) A indústria “madeireira”, com trinta correlações; iii) A indústria “farmacêutica”, com vinte e nove

relações semânticas; iv) A indústria “alimentícia” com quinze relações; v) O “turismo” e a indústria “siderúrgica” com onze correlações semânticas cada.

Em suma, os principais desafios encontrados para a natureza “Econômico-tecnológica”, no âmbito do acervo do Prêmio Professor Samuel Benchimol, identificados por meio da mineração de textos, são relacionados ao desenvolvimento sustentável da Amazônia por meio do empreendedorismo consciente para desenvolvimento da região, com o envolvimento das comunidades locais. Têm destaque também as propostas para uso de tecnologias sustentáveis para aumento da qualidade e da quantidade da produção, além da necessidade de promoção e ampliação da indústria sustentável da região.

5.2.3 Natureza Social

A análise da natureza “Social”, apresenta as principais associações e relações semânticas vinculadas às questões sociais na Amazônia. Após o processamento total do acervo do Prêmio Professor Samuel Benchimol, o software de mineração de textos detectou automaticamente um conjunto de um mil trezentos e cinquenta relações semânticas com o termo “social”. A Tabela 9, a seguir, torna claras as associações mais relevantes determinadas pelo sistema:

Tabela 9 - Correlações semânticas da classe "social"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
inclusão	167		social		65	**ambiental**
responsabilidade	44				31	econômico
organização	30				22	**comunidade**
vulnerabilidade	16				14	**projeto**
projeto	13				9	**política**
contexto	11				8	**família**
assistência	10				7	**criança**
integração	9				5	**indígena**
mobilização	8				5	**saúde**
sustentabilidade	7				4	**idoso**

Fonte: Extraído do software Tropes (2016) - Adaptado pelo Autor - Apêndice VIII.

Ao analisar o lado esquerdo da Tabela 9, percebe-se como melhor ranqueado o termo “inclusão”, com cento e sessenta e sete correlações semânticas. Infere-se,

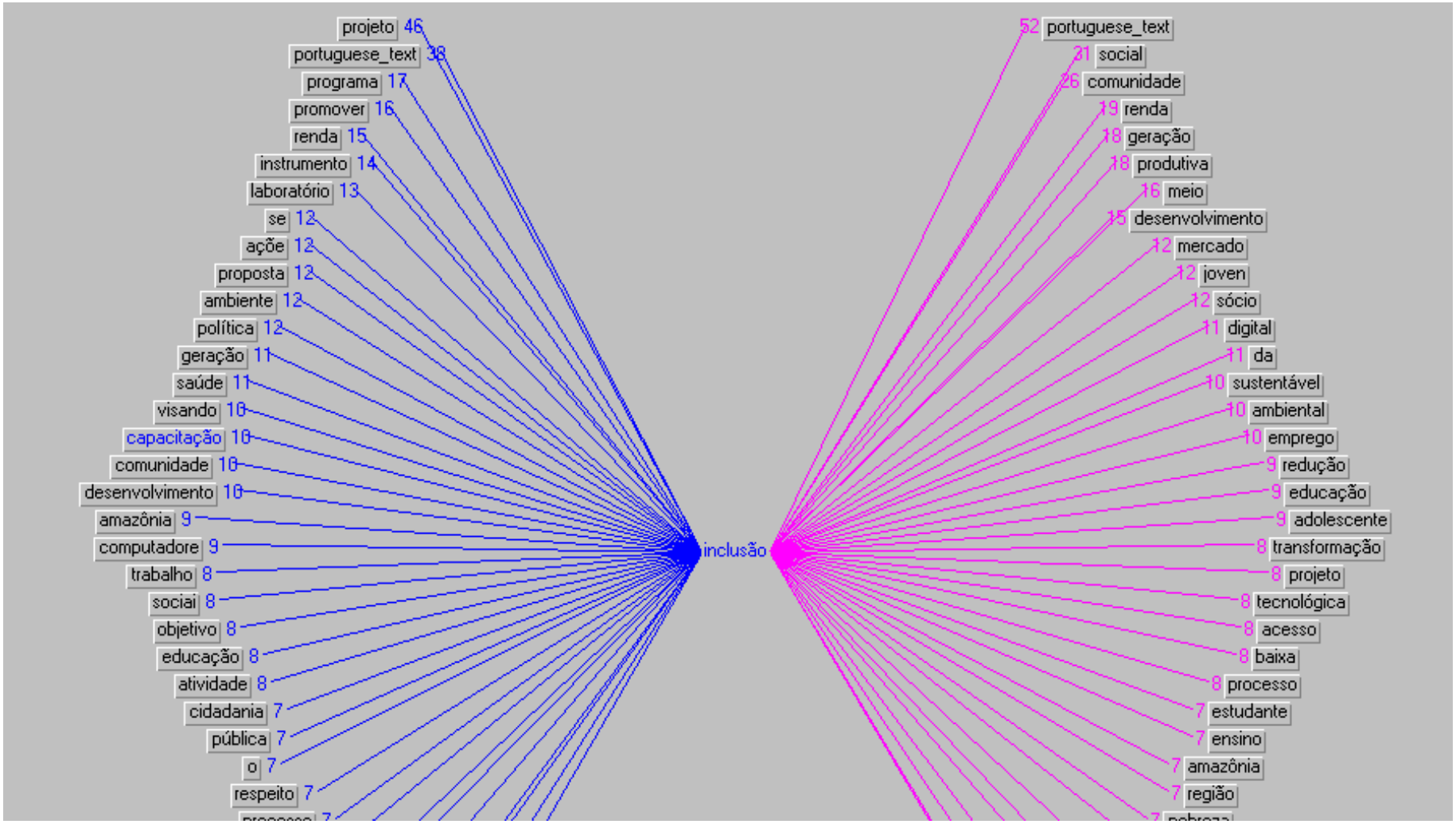
então, que as principais questões, problemas projetos e ações, contidas no corpus da pesquisa, no que se refere à natureza “Social”, tratam de iniciativas para “inclusão social”. O Prêmio Professor Samuel Benchimol recebeu ao longo dos anos inúmeras propostas que sugerem meios e ações para combater à exclusão aos benefícios da vida em sociedade. As propostas perpassam as questões de classe social, de educação, de deficiência, de etnia, de sexualidade, de preconceito social ou de preconceitos raciais.

A Tabela 9, de maneira geral, mostra várias trilhas para análise de outras questões como: a “responsabilidade” social, com quarenta e quatro correlações semânticas; a “vulnerabilidade” social, com dezesseis relações semânticas; e o “controle” social, com quatorze correlações semânticas. Mudando o prisma, ao analisarmos o lado direito do gráfico, destacam-se as questões socioambientais como: o termo “ambiental” correlacionado sessenta e cinco vezes; a classe “comunidade”, com vinte e duas relações semânticas; o termo “projeto”, que aparece com quatorze relacionamentos lógicos; a classe “política”, com nove relações; entre outras.

Dentre os vários caminhos possíveis, o autor dessa pesquisa identificou a seguinte trilha: vulnerabilidade/ social/ família/ criança/ idoso. Essa trilha induz o pensamento que as famílias, principalmente em seus entes mais vulneráveis (crianças e idosos), foram objeto de propostas e pesquisas no âmbito do corpus analisado. Essa linha de pensamento não exclui outras relações semânticas geradas pelo sistema, como a necessidade de inclusão social de jovens e adolescentes, conforme será demonstrado no Gráfico 15 e Tabela 10, à frente.

Ao navegar no software pelas correlações semânticas da categoria “social”, percebem-se inúmeras oportunidades para posteriores explorações e estudos, principalmente relacionados à “inclusão” social. Ao centralizar a categoria “inclusão” tem-se a seguinte configuração:

Gráfico 15 – Correlações semânticas da classe “inclusão”



Fonte: Tela do software Tropes (2016).

Sobressai no Gráfico 15 e Tabela 10 a classe “projeto”, com quarenta e seis correlações lógicas e, ainda, a categoria “programa” com dezessete relações semânticas. Infere-se, pois, que boa parte das propostas que permeiam a categoria “inclusão”, militam a favor da implantação de “programas” e “projetos” para inclusão social nas comunidades. Uma das formas de inclusão, destacada pelo software, com dez relações semânticas, foi o processo de “capacitação”, intrinsicamente ligado à geração de “emprego” e “renda”, bem como, da “redução” das desigualdades (lado esquerdo).

Tabela 10 - Correlações semânticas da classe "inclusão"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
projeto	46				31	**social**
programa	17				26	**comunidade**
renda	15				19	**renda**
proposta	12				12	**jovem**
geração	11		inclusão		11	**digital**
capacitação	10				9	**educação**
trabalho	8				9	**adolescente**
educação	8				8	**tecnológica**
cidadania	7				8	**acesso**
alternativa	6				7	pobreza

Fonte: Extraído do software Tropes (2016) - Adaptado pelo Autor.

Ainda, no lado esquerdo, nota-se a trilha que liga os pontos: inclusão/ social/ comunidade/ geração/ renda/ jovem/ emprego. Inferem-se, então, como questões-chave para o desenvolvimento social da região amazônica: os projetos e os programas para inclusão social na comunidade com foco na geração de empregos e renda para os jovens. Essa leitura vertical foi gerada automaticamente pelo sistema por meio do cruzamento de dados da base documental do Prêmio Professor Samuel Benchimol.

Completando as análises da categoria “social”, cabe destacar a classe “doença”, seguindo a seguinte trilha: social/ vulnerabilidade/ saúde/ doença. A Tabela 11, a seguir, apresenta as principais correlações semânticas associadas ao termo “doença”.

