UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS SOCIAIS APLICADAS
CURSO DE GESTÃO DA INFORMAÇÃO

CARLOS ALEXANDRE LOURENÇO TABORDA

MINERAÇÃO DE DADOS E A ANÁLISE DE MÉTRICAS E INDICADORES: UM ESTUDO SOBRE METRIAS INFORMACIONAIS

CURITIBA
2011

CARLOS ALEXANDRE LOURENÇO TABORDA

MINERAÇÃO DE DADOS E A ANÁLISE DE MÉTRICAS E INDICADORES: UM ESTUDO SOBRE METRIAS INFORMACIONAIS

Trabalho de conclusão de curso apresentado como critério de avaliação à disciplina de Pesquisa em Informação, do Curso de Gestão da Informação, do Departamento de Ciência e Gestão da Informação, do Setor de Ciências Sociais Aplicadas da Universidade Federal do Paraná

Orientadora: Profa. Dra. Denise Fukumi Tsunoda

CURITIBA
2011

FOLHA DE APROVAÇÃO

AGRADECIMENTOS

"Os homens são bons de um modo só e, maus de muitos modos."
Aristóteles

Com base nessa célebre frase, vale agradecer aqueles que colocaram obstáculos em meu caminho, pois, conforme o filósofo, estes fazem parte da segunda classe de pessoas que proporcionam nosso crescimento. Não se faz necessário a menção aos nomes, mas, fica registrado o meu agradecimento.

Os que têm mais valia merecem o maior espaço nesta seção, sendo estas pessoas que estão na primeira classe indicada pelo Aristóteles.

Aos meus pais pela educação, pelos valores e por todo o conhecimento empírico repassado que, com certeza, me ensinaram a nunca desistir dos meus ideais; a minha irmã pela sinceridade e por sempre torcer e festejar minhas conquistas; ao meu irmão por sua colaboração de forma direta ou indireta, a minha linda e amada avó, por suas orações diárias para que eu seja feliz. Ao meu Amor, por sua compreensão em minhas ausências, por sua atenção e zelo com minha saúde e por meu sono, por seu carinho, dedicação e infinito amor. A Você, o meu eterno amor.

Um agradecimento especial a minha orientadora, Profa. Dra. Denise Fukumi Tsunoda, pela sua incansável paciência, compreensão, sabedoria e por me aceitar como orientando em um momento tão incomum. Saiba que suas palavras, ensinamentos e conselhos serão seguidos por toda a minha vida. Obrigado por seu exemplo de vida.

Seria injusto se não agradecesse aos excelentes professores: Profa. Lígia Kraemer, pois seus ensinamentos me fizeram um especialista em análise de requisitos e elaboração de fluxos de informação; Prof. Mauro José Belli e ao Prof. Marcos Tedeschi, por sempre me apoiar, entender e me aconselhar em questões pessoais e profissionais e, aos demais Professores que me ensinaram a ser o Profissional da Informação que sou hoje. A vocês, dignos mestres o meu muito obrigado.

Aos colegas e amigos (Elisa, Cibele, Fran, Karol e Loru) que fizeram parte desta jornada, compartilho a alegria desta conquista e agradeço pelo apoio.

“Alguns identificam a felicidade como virtude;
outros como a sabedoria prática;
Outros como uma espécie de sabedoria
filosófica;
Outros como uma coisa associada ao prazer,
Mas para mim, é simplesmente: viver”

Adaptado de Aristóteles

RESUMO

Discute as métricas e os indicadores disponíveis no *Google Analytics* aplicáveis a plataforma de gestão editorial de periódicos científicos *Open Journal System* (OJS) com vista a prática de mineração de dados. Baseado em uma revisão da literatura pertinente, realiza uma contextualização temática o qual permeia: a gênese da comunicação científica, abordando em específico o surgimento dos periódicos científicos em meio digital; a discussão do movimento de acesso aberto e sua importância para o desenvolvimento da ferramenta OJS; a possibilidade de monitoramento do uso da web utilizando o *plugin* da ferramenta *Google Analytics*; o conceito de *web analytics* e a relação com o *Google Analytics* no que tange a abordagem das métricas e indicadores; a possibilidade de realizar práticas de mineração de dados com os dados gerados pela ferramenta *Google Analytics*. Provê quadros informacionais que contemplam: a descrição das dezenove métricas ou indicadores disponibilizados pelo *Google Analytics* (baseadas nas métricas estabelecidas pela *Web Analytics Association*), suas respectivas finalidades e estruturas de cálculo; a enumeração das principais tarefas e seus respectivos métodos ou técnicas de mineração de dados; a análise dos modelos de algoritmos para o processo de seleção de atributos. Indica a tarefa e o método de mineração que é considero como exeqüível de acordo com o objetivo proposto pela pesquisa. Propõe os tipos de algoritmos que devem ser utilizados para as fases de busca e avaliação de subconjuntos no processo de seleção de atributos. Fundamenta a idéia de mineração do uso da web, baseado em discussões teóricas realizadas por vários autores. Por fim, enumera possíveis trabalhados futuros que podem ser derivados desta discussão teórica.

Palavras chave: *Open Journal System*. Métricas. Indicadores. *Google Analytics*. Mineração de dados. Web Analytics.

ABSTRACT

This work presents the metrics and indicators, available in Google Analytics, and are applicable to editorial management software Open Journal System (OJS) for the practice of data mining. Based on a theoretical framework, provides a theme which permeates context: the genesis of scientific communication, in particular the begins scientific journals in digital media, the discussion of the Open Access Movement and its importance for the development OJS tool , the possibility of monitoring the User's Web using the Google Analytics plugin, the concept of web analytics and the relation with Google Analytics regarding the metrics and indicators, the possibility data mining with data generated by Google Analytics. This work studies the universe of information management and based its relevance to the administration of digital journals. In continuing, defines methodological route to ensure the achievement of the objective. With approval the methodology, sample informational frameworks that include: a description of the nineteen metrics or indicators provided by Google Analytics (based on metrics established by the Web Analytics Association), their structures and purposes of calculation, the enumeration of the main tasks and their methods or techniques of data mining, analysis of algorithms to model the feature selection process. Further, this work indicates the task and the mining method that is regarded as feasible in accordance with the proposed objective for the research. Therefore, proposes types of algorithms to be used for search and evaluation phases of the process of selecting subsets of attributes. Uphold the idea of using web mining, based on theoretical discussions carried out by several authors. Finally, lists possible future work that can be derived from this theoretical discussion.

***Keywords**: Open Journal System. Google Analytics metrics and indicators. Feature Selection. Data mining. Web Analytics.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

OJS	*Open Journal System*
C&T	Ciência e Tecnologia
TICS	Tecnologia, Informação e Ciências
OA	*Open Access*
OAI	*Open Access Initiative*
PKP	*Public Knowledge Project*
SEER	Sistema de Editoração Eletrônica e Revistas
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
OAI-PMH	*Open Access Initiative - Protocol for Metadata Harvesting*
FINEP	Financiadora de Estudos e Projetos
OECD	*Organization for Economic Co-operation and Development*
ISO	*International Standard Organization*
KDD	*Knowledge Discovery in Databases*

SUMÁRIO

1 INTRODUÇÃO

Discussões sobre o crescente volume de informações oriundas de pesquisas científicas - nesse caso, as informações de produtos da comunicação científica -, têm instigado pesquisadores de todas as áreas do conhecimento. A Gestão da Informação, como disciplina estruturada sobre o tripé das áreas de Ciência da Informação, Administração e Tecnologia da Informação, não poderia ficar sem contribuir com o desenvolvimento de pesquisas nesta temática.