Tabela 11 - Correlações semânticas da classe "doença"

Termo antecessor	n° ocorrências		categoria		n° ocorrências	Termo sucessor
prevenção	31				18	**transmissível**
controle	28				16	**veiculação**
tratamento	11				16	**sexualmente**
malária	9				10	**saúde**
risco	8		doença		10	**degenerativa**
resistência	7				10	**infecciosa**
ocorrência	7				8	**hematológica**
transmissores	6				7	**malária**
proliferação	6				7	**chaga**
reduzir	5				7	**dengue**

Fonte: Extraído do software Tropes (2016) - Adaptado pelo Autor - Apêndice IX.

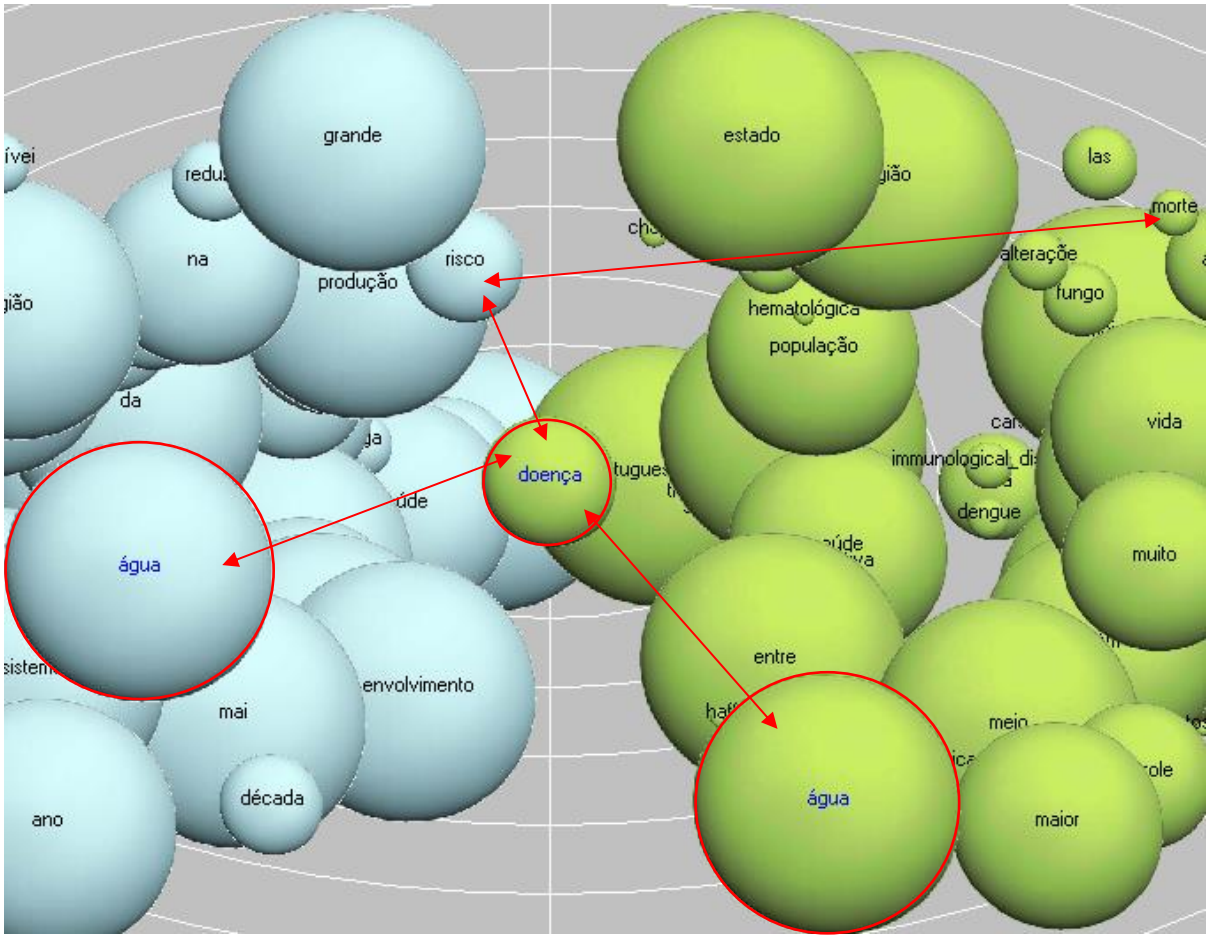
O ranqueamento gerado pelo software de mineração de textos ateve-se às relações lógicas entre as palavras que tiveram maior ocorrência nos textos, seguindo os princípios da Lei de Zipf, com base no algoritmo NaiveBayes. É provável que, sob a ótica de um especialista em saúde da região amazônica, o ranqueamento seria alterado. Todavia, seria humanamente impossível ao especialista ler integralmente todo o corpus da pesquisa para produzir as associações geradas automaticamente pelo sistema, mesmo que com grau superior de fidelidade.

Com base apenas no número de ocorrências das correlações semânticas pôde-se inferir que: um foco importante dado pelos trabalhos apresentados ao Prêmio Professor Samuel Benchimol foi dado aos projetos de “prevenção” e “controle” de doenças e morbidades fortemente frequentes na Amazônia. Aos lados esquerdo e direito da tabela destaca-se a “malária”, com relação de predecessão e sucessão à categoria central, com nove e sete relações semânticas, respectivamente.

A outra inferência feita a partir da tabela foi a constatação da preocupação, presente nas propostas, com a veiculação de doenças sexualmente transmissíveis. Todavia, outras doenças e morbidades também foram destacadas pelo software de mineração de textos: i) Doenças “degenerativas” e “infecciosas”, com dez relações semânticas cada; ii) Doenças “hematológicas”, com oito correlações; iii) Doença de “chagas”, “desordem imunológica” e “dengue”, com sete relações lógicas cada; iv) Cólera, com seis relações lógicas; e v) “Hepatite”, com cinco relações lógicas.

Ao verificar o gráfico de esferas associado à categoria “doença” pode-se perceber a correlação entre as doenças mais comuns e os problemas de saneamento básico e baixa qualidade da água:

Gráfico 16 – Correlações semânticas das classes “doença” e “água”



Fonte: Extraído do software Tropes (2016).

Em resumo, os principais desafios encontrados para a natureza “Social”, no âmbito do acervo do Prêmio Professor Samuel Benchimol, identificados por meio da mineração de textos, são: i) As propostas de iniciativas para “inclusão social” como alternativa para redução da desigualdade; ii) Os problemas de vulnerabilidade social das famílias, em especial as crianças e os idosos; iii) Os projetos e os programas para inclusão social na comunidade, focados na geração de empregos e renda para os jovens; iv) Os projetos de “prevenção” e “controle” de doenças e morbidades fortemente frequentes na Amazônia; e v) A correlação entre as doenças mais comuns e os problemas de saneamento básico, como a baixa qualidade da água.

5.2.4 Questões-chave do desenvolvimento da Amazônia

Após analisadas separadamente as três naturezas escolhidas para a pesquisa (Ambiental, Econômico-tecnológica e Social), pode-se inferir as questões-chave do desenvolvimento da Amazônia sob a ótica do Prêmio Professor Samuel Benchimol. O quadro-síntese a seguir lista as questões:

Quadro 4 - Questões-chave da Amazônia

ID	Questão
01	Promoção da educação ambiental nas comunidades locais e nas escolas, para que o conhecimento traga benefícios no curto, médio e longo prazos, gerando uma cultura de valorização do meio ambiente e de seus recursos. Faz-se necessária a disseminação da consciência ambiental e da valorização do potencial de crescimento econômico-social da região pelo pensamento eco-desenvolvimentista.
02	Implantação de alternativas para preservação e conservação do ecossistema e da biodiversidade de forma a garantir o equilíbrio ecológico da região e subsistência das espécies de fauna e flora. Essas alternativas garantirão a perpetuação das riquezas naturais.
03	A necessidade de criação e manutenção de programas para recuperação das áreas degradadas, principalmente a recuperação das matas ciliares, tendo como principais alternativas os projetos de reflorestamento e evitando a conversão de áreas florestais para agricultura ou criação de gado; mineração degradativa e irregular; exploração ilegal de madeireira e incêndios florestais.
04	Implementação de programas, projetos e iniciativas para “inclusão social” como alternativas para redução da desigualdade com promoção e valorização das comunidades locais, sua cultura e potenciais.
05	Intervenção do Estado com a Implantação de políticas públicas para tratar os inúmeros problemas de vulnerabilidade social das famílias, em especial as crianças e os idosos.

06	Criação de projetos e programas para inclusão social na comunidade, focados na geração de empregos e renda para os jovens, evitando o êxodo e criando condições dignas de qualidade de vida para que os filhos da Amazônia não precisem abandonar seu lar.
07	Ações de prevenção e controle de doenças e morbidades fortemente frequentes na Amazônia. Faz-se necessário o desenvolvimento de ações conjuntas dos órgãos públicos estaduais, municipais e federais, em especial do Ministério da Saúde e da Fundação Nacional de Saúde – Funasa com campanhas específicas à gestão da saúde na região.
08	Execução de obras de infraestrutura devido à grande correlação entre as doenças mais comuns e os problemas de saneamento básico, como a baixa qualidade da água.
09	Incentivo ao desenvolvimento sustentável da Amazônia por meio do empreendedorismo consciente para desenvolvimento da região, com o envolvimento das comunidades locais. Percebeu-se um grande potencial para a indústria cosmética, farmacêutica e extrativista.
10	Implantação de programas para o uso e a disseminação de tecnologias sustentáveis para aumento da qualidade e da quantidade da produção agropecuária sem a necessidade do acréscimo de área ocupada, além da necessidade de promoção e ampliação da indústria sustentável da região.

Fonte: Elaborado pelo autor (2016).

CONCLUSÕES

Samuel Benchimol foi um amazônida apaixonado, um grande humanista, um professor cativante e um cientista brilhante. O acervo documental do Prêmio que leva o nome desse grande homem mostrou-se uma inestimável fonte de conhecimento. Ao analisar quantitativamente e qualitativamente as candidaturas ao Prêmio Professor Samuel Benchimol pode-se extrair um amplo conjunto relevante de informações relativas às questões-chave do desenvolvimento sustentável da região amazônica.

Entre os anos de 2004 e 2015, foram recebidas no Prêmio um total de 1856 (um mil oitocentas e cinquenta e seis) propostas inovadoras, uma média de cento e cinquenta e cinco candidaturas por ano e mais de cinquenta propostas para cada categoria, em média. Esses resultados tornam possível concluir que os objetivos do Prêmio têm sido alcançados, não apenas pelos números, mas pela relevância dos pesquisadores, instituições, empresas, institutos de pesquisa e universidades que recorrentemente participam da comenda.

A aplicação das técnicas, algoritmos bayesianos e software de mineração de textos, mostrou-se eficaz no processo de descoberta de conhecimento nessa grande base de dados textual e não estruturada de propostas de desenvolvimento da Amazônia. A plataforma Zoom e o software Tropes foram amplamente utilizados e apresentaram desempenho adequado para este tipo de pesquisa.

Um dos principais desafios encontrados durante a realização da pesquisa foi a limpeza dos dados. Devido ao grande volume de textos processados o software de mineração gerou um emaranhado de correlações irrelevantes que se misturavam às correlações semânticas relevantes, tornando a compreensão e leitura dos gráficos muito complexa, todavia, com a exclusão das *Stop-Words* e reprocessamento dos textos, pode-se avançar nas análises com um grau muito confiável de precisão.

O ranqueamento gerado pelo software de mineração de textos resultou em uma distribuição compatível com os princípios da Lei de Zipf, em que um conjunto limitado de palavras ocorre com grande frequência enquanto um amplo grupo de palavras ocorre poucas vezes. Pelas análises quantitativas também percebeu-se mantida a correlação da dispersão proposta por Bradford, segundo a qual um seleto grupo de

pesquisadores respondem por um elevado número de pesquisas na temática específica.

A utilização da análise de conteúdo, como ferramenta para organizar os processos de interpretação e inferência, provou-se acertada. Por meio dessa ferramenta as análises qualitativas dos dados foram desenvolvidas naturalmente, de forma a complementar as análises quantitativas, em especial, em complemento à bibliometria, tornando possível elencar, identificar e classificar as questões-chave do desenvolvimento da região amazônica.

O rigor no seguimento das etapas metodológicos foi importante para garantir a fidelidade e confiabilidade da pesquisa. A etapa mais trabalhosa foi a padronização dos formatos de arquivos, pois além do volume do corpus pesquisado, uma grande variedade, formatos e tipos de arquivos foi manipulada. A metodologia trouxe baixo grau de complexidade, todavia, grande esforço braçal.

Foram identificadas as questões-chave do desenvolvimento da Amazônia, sob a ótica do Prêmio, para as três naturezas cobertas pelo espectro dessa pesquisa (Ambiental, Econômico-tecnológica e Social). Para categoria “Ambiental” as principais questões encontradas estão correlacionadas à implantação de alternativas para a educação ambiental nas comunidades locais e na escola, a valorização do meio ambiente, a preservação e conservação do ecossistema e da biodiversidade e a recuperação das áreas degradadas, em especial das matas ciliares.

Na natureza “Econômico-tecnológica”, os pontos principais estão correlacionados ao desenvolvimento da Amazônia pelo empreendedorismo consciente, com o envolvimento das comunidades locais, às propostas para uso de tecnologias sustentáveis para aumento da qualidade e da quantidade da produção, além da necessidade de promoção e ampliação da indústria sustentável da região. Já para a natureza “Social”, as questões-chave consistem nas propostas de iniciativas para inclusão social, nos problemas de vulnerabilidade social das famílias, em especial as crianças e os idosos, nos projetos e nos programas para geração de empregos e renda para os jovens, nos projetos de prevenção e controle de doenças e morbidades frequentes na Amazônia; e nos problemas de saneamento básico e baixa qualidade da água.

Concluídas as análises pertencentes ao escopo deste, restou um amplo campo de estudos para pesquisas futuras. Centenas de outras relações semânticas foram identificadas automaticamente pelo software de mineração de textos, em diversas naturezas, tais como: cultural, científica, étnica, demográfica, política, local, regional, empresarial entre outras. Esses estudos serão importantes para construir uma visão mais ampla das necessidades e questões-chave para o desenvolvimento da Amazônia.

REFERÊNCIAS

ACHARD, F., EVA, H. D., STIBIG, H. J., MAYAUX, P., GALLEGO, J., RICHARDS, T., and MALINGREAU, J. P.: 2002, '**Determination of deforestation rates of the world's humid tropical forests'**, Science 297, 999–1002.

ALVARES, L.; FERREIRA, J.R.; LOBATO NETO, O.; GRAÇA, H.; ISRAEL, M. B. **Retrospectiva histórica das estratégias de desenvolvimento da Amazônia por meio do Prêmio professor Samuel Benchimol e do Prêmio Banco da Amazônia de Empreendedorismo Consciente**. In: FERREIRA, J.R; ALVARES, L. (Orgs). Prêmios Professor Samuel Benchimol e Banco da Amazônia de Empreendedorismo Consciente 2010. Manaus: Federação das Indústrias do Estado do Amazonas (Fieam)/Banco da Amazônia. 2010. 413p.

ALVARES, L.; FERREIRA, J.R.; **Prêmio Prof. Samuel Benchimol: análise quantitativa das propostas vencedoras.** IBICT. 2014.

AMADO, J. S. **A técnica de Análise de Conteúdo**. Revista Referência, 5, 53-63. 2000.

ANDER-EGG, E. **Análise de conteúdo: metodologia da ciência**. 3. ed. Rio de Janeiro: Kennedy. 1974.

ANDRADE, Maria Margarida de. **Introdução a Metodologia do Trabalho Cientifico**. 4 ed. São Paulo: Atlas, 1999.

ARAÚJO JÚNIOR, R. H.; TARAPANOFF, K. **Precisão no processo de busca e Recuperação da Informação:** uso da mineração de textos. Ciência da Informação, Brasília, v. 35, n. 3, p. 236-247, set./dez. 2006.

BABBIE, E. **The practice of social research**. New York: Macmillan. 1998.

BARDIN, L. **Análise de Conteúdo.** Lisboa: Edições 70, 1977.

BARION, E. C. N. Mineração de Textos. **Revista de Ciências Exatas e Tecnologia**. Vol. III, Nº. 3, Ano. São Paulo. 2008.

BATES, Marcia J. **How to use controlled vocabularies more effectively in online searching**. Online, v. 12, n. 6, p. 45-56, nov. 1988.

BECKER, Bertha K.. Geopolítica da Amazônia. Estud. av., São Paulo , v. 19, n. 53, p. 71-86, Apr. 2005. Disponível em: <http://www.scielo.br/scielo.php? script=sci_arttext&pid= S0103-40142005000100005&lng=en&nrm=iso>. acesso em: 2 Jun. 2016.

BELKIN, N. J.; ODDY, R. N.; BROOKS, H. M. **ASK for information retrieval**: part I. Background and theory. The JournalofDocumentation, v. 38, n. 2, jun. p. 61-71, 1982.

BELKIN, N. J; CROFT, W. B. **Retrieval tecniques**. Annual Review of Information Science and Tecnology, v. 22, p. 112-119, 1987.

BENCHIMOL, Samuel. **Zênite ecológico e nadir econômico-social: análisis e propostas para o desenvolvimento sustentável da Amazônia**. Editora 247 SA, 2001.

BENCHIMOL, Samuel. **A Amazônia e o terceiro milênio. Parcerias Estratégicas**, v. 5, n. 9, p. 22-34, 2010.

BEPPLER, M; FERNANDES, A**. Aplicação de text mining para a extração de conhecimento jurisprudencial.**in: primeiro congresso sul catarinense de educação, 2005.

BERRY, D. The Computational Turn: **Thinking About the Digital Humanities, Culture Machine**. 2011.

BERTALANFFY, V. **General System Theory.** Foundations, Development, Applications. New York, Braziller. 1968.

BERTALANFFY, V. **Outline of General System Theory**. British Journal for the Philosophy Science Vol. 1 No. 2 Agosto, 134-165. 1950.

BOLLIER, D. **'The promise and peril of Bigdata'**. 2010.

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

BOYD, D. **Critical Questions for Bigdata:** Provocations for a Cultural, Technological, and Scholarly Phenomenon danahboyd. Microsoft Research. 2011.

BRASIL. **Escola.** Disponível em http://brasilescola.uol.com.br/geografia/eco-92.htm. Acesso em março de 2016.

BRASIL. Secretaria de Assuntos Estratégicos. **Seminário de Segurança da Amazônia**. Período de 11 a 15 de agosto de 2010, Manaus-AM: 2010. Disponível em: <http://www.sae.gov.br/seminarioamazonia/>. Acesso em 25 maio 2016.

BRYNJOLFSSON, E. **The Bigdata Boom is the Innovation Story of Our Time.**The Atlantic. Novembro. 2011.

BURKE, M.A. **Organization of multimedia resources:** principle and practice of information retrieval. Aldershot: Gower, 2005.

BUSH, V. **As we may think**. .4tlantIc Monthly, 176, (1), p. 101-108. 1945.