Com o surgimento dos serviços web, o processo de comunicação científica passou a ter uma dinâmica mais ágil e com abrangência internacional. Sistemas de Gestão Eletrônica de Periódicos, como por exemplo, o *Open Journal System* (OJS), impulsionaram as comunidades acadêmicas a reverem seus papéis no processo editorial de periódicos científicos.

Ferramentas de *web analytics* que podem ser integradas ao OJS - como o *Google Analytics* (GA) -, tem disponibilizado aos administradores de periódicos científicos digitais métricas e indicadores relacionados ao comportamento dos atores (sejam autores, editores, leitores ou pesquisadores) envolvidos em todo o processo de comunicação científica.

Entretanto, a falta de aplicação de técnicas para análise dessas métricas informacionais pode prejudicar a busca por fomento em projetos. Fato posto é que sem a análise das métricas informacionais, não há a possibilidade de identificar os índices de visibilidade do periódico, por exemplo, e até mesmo, realizar uma interpretação do comportamento de navegação dos usuários. É por esta necessidade de análise das métricas e dos indicadores que esta pesquisa está direcionado.

Este relatório de pesquisa é composta por cinco capítulos. No primeiro, são apresentados os elementos básicos, constituintes do estudo: introdução, formulação do problema da pesquisa, justificativa e objetivos.

No Capítulo 2, apresenta-se a fundamentação teórica. Optou-se por subdividir este Capítulo de acordo com o Diagrama de Relacionamento de Temáticas (Apêndice A), sendo este desenvolvido para possibilitar uma visão direta do objeto de pesquisa deste trabalho.

O Capítulo 3 trata dos procedimentos metodológicos utilizados, tais como natureza do estudo, fases de pesquisa e, para fins didáticos, preferiu-se detalhar as ações realizadas em cada objetivo específico estabelecidos.

Os três primeiros capítulos são importantes, pois, são a sustentação para a compilação das informações apresentadas no Capítulo 4, o qual apresenta os resultados conquistados com esta pesquisa.

O Capítulo 5 apresenta uma discussão teórica em relação à proposta de aplicação da técnica de mineração de dados indicada como ideal para este trabalho o que, de certa forma, caracteriza-se como as considerações finais. Para finalizar o capítulo é apresentado um consolidado sobre o atendimento aos objetivos e as propostas de continuidade desta pesquisa.

Por fim, consideram-se como principais produtos informacionais desta pesquisa: a revisão teórica e o relacionamento entre as temáticas estabelecidas; a construção teórica em relação às métricas e aos indicadores possíveis para a ferramenta *Google Analytics*; e a discussão sobre a prática de mineração de dados para extrair conhecimento das métricas e dos indicadores possíveis para um periódico científico.

1.1 PROBLEMA

Um marco histórico para o desenvolvimento da Ciência e Tecnologia (C&T), de maneira global, foi a Segunda Guerra Mundial. Tal acontecimento do mundo moderno é tem sido considerado fundamental para o desenvolvimento sócio-econômico e fator gerador da competitividade entre países.

A partir da década de 1950 iniciou-se uma relação de cooperação entre Ciência (com a visão acadêmica) e Tecnologia (com a visão empírica) a qual deu origem a pesquisas em diversas áreas do conhecimento. Ou seja, a academia passou a se preocupar em realizar pesquisas não somente de ciência pura (para a construção de novas teorias), mas também passou se preocupar com o desenvolvimento de novos produtos e serviços para a sociedade. Para que tais pesquisas fossem desenvolvidas, houve a necessidade de se disponibilizar incentivos à pesquisa por meio do estabelecimento de organismos que, a seu ritmo, atuaram na formulação e implantação de políticas de C&T.

Durante a década de 70, com a liberação do acesso da ARPAnet[1], as universidades e instituições de pesquisa iniciaram um processo de desenvolvimento e identificação de novas formas de comunicação. Este processo estendeu-se até meados de 1990 quando a ARPAnet foi então transformada em NSFnet. Com o aumento significativo no número de usuários, a administração da "rede" foi transferida para instituições não-governamentais, como por exemplo, a *Internet Society*. No Brasil, paralelamente a crescente socialização da Internet no mundo, as pesquisas relacionadas à criação de uma infraestrutura adequada para o compartilhamento de informações iniciou-se em 1995 (MONTEIRO, 2001).

No ano de 2000 o Governo Federal Brasileiro organizou a Conferência Nacional de Ciência, Tecnologia e Inovação. Desse evento nasceu o "Livro Verde", um documento elaborado pelo Ministério da Ciência e Tecnologia (MCT) e pela Academia Brasileira de Ciência (ABC). O "Livro Verde" propõe a integração, coordenação e fomentação de ações necessárias para a criação da sociedade da informação no Brasil. Um dos itens relacionados no Livro Verde diz respeito à importância da expansão em rede dos centros que produzem e divulgam o

[1] Rede desenvolvida pela empresa *Advanced Research and Projects Agency* que tinha como finalidade inicial: ligar computadores dos departamentos de pesquisa e inteligência militar dos EUA.

conhecimento científico, bem como a democratização do acesso da população às Tecnologias de Informação e Comunicação (TICs) (TAKAHASHI, 2000).

Simultaneamente à popularização dos serviços web na Internet, o Movimento de Acesso Aberto (*Open Acess* ou, ora chamado de OA), proveniente da *Budapest Open Access Initiative*[2], estimulou as pesquisas em softwares de gerenciamento de informações e de publicação de revistas eletrônicas.

Antes mesmo do surgimento do OA, em 1998, o *Public Knowledge Project* (PKP) da *University of British Columbia*, deu início a realização de pesquisas e o desenvolvimento de softwares dedicados em melhorar a qualidade acadêmica e pública da comunicação e divulgação das pesquisas científicas. Um dos softwares desenvolvidos pelo PKP é o *Open Journal System* (OJS) que é um sistema de gerenciamento e publicação de periódicos eletrônicos. O OJS é um software de código aberto, livremente disponível para instituições em todo o mundo que tem o propósito de tornar público o conhecimento produzido por pesquisadores. Em 2002, OJS foi considerado um marco no Movimento de Acesso Aberto (PKP, [20-?]).

O Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) traduziu o OJS que, em sua versão brasileira, intitula-se Sistema de Editoração Eletrônica de Revistas (SEER).

Um estudo realizado pela PKP demonstrou que até janeiro de 2010 na América Latina, 1.537 periódicos científicos utilizavam o OJS como uma ferramenta de gerenciamento dos processos editoriais de publicações científicas. No Brasil, segundo o IBICT, até 2009, 1.349 periódicos científicos utilizavam o SEER. Tais informações são coletadas por meio dos registros realizados pelos administradores dos periódicos ou pelos administradores dos repositórios de periódicos (PKP, [20-?]; IBICT, 2010).

São os administradores e editores chefes que possuem liberações de acesso a funcionalidades específicas de gerenciamento estratégico no OJS. Entende-se por funcionalidades específicas de gerenciamento estratégico a possibilidade de gerar relatórios, estatísticas, métricas e indicadores sobre o periódico ou repositório digital.

---

[2] Definido por Chan (2002) como a disposição livre e pública na Internet, de forma a permitir a qualquer usuário a leitura, *download*, cópia, impressão, distribuição, busca ou *link* com o conteúdo completo de artigos, bem como a indexação ou o uso para qualquer outro propósito legal.

Para Kaushik (2010), uma métrica é uma medida estatística quantitativa que descreve eventos ou tendências. Januzzi (2003) afirma que indicador:

> [...] é a medida em geral quantitativa dotada de significado social substantivo, usado para substituir, quantificar ou operacionalizar um conceito social abstrato, de interesse teórico (para pesquisa acadêmica) ou programático (para formulação de políticas públicas) (JANUZZI, 2003, p.15).

Sendo assim, indicador é uma medida reservada para a descrição ou representação de um dado evento ou fenômeno. Uma métrica pode conter um ou mais indicadores (ROZADOS, 2005)

No OJS, a coleta de dados necessária para a geração de métricas e indicadores pode ser feita utilizando *plugins* que podem ser inseridos na ferramenta. Um dos *plugins* disponíveis no OJS é o *Google Analytics* (PKP, 2008).

O *Google Analytics*[3] (GA) é um serviço (ou ferramenta) gratuito de análise da web do grupo Google Inc. desenvolvida para medir quantitativamente o que acontece em *websites*. Foi inicialmente desenvolvida para auxiliar *webmasters* a otimizarem seus site para campanhas de marketing. Atualmente, sua principal funcionalidade é a de gerar estatísticas de tráfego na Internet sendo capaz de identificar: a) taxa de exibição de uma página; b) localização geográfica do visitante; c) a procedência física; d) sistema operacional utilizado pelo usuário; e) navegador; f) combinação do sistema operacional e navegador, bem como suas versões; g) resolução de tela; h) visitação em períodos diários, semanais, mensais e anuais; i) computador da ação, ou seja, de onde foi gerada a solicitação; j) descritores (ou termos de busca) utilizados pelos usuários nos motores de busca; entre outras funcionalidades (FERREIRA; CUNHA, 2008).

Cutroni (2010) detalha todos os processos relacionados ao funcionamento do *Google Analytics* sendo, para este projeto, o aspecto principal a questão da criação de um banco de dados com todos os dados coletados do tráfego do *website*. Ainda de acordo com o autor, as métricas e os indicadores disponibilizados pelo GA são calculados baseados neste banco de dados.

---

[3] Disponível em: < http://www.google.com/intl/pt-BR/analytics/>

Uma das possibilidades de manipulação de dados em um banco de dados é a prática da extração de conhecimento do mesmo e, para essa prática dá-se o nome de Mineração de Dados, ou seu termo original: *data mining*.

Seguindo a definição de Fayyad (1996), Goldschmidt e Passos (2005) e Tan et. al. (2006), a mineração de dados surge a partir de todas as fases de processamento baseado em uma metodologia e algoritmos que foram desenvolvidos para aprimorar técnicas de busca, tratamento e recuperação, sendo seu principal objetivo: buscar conhecimento que envolva e relacione a informação formando um conjunto de padrões estabelecidos por regras: SE <condições> ENTÃO <conclusões>.

Ora, tendo-se um banco de dados e, consequentemente, uma demanda pela extração de conhecimento do mesmo, pode-se intuir que a aplicação de técnicas de *data mining* pode auxiliar a análise das métricas e indicadores.

Isto posto, considerando-se o crescente uso de recursos web na Internet, a necessidade de disseminação do conhecimento científico e a necessidade de se estabelecer métricas e indicadores para justificar os investimentos realizados, surge um questionamento para esta pesquisa em informação: como o uso da mineração de dados pode auxiliar, na análise e interpretação de métricas e indicadores aplicáveis ao OJS?.

Desta pergunta, subdividiram-se em duas outras questões que devem ser respondidas a fim de complementar o entendimento da questão de pesquisa mais ampla:

1) quais métricas e indicadores gerados pelo *Google Analytics* podem ser utilizados na ferramenta OJS?
2) quais técnicas de mineração de dados podem ser aplicadas na análise das métricas e dos indicadores do *Google Analytics* no OJS?

1.2 JUSTIFICATIVA

Este projeto pode contribuir com discussões na área de informação, com abordagem organizacional, científica e tecnológica; essencialmente para projetos relacionados à identificação, levantamento e análise de métricas e indicadores.

As contribuições para a abordagem nas áreas do conhecimento de informação têm-se como potencial benefício a discussão sobre a relação entre a prática de mineração de dados e métricas e os indicadores voltados para as ferramentas de comunicação científica. Além disso, a discussão das técnicas de mineração de dados pode despertar o interesse pela atuação em um dos campos da Gestão da Informação: o estudo das metrias informacionais. O assunto tem despertado interesse de pesquisadores de diversas áreas do conhecimento, quase sempre tendo como objetivo ou contexto o uso dos indicadores para decisões de fomento.

Neste aspecto encaixam-se os administradores dos repositórios de periódicos científicos, que utilizam a ferramenta OJS. Estes podem se beneficiar com a discussão sobre métricas e indicadores (como técnicas de mensuração), fundamentando e justificando a existência (grau de relevância) dos repositórios perante a sociedade, além de garantir informações gerenciais e estratégicas para solicitar aos órgãos de fomento à pesquisa maior disponibilização de recursos.

Para a plataforma OJS as contribuições são potencialmente previstas, como agregadores de valor no uso da ferramenta de monitoramento *Google Analytics* que pode ser integrado ao OJS e prover dados potenciais a serem insumos para geração das métricas e indicadores que, por sua vez, pode auxiliar na geração de informações estratégicas.

Por fim, para a comunidade científica fica a contribuição e a possibilidade de identificação de tendências de pesquisas em cada área do conhecimento, pois, com a análise das métricas e dos indicadores e, também das questões relacionadas à mineração de dados, poder-se-á aplicar técnicas para a identificação de padrões e projeções de comportamento do usuário ou identificação de grupos de interesse.

1.3 OBJETIVOS

Neste item são apresentados os objetivos deste projeto, sendo estes divididos em um geral e quatro específicos.

1.3.1 Objetivo Geral

O objetivo geral do projeto é discutir as métricas e os indicadores disponibilizados pelo *Google Analytics*, aplicáveis ao *Open Journal System*, para a prática de mineração de dados.

1.3.2 Objetivo Específico

Para atingir o objetivo geral, foram-se estabelecidos os seguintes objetivos específicos:

a) identificar na literatura as métricas e os indicadores sugeridos para o *Google Analytics*;
b) identificar na literatura os métodos e as técnicas de mineração de dados que podem ser aplicados na base de dados gerada pelo *Google Analytics*;
c) estabelecer os critérios para a seleção de atributos da base de dados gerada pelo *Google Analytics;* e,
d) definir um método/técnica de mineração de dados e um conjunto de algoritmos de seleção de atributos que se caracterizam adequados para ser aplicado na descoberta de conhecimento a partir da base de dados gerada pelo *Google Analytics*.

2 LITERATURA PERTINENTE

A fim de fundamentar os princípios e ideais deste relatório de pesquisa, a necessidade de se apoiar em autores e, principalmente, em conceitos consagrados e reconhecidos pela comunidade acadêmica faz-se explícito neste momento da pesquisa.

Desta maneira, é estabelecido um diagrama dos relacionamentos das temáticas (Apêndice A) abordados para a realização desta pesquisa de informação. A seguir, estruturam-se os tópicos pertinentes ao entendimento dos conceitos, acontecimentos/eventos/fatos, ferramentas, técnicas e modelos relacionados a esta pesquisa.