CAMARGO, Flávio Fortes et. al. **Avaliação dos métodos booleano, fuzzy gama e bayesiano na identificação de áreas susceptíveis a movimentos de massa no município de São Sebastião.** São José dos Campos: INPE, 2007.

CAMPOS, C. J. G. **Método de Análise de Conteúdo:** ferramenta para a análise de dados qualitativos no campo da saúde**.** Rev. bras. enferm. [online]. 2004

CAPURRO, R.; **Epistemologia e Ciência da Informação**. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIAS DA INFORMAÇÃO, 8. 2003, Belo Horizonte. Anais. Belo Horizonte: UFMG, 2003.

CAPURRO, R.; HJORLAND, B. **O conceito de informação. Perspectivas em Ciência da Informação**, Belo Horizonte, v.12, n.1, p.148-207, abr. 2007.

CARVALHO JUNIOR, B. H. M.; SANTOS, P. E. A.; ROMAGNA, W. **Sistemas distríbuidos: o Bigdata**. São Paulo. 2013.

CATARINO, M. E.; BAPTISTA, A. A**. Folksonomia: um novo conceito para a Organização dos recursos digitais na Web.**DataGramaZero - Revista de Ciência da Informação, Rio de Janeiro, v. 8, n. 3, jun. 2007.

CERVO, Amado Luiz; BERVIAN, Pedro Alcino. **Metodologia científica:** para uso de estudantes universitários**.** 3. ed. São Paulo: McGraw-Hill do Brasil, 1983.

CHIZZOTTI, **A. Pesquisa em ciências humanas e sociais**, (8a ed.). São Paulo: Cortez. (2006).

CHOO, C. W. **A organização do conhecimento:** como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. São Paulo: Senac, 2003.

CONWAY, M.T. **The subjective precision of computers: A methodological comparison with human coding. Journalism& Mass Communication Quarterly**, 83(1), p. 186–200. 2006.

CORRÊA, A. C. G. **Recuperação de documentos baseada em Informação Semântica no Ambiente** AMMO. UFSCAR, 2003.

CRAWFORD, K**. following you: disciplines of listening in social media', continuum: journal of media & cultural Studies.**Vol 23. 2009.

CUNHA, Murilo Bastos. **Análise de Conteúdo, uma técnica de pesquisa.** Revista de Biblioteconomia de Brasília, v.11, n.2, p.247-256, jul./dez. 1983.

DACONTA, M.C; OBRST, L.J; SMITH, K.T**. The Semantic Web: a guide to the future of XML, Web services, and knowledge management**. Indianapolis: Wiley, 2003.

DATTA, Suman. **A organização de conceitos para Recuperação da Informação**. Ciência da Informação, Rio de Janeiro, v. 6, n. 1, p. 17-28, 1977.

DeFRIES, R. S., HOUGHTON, R. A., HANSEN, M. C., FIELD, C. B., SKOLE, D., and TOWNSHEND, J.: 2002, '**Carbon emissions from tropical deforestation and regrowth based on satellite observations for the 1980s and 1990s'**, PNAS 99, 14256–14261.

DEMO, P. **Pesquisa e construção do conhecimento: Metodologia Científica no caminho de Habermas.** Rio de Janeiro: Tempo Brasileiro, 1994.

DOWNE-Wamboldt, B. **Content analysis: method, applications, and issues.** health care for women international, 13, 313-321. 1992.

FEIJÓ, Bruno Vieira. **A Revolução dos Dados.** Revista Exame PME – Pequenas e Médias Empresas, São Paulo, p. 30-43, set. 2013.

FERNANDES, O. S. L. **Conheça o Prêmio Professor Samuel Benchimol.** out. 2011. Disponível em http://www.amazonia.desenvolvimento.gov.br/conheca/index/item/8. Acesso em: 02 out. 2011.

FERNEDA, Edberto. **Recuperação de informação: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. 2003. 137 f. Tese (Doutorado em Ciências da Comunicação) - Universidade de São Paulo, São Paulo, 2003.

FERREIRA, Berta Weil. **Análise de Conteúdo**. Aletheia, n. 11, p.13-20, jan./jun. 2000.

FIGUEIREDO, Dayana Ester. **Recuperação da Informação: uma análise sobre os sistemas de busca na Web.** 2006. 61 f. Monografia (Graduação em Biblioteconomia) - Universidade de Brasília, Brasília, 2006.

FIGUEIREDO, Nice. **Tópicos modernos em bibliometria.** Brasília: Associação dos Bibliotecários do Distrito Federal, 1977.

GARFIELD, E. **Information Retrieval.** Science. Jun. 1967.

GIL, A. C. **Métodos e técnicas da pesquisa social**. 5. Ed. São Paulo. Atlas. 1999.

GOMES, Helder Joaquim Carvalheira. **Text mining: análise de sentimentos na classificação de notícias**. Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on. Lisboa. 2013.

GRANEHEIN, U.H.; LUNDMAN, B. **Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness.**Nurse EducationToday, 24, 105-112. 2003.

GRINSTEAD, C, M; SNELL, J. L., **The Grinstead and Snell's Introduction to probability, parade magazine**, v. 3, mar. 1996.

GUEDES, V.; BORSCHIVER, S.; **Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica.** Salvador: ici/ufba, 2005.

HENRIQUES, D.A.; COSTA, H, R. **Bigdata – Como Utilizar a Extraordinária Quantidade de Informações Coletadas por Novas Tecnologias para Obter Vantagens**

Competitivas. Disponível em: <http://revistapensar.com.br/tecnologia/pasta_upload/artigos/ a72.pdf> Acesso em: 01 set. 2014.

HOUGHTON, R.A. 2005. **Tropical deforestation as a source of greenhouse gases. In "Tropical Deforestation and Climate Change**" Edited by P. Moutinho and S. Schwartzman. Instituto de Pesquisa Ambiental da Amazônia (IPAM) e Environmental Defense (ED).

HSIEH, Hsiu-Fang; SHANNON, S. E..**Three Approaches to Qualitative Content Analysis**. Qualitative Health Research, Vol. 15 No. 9, 1277-1288. 2005.

IANNI, Octavio. **A metáfora da viagem**. Cultura Vozes: n. 2 (mar/abr), 1994.

INGWERSEN, P. **Information retrieval interaction.**London: Taylor Graham, 1992.

IPCC 2000. **Intergovernmental Panel on Climate Change. Land Use, Land-Use Change and Forestry**. Cambridge University Press.

IPCC, 2007: Summary for Policymakers. In: **Climate Change 2007**: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [B. Metz, O.R. Davidson, P.R. Bosch, R. Dave, L.A. Meyer (eds)], Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

ISACA, **Bigdata – Impactos e Benefícios**. 2013. Disponível em: http://www.isaca.org/Knowledge-Center/Research/Documents/BigData_whp_Por_0413.pdf. Acesso em: mar. 2015.

ISACA, **Privacidade and Bigdata.** disponível em: http://www.isaca.org/knowledge-center/research/documents/privacy-and-bigdata_whp_portuguese_0913.pdf. Acesso em mar. 2015.

JAPIASSU, H. **Interdisciplinaridade e patologia do saber**. Rio de Janeiro: Imago, 1976.

JAPIASSU, H. **Introdução ao pensamento epistemológico**. 2 ed. Rio de Janeiro, Francisco Alves, 1977, 202p.

JAPIASSU, Hilton**. Introdução às Ciências Humanas: análise de epistemológica histórica.** São Paulo: Editora Letras& Letras, 1994. 191 p.

KERLINGER, F.N. **Foundations of behavioral research**. New York: Holt, Rinehart & Winston, 1973.

KOTHARI, C. R. **Research methodology: methods and techniques, New age International**.PvtLtdpublishers. New Delhi. 2004.

LANCASTER, F. Wilfrid. **Indexação e resumos: teoria e prática**. 2. ed. Brasília: Briquet De Lemos, 2004. xviii, 452 p. ISBN 8585637242.

LARA, M. L. G. **Informação, informatividade e lingüística documentária: alguns paralelos com reflexões de Hjorlando e Capurro.**DataGramaZero - Revista de Ciência da Informação, Rio de Janeiro, v. 9, n. 6, dez. 2008.

LASMAR, Dimas José; SOUZA, Euler Guimarães Menezes de.; ARAÚJO FILHO, Guajarino de.;. **Relatório sobre os Determinantes do Sistema Local de Inovação de Manaus**, Brasil [2010]. Disponível em: <http://www.iit-berlin.de/ANIS_ Manaus_ Portuguese.pdf>. Acesso em 15 maio 2016.

LASSWELL, Harold D. **Estructura y function de la comunicaciónen la sociedad**. In: MORAGAS SPÁ, Miquel. Sociologia de lacomunicación de masas. Barcelona : G. Gilli, 1985.

LAVILLE, Christian; DIONNE, Jean. **A construção do saber. Manual de metodologia da pesquisa em ciências humanas.** Trad. Heloísa Monteiro e Francisco Settineri. Porto Alegre: Artes Médicas Sul Ltda; Belo Horizonte: Editora UFMG, 1999.

LIMA, J. L. O.; ALVARES, L. **Orgnização e representação da informação e do conhecimento: Organização da informação e do conhecimento: conceitos, subsídios interdsiciplinares e aplicações**. São Paulo.2012.

LOPES, I. L. **Estratégia de busca na Recuperação da Informação: revisão da literatura**. Ciência da Informação, Brasília, v. 31, n.2, p. 60-71, maio/ago. 2002.

MANOVICH, L. **The Promises and the Challenges of Big Social**. Abril, 2011.

MATTOS. Luis Carlos Gomes. **O Exército Brasileiro na Defesa da Soberania da Amazônia**, 2012.