2.1 A GÊNESE DA COMUNICAÇÃO CIENTÍFICA

Conforme abordado por Biojone (2001), a ciência como é vista hoje teve sua gênese no século XIV a partir da aceitação do método científico proposto por Francis Bacon, bem como com a aplicação das sociedades e academias científicas.

Para vários autores comunicação científica é o conjunto de processos que engloba a produção, a disseminação (incluindo o processo de inserção nos canais de comunicação formais ou informais), a recuperação e o uso da informação (MUELLER, 1994; GARVEY apud MIRANDA, 1996 e; MEADOWS, 1999).

Para Le Coadic (1996), assegurar a troca de informações e conhecimento entre os cientistas é a função da comunicação científica que, de certa forma, sustenta toda a prática da pesquisa científica. Como abordado por Weitzel (2006), promover a evolução das pesquisas da comunicação científica é compreendida como: o processo que abrange a criação, a transmissão e o uso do conhecimento. Desta maneira, as práticas científicas propiciam a organicidade do conhecimento científico.

Os primeiros artefatos informacionais divulgados à sociedade científica são datados do séc. XVI, sendo estes, precursores do modelo moderno de comunicação científica (GONÇALVES, et. al., 2006). Porém, a ênfase maior na cobrança para que

os pesquisadores divulgassem novos resultados de pesquisa surgiu nos séculos XVII e XVIII, fazendo com que os acadêmicos contribuíssem de maneira mais efetiva com o conhecimento científico. Isso propiciou uma expansão do número de trabalhos publicados e o aumento do número de periódicos científicos (BURKE, 2003). No Brasil, o primeiro periódico científico surgiu em 1827 e foi intitulada Propagador das Ciências Médicas ou Anais de Medicina, Cirurgia e Farmácia.

Em 1996 existiam cerca de 25 milhões de artigos científicos, e publicados em cerca de 200.000 periódicos científicos distribuídos entre as diversas áreas do conhecimento (CHARTRON, 1996 apud BIOJONE, 2003).

Com a introdução e adoção das novas tecnologias eletrônicas na comunicação científica, aproximadamente na década de 60, várias alterações foram sendo evidenciadas na história da explicitação do conhecimento científico (BIOJONE, 2003).

A partir dos anos 80, com a socialização dos recursos de tecnologia eletrônico como a Internet e a web, problemas ainda existentes no modelo de fazer ciência são direcionados as novas possibilidades tecnológicas (WEITZEL, 2006).

Em seu livro "A sociedade em rede", Castells apresentação claramente o surgimento da Internet:

> A criação e o desenvolvimento da Internet nas três últimas décadas do século XX foram conseqüência de uma fusão singular de estratégia militar, grande cooperação científica, iniciativa tecnológica e inovação contracultural (CASTELLS, 2005, p.82).

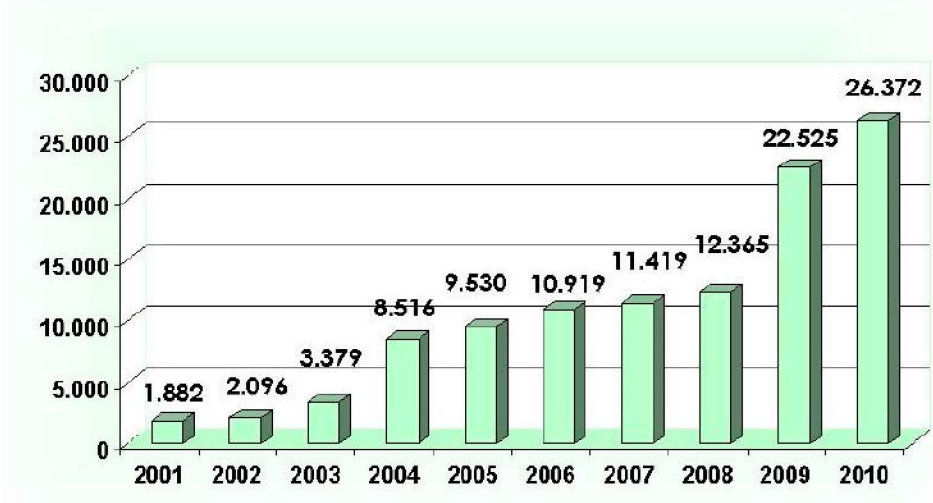
O autor aborda os relacionamentos estabelecidos para a criação da grande rede mundial de computadores:

> A certa altura tornou-se difícil separar a pesquisa voltada para fins militares das comunicações científicas e das conversas pessoais. Assim, permitiu-se o acesso à rede para cientistas de todas as disciplinas e, em 1983, houve a divisão entre ARPANET (ARPA é a sigla para Agência de Projetos de Pesquisa Avançada), dedicada a fins científicos, e a MILNET, orientada diretamente às aplicações militares. [...] todas as demais redes usavam a ARPANET como espinha dorsal do sistema de comunicação. A rede das redes que se formou durante a década de 1980 chamava-se ARPANET-INTERNET, depois passou a chamar-se INTERNET, ainda sustentada pelo Departamento de Defesa e operada pelo National Science Fundation (CASTELLS, 2005. p.82).

Segundo Targino (1999), o *Eletronic Information Exchange System* foi primeiro periódico eletrônico fundado, entre as décadas de 70 e 80 nos Estados Unidos. Já no Brasil, apesar de publicar periódicos desde 1862 (Gazeta Médica do Rio de Janeiro), o primeiro periódico eletrônico, pelo menos a utilizar avaliação por pares foi *o Postmodern Culture*, surgido em setembro de 1990, primeiro em formato de correio eletrônico, depois em disquete; e em janeiro de 1994, surgiu a versão em hipermídia na Internet (REIS; GIANNASI-KAIMEN, 2007).

Em busca realizada no Portal de Periódicos da CAPES, encontrou-se a FIGURA 1 que ilustra a evolução dos periódicos online disponibilizados para acesso.

FIGURA 1 – Evolução dos Títulos de Periódicos no Portal CAPES - 2001-2010



Fonte: Portal CAPES, 2011

Vale ressaltar que existem diferenças entre os periódicos científicos impressos e os periódicos científicos eletrônicos. Segundo Barreto (1998) caracteriza as principais diferenças entre o modelo tradicional e o modelo de comunicação científica eletrônica. O primeiro tem uma visão envelhecida baseado no acesso a um documento físico, outro modelo possibilita a interação direta do receptor sem intermediários.

Weitzel afirma que:

> A consolidação de estruturas de redes e sistemas de informação científica [...] desencadeou as diversas iniciativas de uma parte considerável da comunidade científica, visando à legitimação de novas formas de comunicação científica na Internet (WEITZEL, 2006b, p.54).

Baseado nisso e no crescimento vertiginoso dos periódicos científicos eletrônicos, a comunidade acadêmica vivenciou problemas com o modelo de negócio adotado pela maior parte das editoras de periódicos.

Para muitos pesquisadores, o modelo de negócio aplicado pelos periódicos científicos parecia não mais corresponder aos ideais da comunicação científica. Os editores, envolvidos no processo de comunicação de pesquisas científicas, passaram a exigir pagamento de assinaturas aquém do ideal. A justificativa para cobrança seria que o valor arrecadado custearia o processo de editoração e publicação de periódicos.

A cobrança de assinaturas para periódicos científicos impressos é compreensível, pois, há toda uma logística envolvida no processo de publicação das mesmas. Porém, no caso dos periódicos científicos eletrônicos os serviços web proporcionaram mudanças consideráveis nesta logística. Não há mais gastos com impressão, transporte e entrega dos periódicos científicos, pois, eles estão acessíveis a qualquer momento na web. Sendo assim, entende-se que a prática de comunicação do conhecimento científico passou a ser uma prática de comércio do conhecimento. Esse paradigma é conhecido na literatura como a crise dos periódicos científicos (BIOJONE, 2001; BAPTISTA et al., 2007).

2.2 OS MOVIMENTOS PARA O ACESSO ABERTO

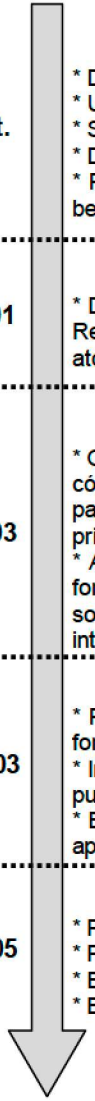
Considerando a evolução das tecnologias da informação e a crise dos periódicos científicos, discussões sobre Acesso Aberto (ora denominado também de Acesso Livre) começaram a surgir na comunidade científica.

Conforme abordado por Baptista et al. o Acesso Livre é:

> O resultado de (1) uma reação dos pesquisadores ao modelo de negócios de editoras comerciais de revistas científicas (e seus preços cada vez mais altos preços de assinatura); e da (2) crescente conscientização do aumento de impacto provocado pela disponibilização de documentos científicos livres de barreiras ao acesso (BAPTISA, et. al., 2007, p.2)

Há várias iniciativas de sucesso sobre Acesso Aberto/Livre que são mencionadas em uma vasta literatura relacionada ao tema. Podemos citar alguns autores: Garcia e Sunye (2003); Harnad et al. (2004), Silva, Ramos e Noronha (2006); Borges (2006); Baptista et. al. (2007); Silva e Tomaél (2008) e, Guedes (2010). Contudo, existem alguns marcos importantes para o movimento que são descritos na linha do tempo da FIGURA 2.

FIGURA 2 – Linha do tempo dos principais acontecimentos relacionados ao Acesso Aberto/Livre



21 e 22 out. 1999

Convenção de Santa Fé (México)

* Desenvolvimento de modelos e planos de negócio sustentáveis para o auto-arquivo e a publicação em Open Access (OA);
* Uso das redes de bibliotecas para mobilizar e suportar o OA globalmente;
* Suporte aos investidores em países menos desenvolvidos para que possam publicar em revistas científicas em OA que cobrem taxas de publicação;
* Desenvolvimento de software, incluindo *templates*, para a publicação em OA, auto-arquivo, indexação e navegação e,
* Promoção da filosofia do OA entre Fundações e outros doadores, agências de financiamento de ciência e investigação, bibliotecas e universidades, bem como governos, governantes e organizações internacionais mundiais (BOAI, 2002).

01 dez. 2001

Iniciativa de Acesso Livre de Budapeste

* Definiram-se duas estratégias básicas, ambas baseadas no uso do protocolo *Open Archives Initiative – Protocol for Metadata Harvesting* (OAI-PMH). Representa uma das primeiras fortes reações da comunidade científica que ocorreu de forma planejada, organizada e envolvendo uma variedade de atores (COSTA, 2006).

11 abr. 2003

Declaração de Bethesda

* O(s) autor(s) ou detentor(s) de direitos de autor outorga aos utilizadores um uso livre, irrevogável, mundial e perpétuo de acesso, e uma autorização de cópia, uso, distribuição, transmissão e apresentação pública do trabalho e ainda para fazer e distribuir trabalhos derivados, em qualquer meio digital para qualquer objetivo responsável, sujeito à atribuição de autoria, bem como o direito de fazer um pequeno número de cópias empresas para uso privado;
* A versão completa do trabalho e de todos os materiais suplementares, incluindo uma cópia da permissão nos termos enunciados, é depositada num formato eletrônico apropriado imediatamente a seguir à publicação inicial em, pelo menos, um repositório suportado por uma instituição acadêmica, sociedade científica, agência governamental ou qualquer outra organização bem estabelecida que procure promover o OA, distribuição sem restrições, interoperabilidade e arquivo ao longo do tempo (BORGES, 2006).

22 out. 2003

Declaração de Berlim

* Promover a Internet como o instrumento funcional para uma base global de conhecimento científico e reflexão humana, e a especificar medidas que formuladores de política, instituições de pesquisa, agências de fomento, bibliotecas e museus precisam considerar;
* Implementar uma política para requerer de seus pesquisadores que depositem em um repositório de acesso aberto uma cópia de todos os seus artigos publicados e,
* Encorajar seus pesquisadores a publicar seus artigos de pesquisa em periódicos de acesso aberto – onde haja um periódico apropriado – e prover o apoio necessário para que isso aconteça (COSTA, 2006).

13 set. 2005

Manifesto Brasileiro de Apoio ao Acesso Livre à Informação Científica

* Promover o registro da produção científica brasileira em consonância com o paradigma do acesso livre à informação;
* Promover a disseminação da produção científica brasileira em consonância com o paradigma do acesso livre à informação;
* Estabelecer uma política nacional de acesso livre à informação científica e,
* Buscar apoio da comunidade científica em prol do acesso livre à informação científica (IBICT, 2005).

Fonte: o autor, 2011

Segundo Weitzel (2006), o Open Archives Iniaitive (OAI) pode ter contribuído para o aparecimento do Movimento de Acesso Livre, pois, o OAI é uma iniciativa que surgiu com a Convenção de Santa Fé em 1999. O Movimento de Acesso Livre teve origem com a Declaração de Budapest em 2001. Ambos foram inseridos em um contexto mais amplo denominado Modelo Open Acess (OA).

O OAI tem como objetivo central a promoção do desenvolvimento de arquivos de artigos científicos eletrônicos, onde cada autor é o responsável por depositar ele mesmo seus resultados de pesquisa. De certa forma, o OAI deve contribuir para a transformação do processo de comunicação científica, definindo padrões técnicos e organizacionais para que seja possível a comunicação e o intercâmbio de informação entres esses arquivos (VAN DE SOMPEL, 2000).

2.3 O OPEN JOURNAL SYSTEM (OJS)

Em 1998, o *Public Knowledge Project* (PKP) da *University of British Columbia*, deu início a realização de pesquisas e o desenvolvimento de softwares dedicados em melhorar a qualidade acadêmica e pública da comunicação e divulgação das pesquisas científicas.

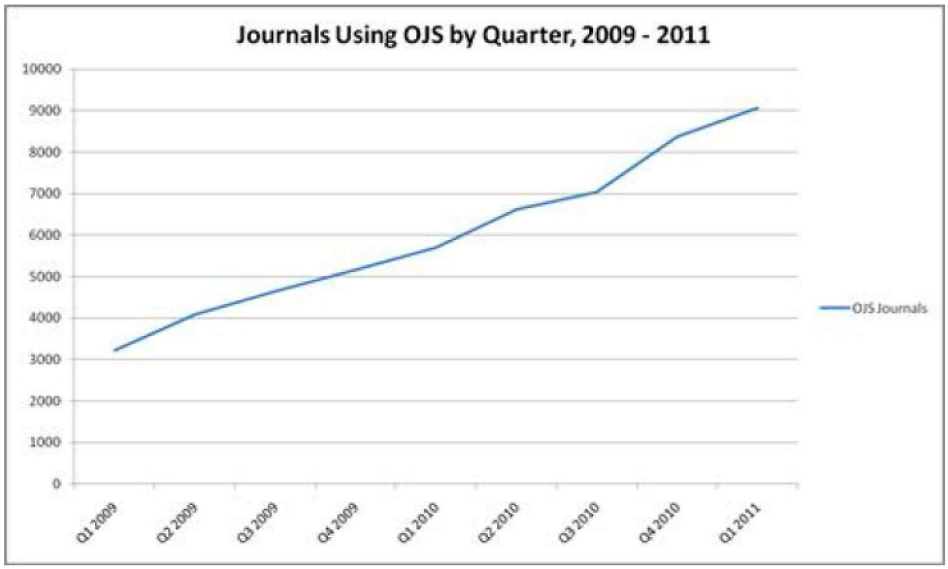
Estas pesquisas e o desenvolvimento de softwares pelos colaboradores do PKP apontam para a necessidade de investigar a contribuição destas novas tecnologias potenciais para o conhecimento público/compartilhado, especialmente as organizações acadêmicas (CHAN et al. 2002).