MINAYO, Maria Cecília de Souza. **O desafio do conhecimento: pesquisa qualitativa em saúde**. 4.ed. São Paulo. HUCITEC-ABRASCO, 1996.

MOOERS, C. N. **The theory of digital handling of non - numerical information and its imp lications to machine economics, in Association for Computing Machinery Conference**, Rutger University, 1950.

MORAES, R. **Análise de Conteúdo**. Educação, Porto Alegre, v.22, n.37, p.7-32, 1999.

MORAIS, E. A.; AMBRÓSIO, A. P. L. **mineração de textos**. Relatório Técnico. Instituto de Informática; Universidade Federal de Goiás, 2007.

MORGAN, D. L. **Qualitative content analysis: A guide to paths not taken.**Qualitative Health Research, 3, p. 112-121. 1993.

MOURA, M. F. **Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos.** Embrapa Informática Agropecuária, 2004, ISSN 1677-9274, 2004.

MOURA, Patrícia Garcia de; BATISTA, Luciana Rodrigues Vieira e MOREIRA, Emilia Addison Machado. **População indígena: uma reflexão sobre a influência da civilização urbana no estado nutricional e na saúde bucal**. Rev. Nutr. [online]. 2010, vol.23, n.3, pp.459-465. ISSN 1415-5273.

MOUTINHO, Paulo. Desmatamento na Amazônia: desafios para reduzir as emissões de gases de efeito estufa do Brasil. Disponível em: **www.ipam.org.br/biblioteca**, acesso em: 15 maio 2016.

MOUTINHO, P. and S. SCHWARTZMAN (2005). **Tropical Deforestation and Climate Change.** Instituto de Pesquisa Ambiental da Amazônia (IPAM) and Environmental Defense (ED).

NANDY, B. R., SARVELA, P. D. **Content analysis reexamined: A relevant research method for health education.**American Journalof Health Behavior, 21, p. 222-234. 1997.

NAVARRO, E. M. A. **Los lenguajesdocumentales ante elpaso de laorganización de larealidad y el saber a laorganizacióndelconocimiento**. Scire, Zaragoza, v. 1, n. 2, 1995.

NAZARETH, Tayana; BRASIL, Marília; TEIXEIRA, Pery. Manaus: **Crescimento Populacional E Migrações Nos Anos 1990**. VII Encontro Nacional Sobre Migrações de Tema Central: Migrações, Políticas Públicas e Desigualdades Regionais, realização de 10 a 12 de Outubro de 2011, Curitiba/PR.

NICHOLAS, David; RITCHIE, **Maureen. Literature and bibliometrics**. London: Clive Bingley, 1978.

O CICLO DA BORRACHA. Disponível em: **<www.suapesquisa.com/historiadobrasil /ciclo_borracha.htm** >. Acesso em 20 Maio 2016.

OLABUENAGA, J.I. R., ISPIZUA, M.A. **La descodificacion de la vida cotidiana: metodos de investigacioncualitativa**. Bilbao, Universidad de deusto, 1989.

OLIVEIRA, A. **Data Science and Data Analytics**. 2013. Palestra apresentada no 1o. EMC Summer SchoolonBigdata. EMC/NCE/UFRJ. Rio de Janeiro. 2013.

OLIVEIRA, D. de P. R. **Sistemas de informação gerenciais: estratégias, táticas, operacionais**. 8. ed., São Paulo: Atlas, 1992.

OLIVEIRA, R. R., DE CARVALHO, C. L.. **Implementação de interoperabilidade entre repositórios digitais por meio do protocolo OAI-PMH**. Goiânia: UFG, 2009.

OLIVEIRA, Márcia Maria. **A mobilidade humana na tríplice fronteira: Peru, Brasil e Colômbia**. Estud. av. vol. 20, nº. 57. São Paulo May/Aug. 2006. Disponível em: <www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40142006000200014>. Acesso em 10 maio 2016.

ORTEGA, C.D. **Informática Documentária: estado da arte.** 2002. 234 f. Dissertação (Mestrado em Ciências da Comunicação) - Escola de Comunicação e Artes, Universidade de São Paulo. 2002.

PARISER, E. **The Filter Bubble: What the Internet is Hiding from You.**Penguin Press, New York, NY. 2011.

PAUTOSO, A. **Metodologia para revisão crítica de literatura**. 2013.

PAYNE, G. and J. PAYNE. **Key Concepts in Social Research**. London: Sage. 2004.

PICHILIANI, Mauro. **Data mining na prática:** regras de associação. 2008.

PIEDADE, M. A. R., **Introdução à teoria da classificação**. Rio de Janeiro: Interciência, 1977.

PIGNATARI, D. **Informação. Linguagem. Comunicação.** São Paulo: Perspectiva, 1993.

PINHEIRO, L. V. R. A **Ciência da Informação no Brasil: historiografia de uma área do conhecimento contemporânea no cenário nacional.** Projeto de pesquisa. Rio de Janeiro: IBICT, ago. 2002 –fev. 2006.

POMBO, O. **Da classificação dos seres à classificação dos saberes**. Lisboa, 1998.

RANGANATHAN, S.R. **Prolegomena to Library Classification.** Bombay: Asia Publishing House, 1967.

RENNIE, J. D. et al. **Tackling the poor assumptions of naive bayes text classifiers**. In:ICML. 2003. p. 616-623.

RIJSBERGEN, V. C. J., **INFORMATION RETRIEVAL**. Melvin J. Voigt, p. 91-118. 1979.

ROBERTSON, S.E; JONES, K.S. **Relevance weighting of earch terms**. Journal of the Americam Society for Information Science, v. 27, n. 3, p.129-146, 1976.

RODRIGUES, B. C.; CRIPPA, G.; **A Recuperação da Informação e o conceito de informação: o que é relevante em mediação cultural?** Perspectivas em Ciência da Informação, Belo Horizonte, v. 16, n. 1, p. 45-64, jan./mar. 2011.

ROSENGREN, K. E**. Advances in Scandinavia content analysis**: An introduction. In K. E. Rosengren (Ed.), Advances in content analysis (pp. 9-19). Beverly Hills, CA: Sage. 1981.

ROSSI, G. B.; SERRALVO, F. A.; JOÃO, B. N. **Análise de Conteúdo**. Revista Brasileira de Marketing, v. 13, n. 4, p. 39-48, 2014.

SALATI, E., and R. VOSE (1984), **Amazon basin**: A system in equilibrium. Science, 225, 129-138.

SALTON, G. **The SMART retrieval system: experiments in automatic document processing**. Nova Jersey: Prentice-Hall, 1971.

SALTON, G., FOX, E.A., WU, H. **Extended Boolean Information Retrieval**. Communication ofthe ACM, v. 26, n. 11, p.1022-1036.1983.

SAMUEL ISAAC BENCHIMOL – BIOGRAFIA. Disponível em: < **www.benchimol.com.br /html/biografia_samuel.htm** >. Acesso em 20 maio 2016.

SANTILLI, M., P. MOUTINHO, S. SCHWARTZMAN, D. NEPSTAD, L. CURRAN, C. NOBRE. 2005. **Tropical deforestation and the Kyoto Protocol: an editorial essay.** Climate Change 71: 267-276.

SANTOS, Breno Santana et al. Comparing Text Mining Algorithms for Predicting Irregularities in Public Accounts. **In: Proceedings of the annual conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective-Volume 1.** Brazilian Computer Society, 2015. p. 89.

SANTOS, M. A. M. R. **Extraindo Regras de Associação a partir de Textos**. PUC. 2002.

SARACEVIC, T. **Interdisciplinary nature of information science**. Ciência da Informação, v.24, n.1, p.36-41, 1995.

SHAW, I.S.; SIMÕES, M.G. **Controle e modelagem fuzzy**. São Paulo: Edgard Blücher, 1999.

SOARES-FILHO, B., D. NEPSTAD, L. CURRAN, et al. **Modeling Amazon conservation**. 2006. Nature 440:520-523.

TAGUE-SUTCLIFFE, Jean. **An introduction to informetrics. Information processing & management**, Oxford, v. 28, n. 1, p. 1-3, 1992.

TORRACO, R. J. **Writing integrative literature reviews: Guidelines and examples**. HumanResourceDevelopmentReview, (4), 356. 2005.

TRAQUINA, Nelson. **Teorias do Jornalismo - Porque as notícias são como são.** 2. ed. Florianópolis: Insular, 2005.

URBIZAGASTEGUI ALVARADO, R.; **Produtividade dos autores na literatura de enfermagem um modelo de aplicação da lei de lotka**. São Paulo, 2002.

VANTI, N. A. P. **Da bibliometria à Webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento.** Ciência da Informação, Brasília, v. 31, n. 2, p. 152-162, maio/ago. 2002.

VIEIRA, M. R.; FIGUEIREDO, J. M.; LIBERATTI G.; VIEBRANTZ, A. F. M. **Bancos de Dados NoSQL: Conceitos, Ferramentas, Linguagens e Estudos de Casos no Contexto de Bigdata**. Simpósio Brasileiro de Bancos de Dados – SBBD. 2012.

VILEJA JUNIOR, G. B.; **Análise de Conteúdo.** Centro de Pesquisas Avançadas em Qualidade de Vida, disponível em http://www.cpaqv.org/epistemologia/analiseconteudo.pdf. Acessado em fev. 2015.

WHITE, M. D.; MARSH, E. E. **Content analysis: a flexible methodology.**Library Trends, v. 55, n.1, p.22-45, Sum. 2006.

WIVES, L. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva.** Exame de Qualificação EQ-069, PPGC-UFRGS, 2002.