Um dos softwares desenvolvidos pela PKP é o *Open Journal System* (OJS) um sistema de gerenciamento e publicação de periódicos eletrônicos. O OJS é um software de código aberto, livremente disponível para instituições em todo o mundo que tem o propósito de tornar público o conhecimento produzido por pesquisadores. Em 2002, OJS foi considerado um marco no Movimento de Acesso Aberto (PKP, [20-?]).

O Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) traduziu o OJS que, em sua versão brasileira, intitula-se Sistema de Editoração Eletrônica de Revistas (SEER).

De acordo com o monitoramento realizado pela PKP[4], até abril de 2011 em todo o mundo, foram contabilizadas 9 mil instalações da ferramenta, para a gestão de periódicos científicos, monografias ou outras práticas de ensino. A evolução do crescimento de instalações do OJS em todo o mundo pode observada na FIGURA 3.

FIGURA 3 - Evolução por trimestre do número de instalações do OJS pelo mundo - 2009-2010



Fonte: PKP, 2011

Ainda segundo o PKP[5], na América Latina, até janeiro de 2010, 1.537 periódicos científicos utilizavam o OJS como uma ferramenta de gerenciamento dos processos editoriais de publicações científicas. No Brasil, segundo o IBICT, até 2009 1.349 periódicos científicos utilizavam o SEER. Tais informações são coletadas por meio dos registros realizados pelos administradores dos periódicos ou pelos administradores dos repositórios de periódicos (PKP, [20-?]; IBICT, 2010).

Uma das grandes vantagens do OJS (hora denominado também de SEER neste projeto), conforme abordado por Batista (2007) é de que esta ferramenta permite a compatibilidade com o protocolo OAI-PMH para a consulta de metadados, permitindo, assim, aumentar a visibilidade dos artigos científicos publicados nos periódicos que utilizam o OJS.

[4] Resultado do monitoramento está disponível em <http://pkp.sfu.ca/ojs-user-numbers>
[5] Estudo por continente está disponível em <http://pkp.sfu.ca/ojs-geog>

Outro aspecto que difere o OJS das demais ferramentas de gerenciamento de periódicos científicos, diz respeito à possibilidade de inclusão nos artigos no que denomina o PKP (2007) "Ferramentas de Leitura":

> As “Ferramentas de Leitura” foram desenvolvidas para auxiliar na leitura de pesquisadores experientes e novatos, proporcionando um rico contexto de material relacionado de uma variedade de fontes e recursos de acesso livre. As ferramentas utilizam as palavras-chave do autor para busca automática em bases de dados de acesso livre relevantes. O conteúdo relacionado é apresentado em outra janela do navegador. Os leitores possuem a escolha de um conjunto de ferramentas de bases de dados, bem como acesso às informações sobre a base em questão (PKP, 2007).

Uma das funcionalidades que o OJS disponibiliza para os administradores dos periódicos científicos online é a possibilidade da inserção de *plugins* externos a ferramenta.

Um *plugin* (conhecido também por *plug-in*, *add-in* ou *add-on*) é uma extensão de uma parte de um software ou ferramenta que possibilita realizar atividades além do que era possível. Uma dessas possibilidades é a integração entre outras partes de outro software (ferramenta ou serviços web) ou dele mesmo. Visa à reutilização de componentes para construir outros componentes. Geralmente são desenvolvidos por terceiros, mas também podem ser criados pelos próprios desenvolvedores do software original.

O OJS disponibiliza, em sua versão 2.3.1, um sistema de gerenciamento de *plugins* que possibilita ao administrador do periódico científico a configuração de diversos *plugins* tais como: COUNTER Statistics, TinyMCE, *Web Feed*, *JQuery*, *External Feeds*, *Thesis Abstracts*, *Rounded Corners* e o qual será foco de estudo deste projeto o *Google Analytics plugin*.

Pelo fato de ocorrer essa interoperabilidade entre OJS e *Google Analytics* é que se poderá coletar os dados de navegação dos usuários de um periódico científico digital e realizar práticas de *web analytics*.

2.4 *WEB ANALYTICS* E O *GOOGLE ANALYTICS*

Para verificar a existência de uma possível correlação entre *web analytics* (denominado igualmente como web analítica) e o *Google Analytics*, há a necessidade de se compreender ao que o primeiro termo remete.

De acordo com a *Web Analytics Association*[6], *web analytics* é a medição, coleta, análise e comunicação de dados da Internet para objetivar a compreensão e otimização do uso da Web. Sendo conceito também adotado pelo Comitê *de Web Analytics* do Brasil[7]. Sendo assim, a *web analytics* é uma prática (ou metodologia) e não uma ferramenta.

Conforme abordado por Cutroni (2010), o *Google Analytics* não tem o mesmo significado que *web analytics*, sendo esta definida por Kaushik como:

> [...] a análise de dados qualitativos e quantitativos de seu website e de sua concorrência, para motivar uma melhora contínua da experiência online que seus clientes e clientes potenciais experimentam, traduzindo-se nos resultados desejados (online e offline) (KAUSHIK, 2007, p.5).

Entende-se que o *Google Analytics* é uma ferramenta para a aplicação da w*eb analytics* e não, propriamente, a w*eb analytics* em si (FANG, 2007).

A definição de Kaushik (2007) direciona o pensamento dos pesquisadores para três possíveis princípios:

a) a mensuração dos dados qualitativos e quantitavos;
b) a melhoria contínua do website analisado e,
c) o alinhamento das estratégias de mensuração com as estratégias de negócio relacionadas.

O terceiro princípio abordado pelo autor conduz a idéia de que: as métricas e indicadores estabelecidos, por meio de um processo de mensuração, devem ser aplicados e, consequentemente, analisados sob a ótica das estratégias de negócio.

Na gênese da *web analytics,* aproximadamente em 1995 com o *Analog*[8], o tratamento dos dados para a geração dos relatórios de análise dava-se pela leitura

---

[6] Conceito em seu idioma original disponível em <http://www.webanalyticsassociation.org>
[7] Para integrante, desde 2005, da *Interactive Advertising Bureau*. Disponível em < http://www.iabbrasil.org.br/arquivos/doc/glossario_de_metricas.pdf>
[8] Analog, desenvolvido pelo Dr. Stephen Turner, foi um dos primeiros programas de análise do arquivo de log amplamente disponível na Web (KAUSHIK, 2007).

de *log*[9] gerados pelo servidor HTTP. Os *logs* do servidor HTTP capturavam não somente a idéia que de um usuário acessava um website, mas informações adicionais, como por exemplo: o nome do arquivo, hora, referência (página que faz a solicitação), endereço Internet Protocol (IP), identificador do navegador, sistema operacional, entre outros registros (KAUSHIK, 2007). Este, segundo alguns autores como Kaushik (2007) e Cutroni (2010) seria o auge da medição do *hits*[10] e da análise dos tráfegos e do uso de bandas.