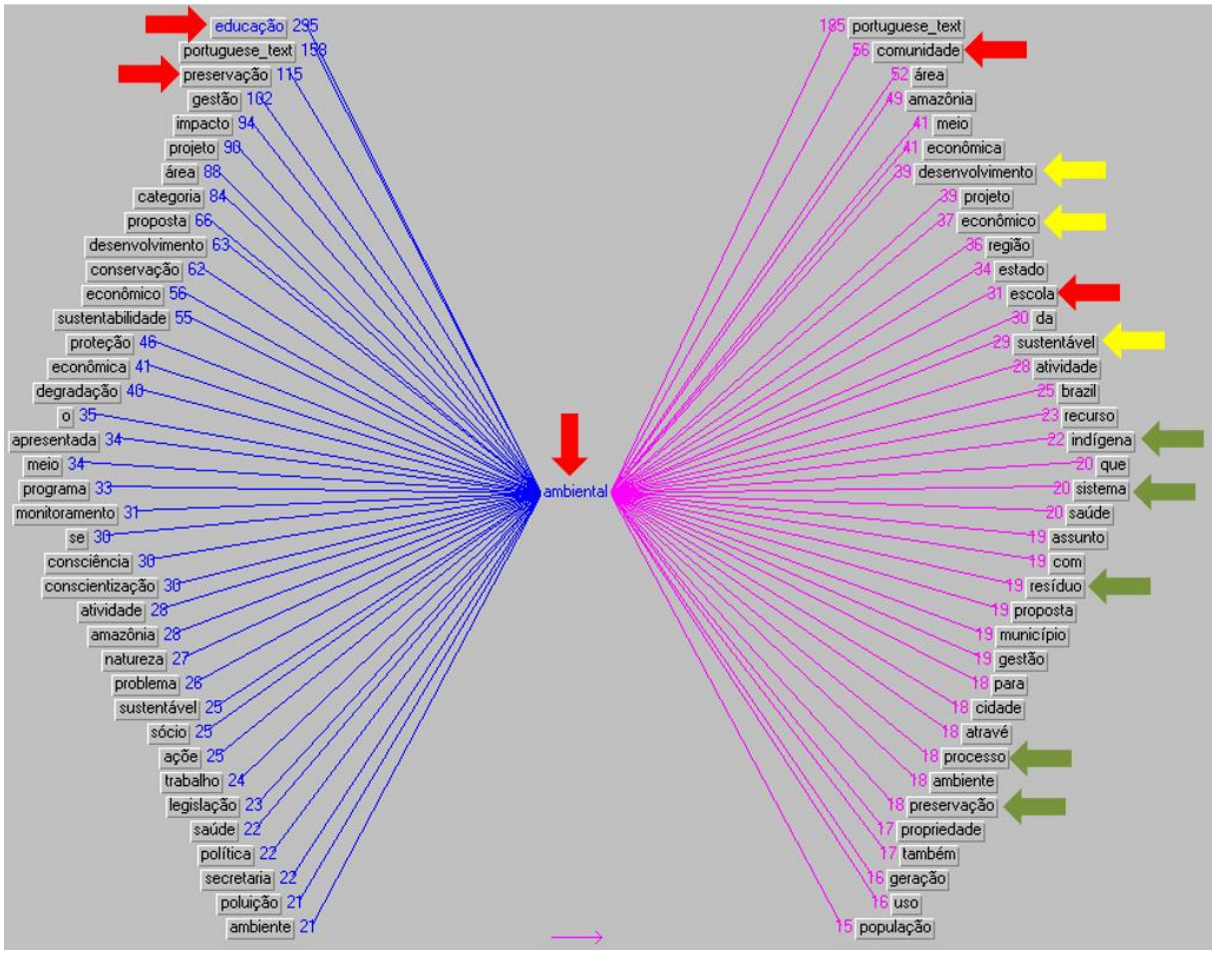
WOLF, Mauro. **Teorias da Comunicação**. Presença: Lisboa. 1985.

ZOETER, Onnoet al**. Information retrieval system**. U.S. Patent n. 8,037,043, 11 out. 2011.

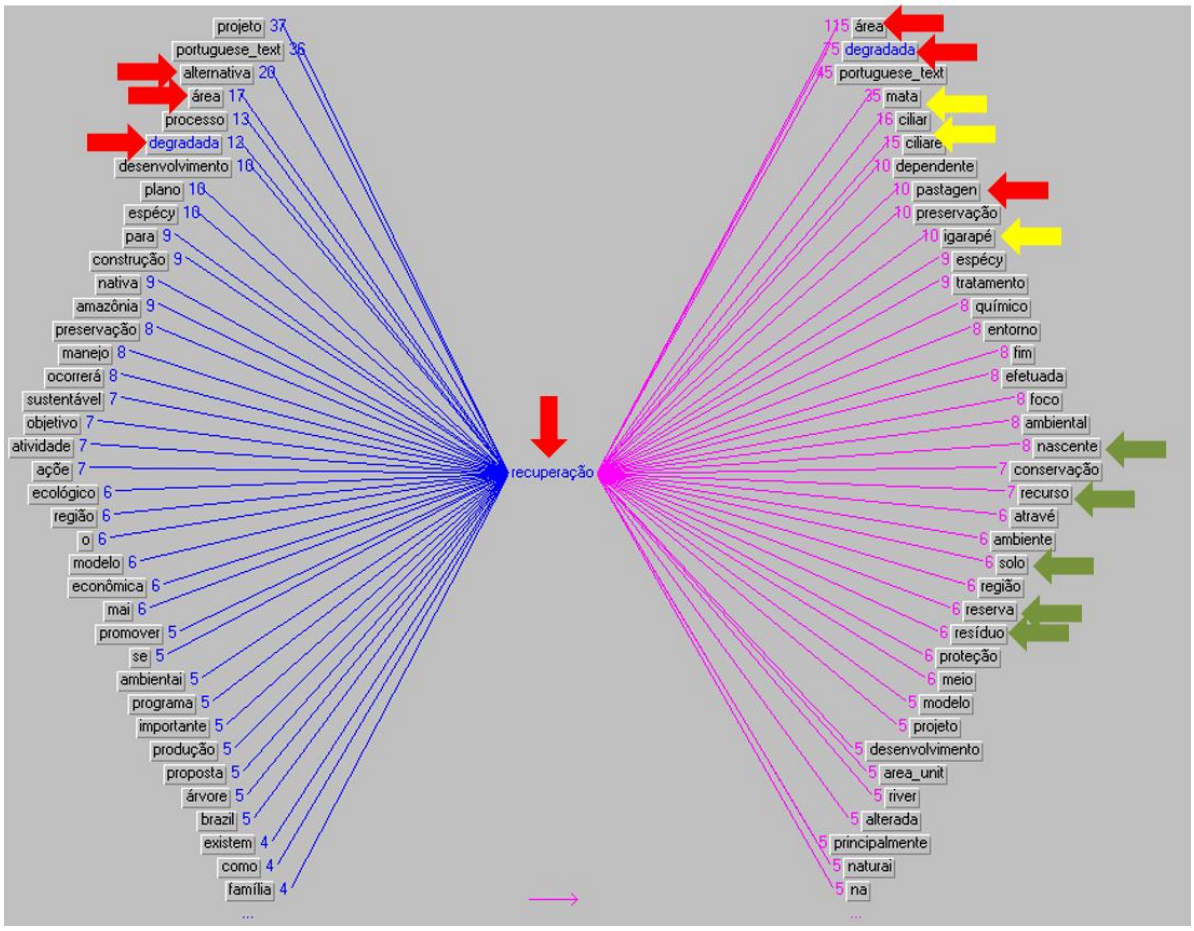
APÊNDICE I – Lista De Stopwords Utilizadas

a, agora, ainda, alguém, algum, alguma, algumas, alguns, ampla, amplas, amplo, amplos, ante, antes, ao, aos, após, aquela, aquelas, aquele, aqueles, aquilo, as, até, através, cada, coisa, coisas, com, como, contra, contudo, da, daquele, daqueles, das, de, dela, delas, dele, deles, depois, dessa, dessas, desse, desses, desta, destas, deste, deste, destes, deve, devem, devendo, dever, deverá, deverão, deveria, deveriam, devia, deviam, disse, disso, disto, dito, diz, dizem, do, dos, e, é, ela, elas, ele, eles, em, enquanto, entre, era, essa, essas, esse, esses, esta, está, estamos, estão, estas, estava, estavam, estávamos, este, estes, estou, eu, fazendo, fazer, feita, feitas, feito, feitos, foi, for, foram, fosse, fossem, grande, grandes, há, isso, isto, já, la, lá, lhe, lhes, lo, mas, me, mesma, mesmas, mesmo, mesmos, meu, meus, minha, minhas, muita, muitas, muito, muitos, na, não, nas, nem, nenhum, nessa, nessas, nesta, nestas, ninguém, no, nos, nós, nossa, nossas, nosso, nossos, num, numa, nunca, o, os, ou, outra, outras, outro, outros, para, pela, pelas, pelo, pelos, pequena, pequenas, pequeno, pequenos, per, perante, pode, pude, podendo, poder, poderia, poderiam, podia, podiam, pois, por, porém, porque, posso, pouca, poucas, pouco, poucos, primeiro, primeiros, própria, próprias, próprio, próprios, quais, qual, quando, quanto, quantos, que, quem, são, se, seja, sejam, sem, sempre, sendo, será, serão, seu, seus, si, sido, só, sob, sobre, sua, suas, talvez, também, tampouco, te, tem, tendo, tenha, ter, teu, teus, ti, tido, tinha, tinham, toda, todas, todavia, todo, todos, tu, tua, tuas, tudo, última, últimas, último, últimos, um, uma, umas, uns, vendo, ver, vez, vindo, vir, vos, vós

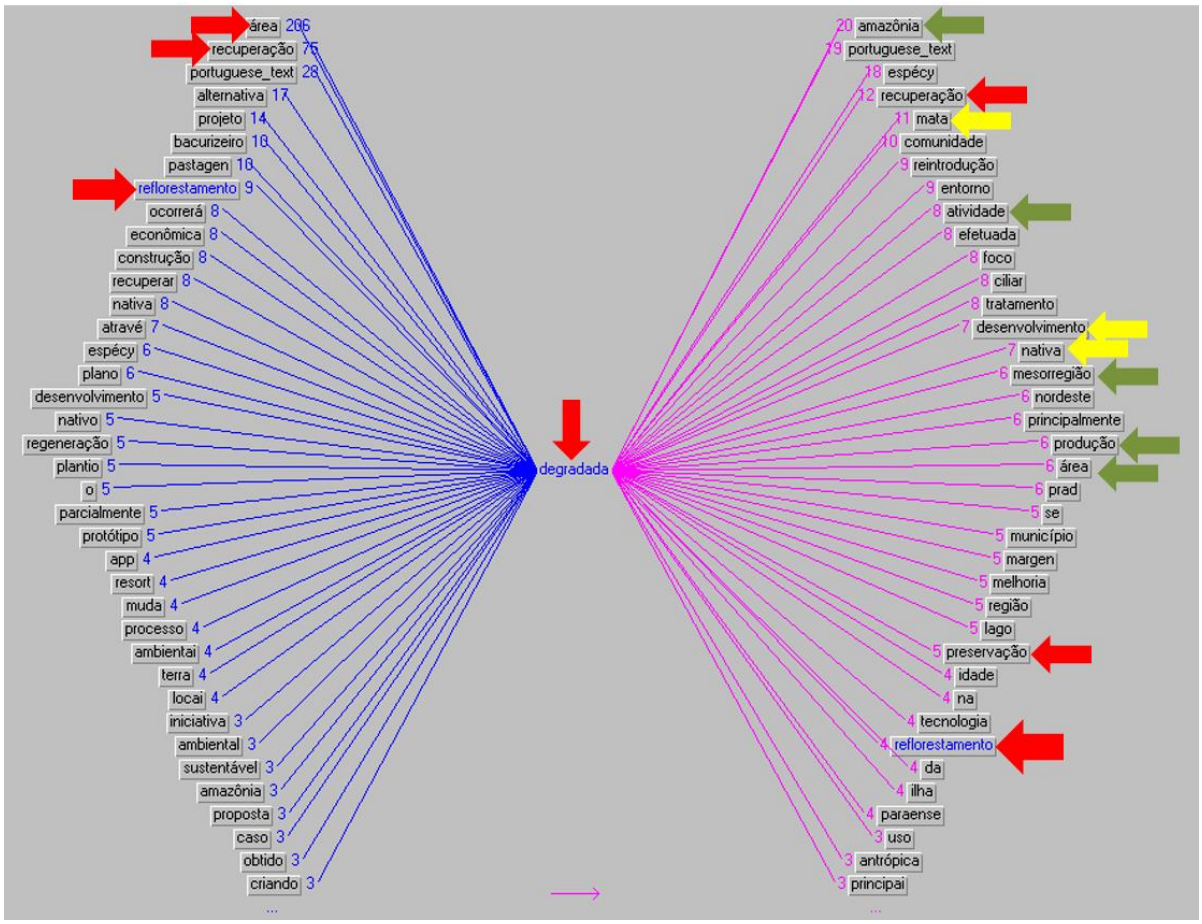
APÊNDICE II - Gráfico das correlações semânticas da categoria “ambiental”



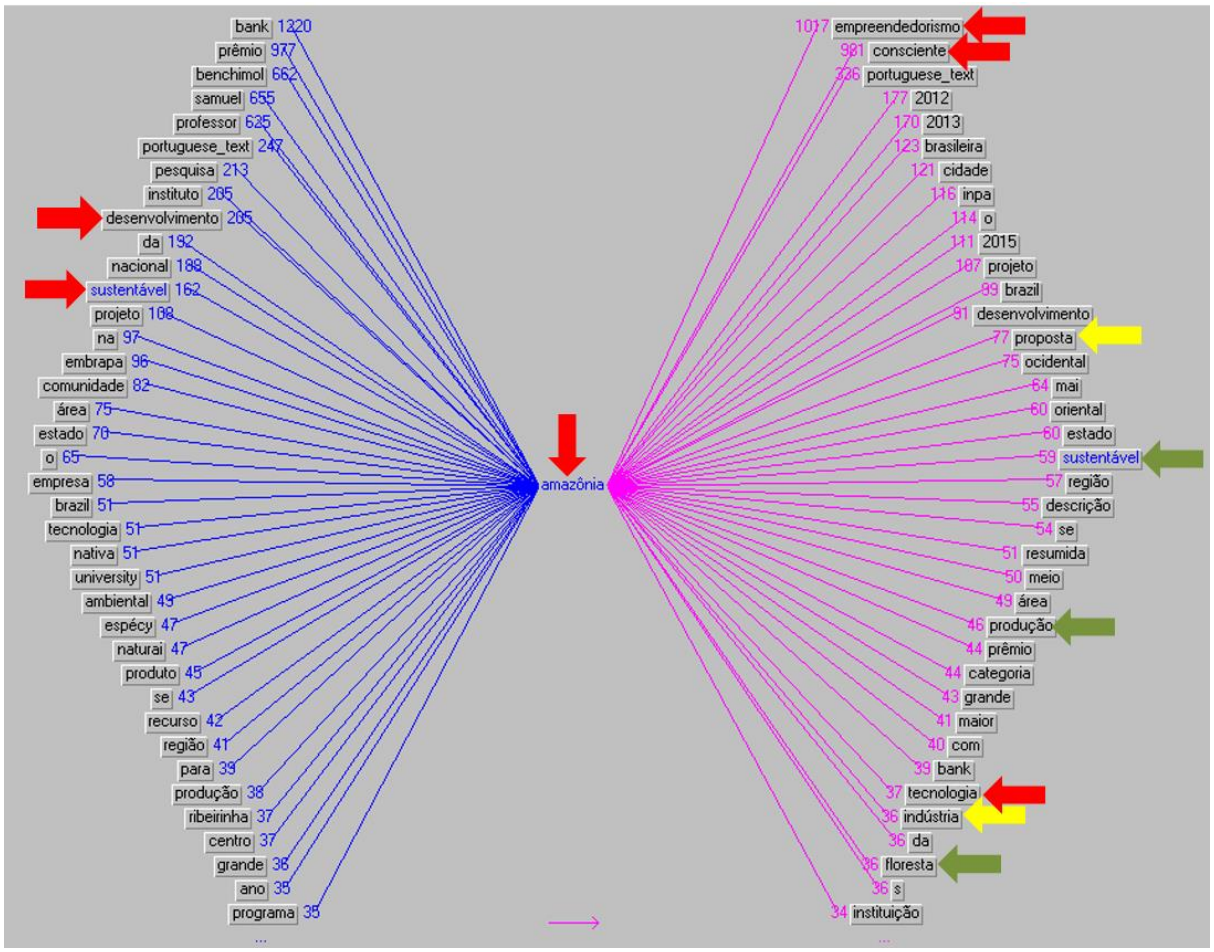
APÊNDICE III - Gráfico das correlações semânticas da categoria “recuperação”



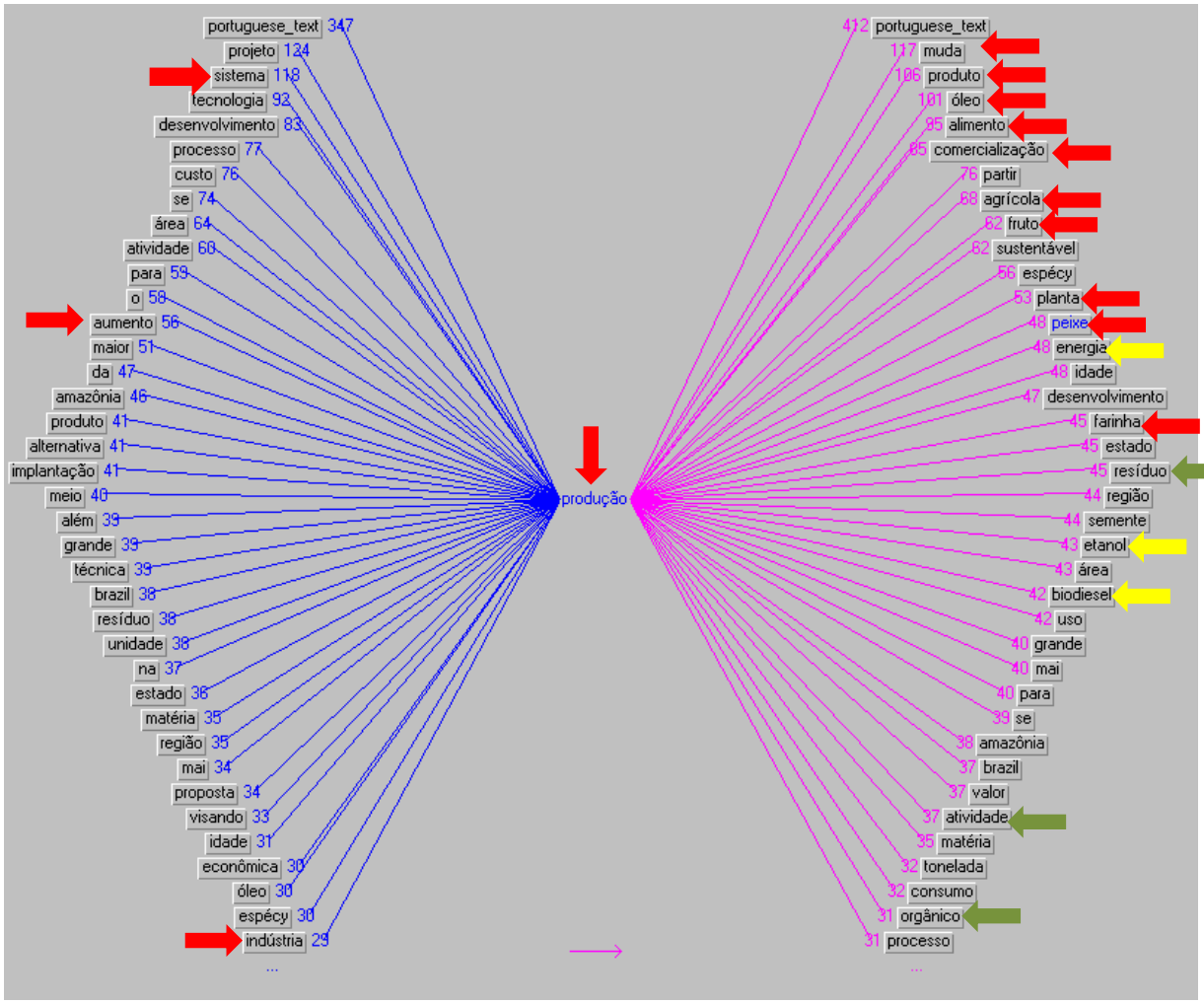
APÊNDICE IV – Gráfico das correlações semânticas da classe “degradada”