Devido algumas dificuldades que passaram a surgir na análise dos *logs* dos servidores *web*, o conceito de análise por *tags JavaScript*[11] passou a ser o novo padrão para coletar dados dos *websites*. Grandes empresas como *Coremetrics*, *Omniture*, *WebTrends* e *Urchin* passaram a disponibilizar para seus clientes ferramentas de análise web. O *Google Analytics* surge quando a *Urchin Software Corporation* é adquirida pelo Google Inc. em abril de 2005.

O *Google Analytics* é baseado no sistema que foi estruturado com o módulo de estatísticas Urchin, porém ainda hoje o sistema original de instalação em computadores é vendido pelo Google como um produto diferenciado.

Atualmente, o Google Inc. disponibiliza a ferramenta para aplicação de web analytics denominada *Google Analytics* de maneira gratuita para toda a sociedade em 31 idiomas.

Da mesma maneira que as outras ferramentas do mercado se beneficiam com o uso da metodologia de análise de *logs* gerados por *tags JavaScritps*, o *Google Analytics* coleta os dados para geração de relatórios em sua plataforma de acesso web.

A análise do *log* gerado pelo código *JavaScript* que deve ser incluído na página inicial a ser analisada, possibilita o envio de dados ao *Google Analytics*, que, por sua vez, o reproduz na conta do website (ou período online, para o contexto desta pesquisa) em análise. Desse modo, todos os visitantes que acessarem o periódico científico online, por exemplo, terão seus dados capturados e arquivados na ferramenta (FERREIRA; CUNHA, 2008).

---

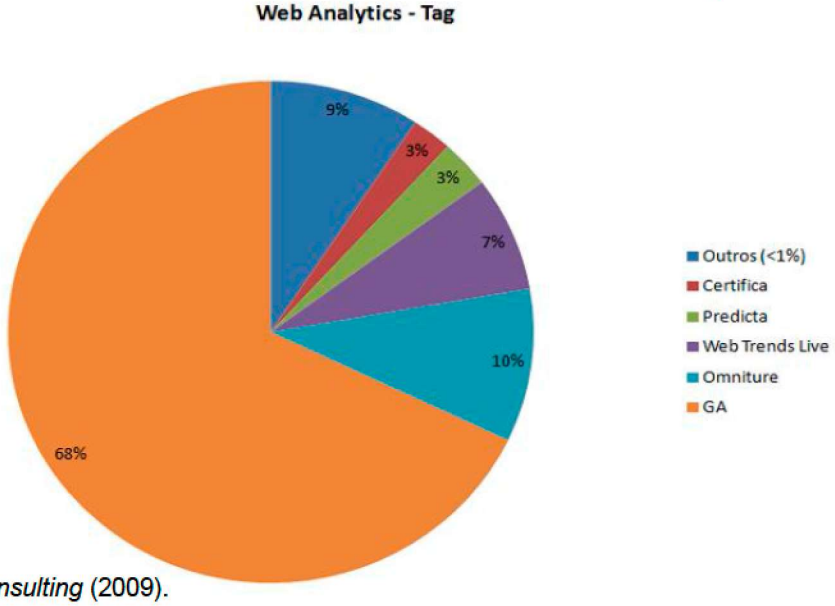
[9] *Logs* são arquivos-texto gerados por um software que registra eventos (por exemplo, à hora em que ocorrem esses eventos, que tipo de evento, o que ocorreu, entre outras informações) (FERREIRA; CUNHA, 2008).
[10] *Hits* é toda troca de dados realizada entre um cliente e um servidor web. Exemplo: um usuário solicita, através de seu navegador, (*browser*) uma páginaweb. Neste caso, tem-se um hit. (DIAS, 2002).
[11] Linhas de código em *JavaScript* são adicionadas a cada página e são iniciadas quando a página é carregada, e enviam dados para um servidor de coleta de dados (KAUSHIK, 2007).

As autoras, ao explicarem seu estudo realizado no Portal Revcom, apresentam de maneira clara a importância da geração e análise dos dados realizada pelo *Google Analytics*:

> A geração de arquivos de log é feita sem obstrução e, se processados apropriadamente, podem fornecer estatísticas de uso, com dados úteis para estudos de usuários e identificação de seu perfil, ferramentas utilizadas e procedimentos de busca e uso de informação. Com base em tais resultados e traçando um paralelo do desempenho do próprio serviço, é possível avaliar as condições mais adequadas para construção da interface do Portal, oferecer melhores serviços, implementar mecanismos para auxiliar os usuários na busca por informações, rever a usabilidade e interface de acesso, dentre outras ações. A ferramenta em questão é de simples e fácil implementação (FERREIRA; CUNHA, 2008, p.47).

De acordo com um estudo realizado por *WA Consulting* (2009) e, apontado na FIGURA 4, pode-se constatar que no Brasil o *Google Analytics* é utilizado como ferramenta predominante para a prática de *web analytics*.

FIGURA 4 - O Market Share das ferramentas de web analytics no Brasil



Fonte: WA *Consulting* (2009).

Isso significa que o mercado para pesquisas com as ferramentas de *web analytics* está crescendo vertiginosamente, além de, as pesquisas - com resultados positivos ou negativos - realizadas com a ferramenta *Google Analytics* podem

impactar mais do que os 68% do *market share* de usuários da ferramenta no Brasil; pode impactar os milhões de usuários no mundo todo.

2.5 MÉTRICAS E INDICADORES

"*Se algo não pode ser medido, ele realmente não existe*"
Willian Thompson - Lorde Kelvin (1824 - 1907),
Físico-matemático Irlandês

Baseado na frase célebre do físico-matemático irlandês, Willian Thompson, a necessidade de se medir objetos, eventos e razões não é apenas uma preocupação da ciência, mas é também uma forma de aplicar sentido a fenômenos empíricos. É consenso para a comunidade científica que as formas de mensuração são realizadas pela elaboração de métricas e indicadores.

Para Lovelle (2004), métrica é um valor numérico ou o valor nominal atribuído a características ou atributos de um ser (tangível ou intangível) computados a partir de um conjunto de ações observáveis e consistentes, podendo ser: direta ou indireta, interna ou externa, objetiva ou subjetiva.

Kaushik (2010) define métrica como uma medida estatística quantitativa que descreve eventos ou tendências. Contudo, ambos os autores acima não contextualizam o relacionamento com o termo "indicador".

Geisler (*apud* MUELLER, 2008), ao abordar seu estudo sobre a mensuração da ciência e tecnologia, oferece uma definição para ambos os termos, dizendo que:

> [...] indicadores da ciência como um termo genérico que se aplica a um amplo espectro de medidas quantitativas utilizadas para medir atividades, insumos, e resultados da pesquisa, desenvolvimento e inovação; e define o termo métricas como um sistema de medidas que inclui o item objeto da medida, a unidade de medida e o valor da unidade (GEISLER apud MUELLER, 2008, p.27).

Percebe-se, então, que métrica é um sistema ou conjuntos de objetos, unidades ou valores e, indicador é uma medida quantitativa de um objeto. Sendo assim, um indicador é uma medida descritiva e representativa de eventos ou fenômenos e, uma métrica pode conter um ou mais indicadores.