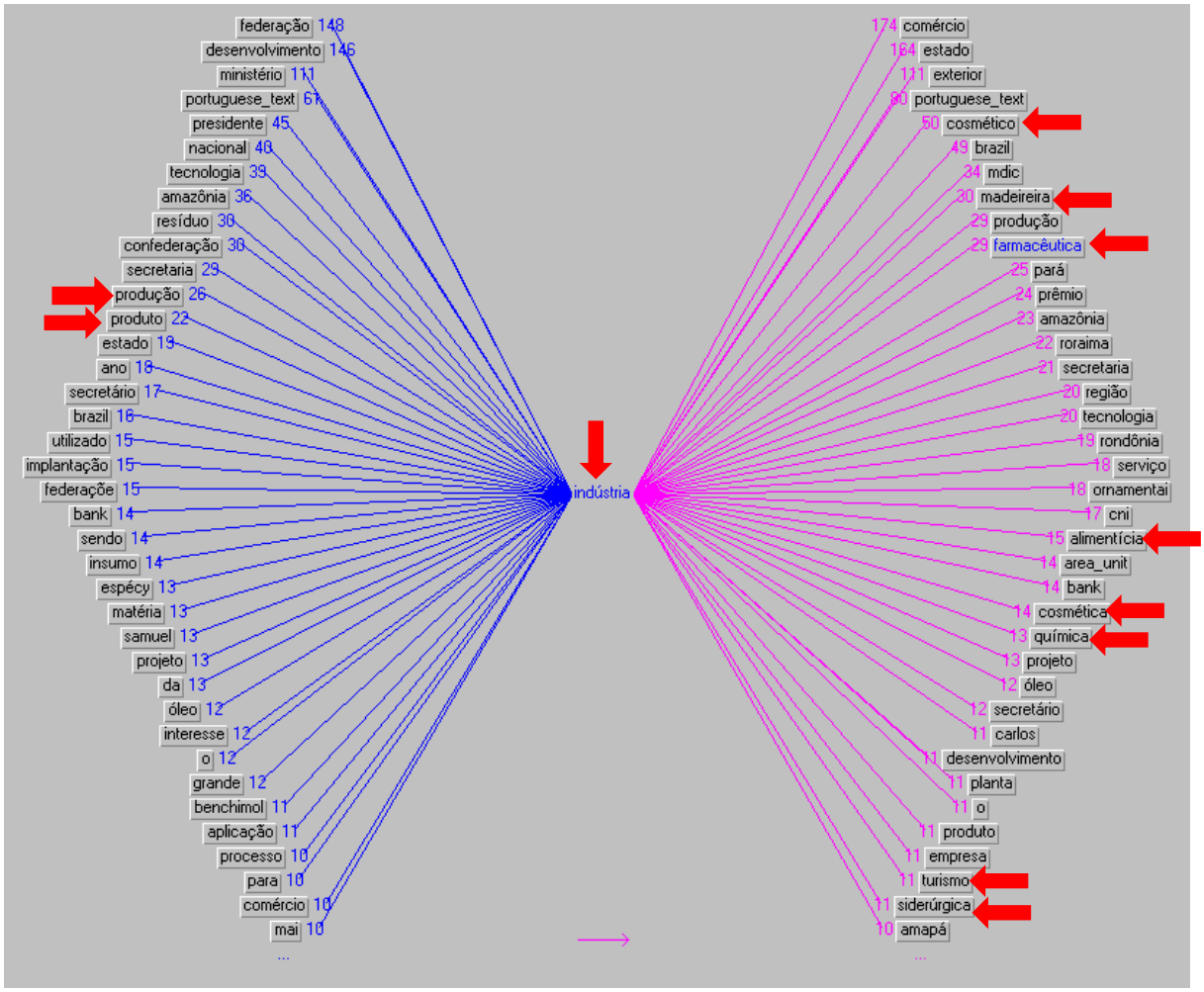
APÊNDICE V – Gráfico das correlações semânticas da classe “amazônia”



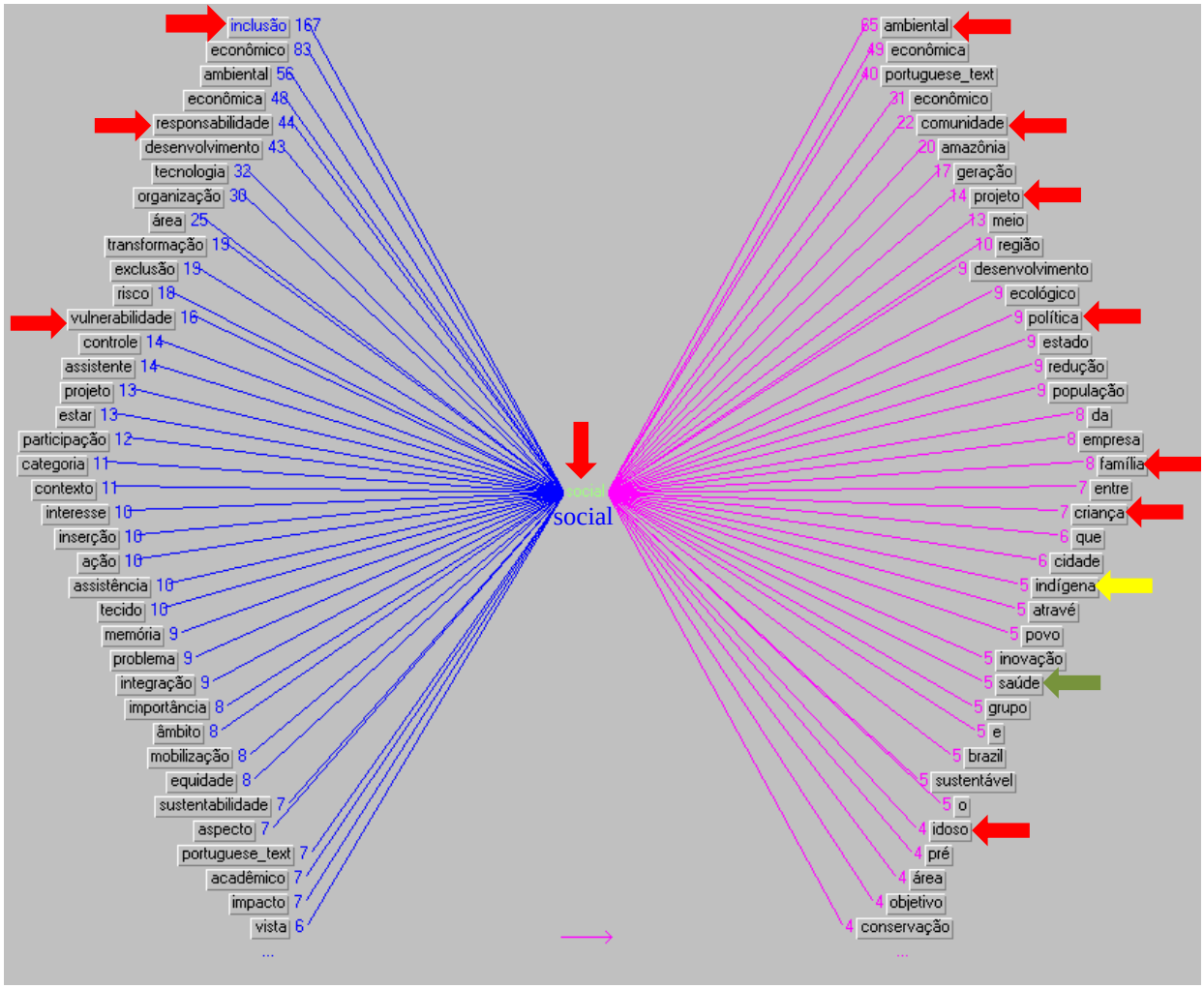
APÊNDICE VI – Gráfico das correlações semânticas em esferas da classe “produção”



APÊNDICE VII - Gráfico das correlações semânticas da classe “indústria”



APÊNDICE VIII – Gráfico das correlações semânticas da classe “social”



APÊNDICE IX – Gráfico das correlações semânticas da classe “doença”