Outros conceitos de indicador são encontrados na literatura. Rozados (2005) apresenta os conceitos estabelecidos da FINEP, OECD e da ISO conforme apresentado no diagrama de conceitos sobre indicadores (QUADRO 1):

QUADRO 1 - Relação de conceitos para o termo "indicador"

Autor ou Instituição	Conceito
Financiadora de Estudos e Projetos (FINEP)	Especificação quantitativa e qualitativa para medir o atingimento de um objetivo.
Organisation for Economic Co-operation and Development (OECD)	Uma série de dados definidos para responder perguntas sobre um fenômeno ou um sistema dado.
A International Standart Organization (ISO)	Expressão (numérica, simbólica ou verbal) empregada para caracterizar as atividades (eventos, objetos, pessoas), em termos quantitativos e qualitativos, com o objetivo de determinar o valor

Fonte: adaptado de ROZADOS, 2005

De maneira pertinente, Rozados (2005) constata que "medida, qualitativo e quantitativo" são conceitos comuns para todas as definições acima citadas. Isso quer dizer, segundo a autora, que:

> [...] indicadores nada mais são do que unidades que permitem medir – caso de elementos quantitativos, ou verificar – caso de elementos qualitativos, se estão sendo alcançados os objetivos ou as mudanças previstas. Também possibilitam conhecer melhor os avanços em termos de resultados ou de impactos. Um indicador é, portanto primordialmente, uma ferramenta de mensuração, utilizada para levantar aspectos quantitativos e/ou qualitativos de um dado fenômeno, com vistas à avaliação e a subsidiar a tomada de decisão (ROZADOS, 2005, p.62)

Sob a perspectiva das características dos indicadores, Martinez et al (1998) sugerem que a validade dos dados deve seguir os seguintes preceitos: a) generalidade; b) possibilidade de correlação entre as distintas variáveis ou os diferentes contextos e, c) temporalidade.

Já nos aspectos relacionados à classificação dos indicadores, Rozados (2005) apresenta um constructo baseado em três autores: Briand, Sutter e Moore, separando os indicadores em dois grupos sendo eles quantitativos ou qualitativos.

Os quantitativos são as unidades de contagem e os qualitativos são as resultados de pesquisas de qualidades e avaliação. Sendo assim, por serem instrumentos de avaliação os indicadores, são instrumentos de gestão.

Ainda de acordo com a autora, é válido ressaltar o aspecto de que para que os indicadores sejam ferramentas úteis, devem seguir uma temporalidade regular permitindo a percepção das tendências no tempo e nos dados, além de possibilitar comparações internacionais quando for o caso.

Seadi et. al. (2002) afirmam que toda a avaliação parte de um princípio de comparação, sendo, portanto, necessário que haja dados passíveis de serem comparados, que permitam ser coletados de forma semelhante em todos os casos. Sendo assim, torna-se efetivo o uso de múltiplos indicadores, pois são necessários para proporcionar uma adequada cobertura de dimensões e aspectos de processos complexos, atividade e resultados. A comparação entre os dados que geram os indicadores, a fim de se abstrair novas perspectivas de análise poder-se-á ser feita pela aplicação de técnicas de *data mining*.

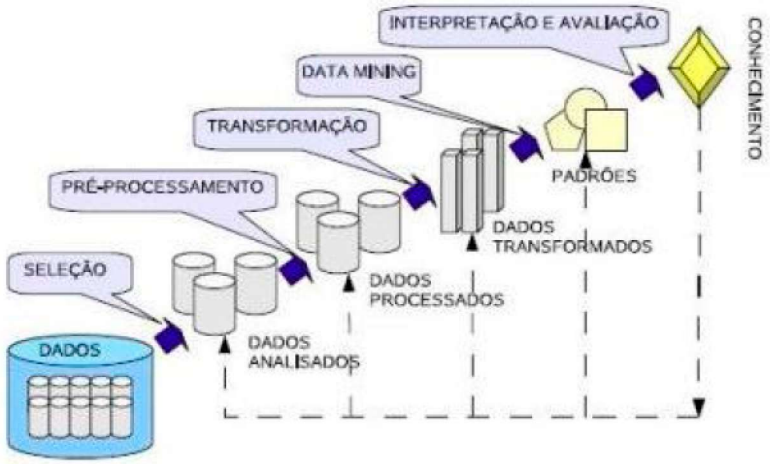
2.6 MINERAÇÃO DE DADOS

Para Tonus e Costa (2011), desde quando o conhecimento passou a ser sinônimo de poder, o ser humano e, principalmente, suas organizações têm buscado formas de entender, de maneira correta, como é o comportamento dos usuários. Isso se tornou tangível quando se passou a usufruir dos benefícios da informática para a mineração, monitoramento e mensuração dos dados.

Aproximadamente em 1995, o *data mining* foi o resultado da união entre três áreas: estatística clássica, inteligência artificial e aprendizado de máquina, sendo a primeira a mais antiga das três.

Além disso, o *data mining* é parte de um processo maior conhecido como *Knowledge Discovery in Databases* ou KDD (FIGURA 5) que tem por finalidade permitir a extração não trivial de conhecimento previamente desconhecido e, potencialmente útil, de um banco de dados (SFERRA; CORRÊA, 2003).

FIGURA 5 - Diagrama de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases - KDD)



Fonte: MORELLATO, 2008

Fayyad (1996 apud BRAGA, 2005), um dos maiores conhecedores do tema, indica que a mineração de dados é aplicada para a descoberta do conhecimento em bases de dados, ou seja, de dados imperceptíveis extraem-se padrões que possibilitarão construir o conhecimento mediante a transformação dos dados em informação.

Fayyad (2011) diz que *data mining* é encontrar estruturas interessantes em dados. Para o autor, as estruturas referem-se a padrões estatísticos, modelos preditivos e relacionamentos ocultos. O termo "interessante" fica de maneira subentendida como um critério diferencial para as pesquisas de *data mining*: os objetivos. Logo, é a finalidade da pesquisa que irá direcionar a aplicação das técnicas de mineração de dados.

Braga (2005) estabeleceu uma sequência genérica que consiste em um conjunto de etapas que são as mais adequadas para a prática de um projeto de mineração de dados: a) definição do problema; b) aquisição e avaliação dos dados; c) extração de características e realce; d) Plano de prototipagem, prototipagem e desenvolvimento de modelo; e) avaliação de modelo; f) implementação e, g) avaliação do ROI (pós-projeto).

Para entendimento do KDD, faz-se necessário a explicação clara de cada uma das etapas desse processo. Para tal, elaborou-se um diagrama (QUADRO 2) contendo um resumo, baseado em De Amo (2004) e Braga (2005)

QUADRO 2 – RESUMOS DAS FINALIDADES DAS ETAPAS DO KDD

Etapa	Resumo	Resultado
Primeira Etapa: SELEÇÃO	Tem início com o entendimento do domínio da aplicação e dos objetivos a serem atingido. Fase onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados (DE AMO, 2004; BRAGA, 2005).	Dados preparados e selecionados
Segunda Etapa: PRÉ-PROCESSAMENTO	A limpeza dos dados (ou Data Cleaning) é realizada por meio de um pré-processamento, visando assegurar a qualidade dos dados selecionados. Fase onde são eliminados ruídos e dados inconsistentes. Etapa onde são selecionados os atributos que interessam ao usuário (BRAGA, 2005).	Dados sem ruídos e pré-processados
Terceira Etapa: TRANSFORMAÇÃO	Etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de mineração (DE AMO, 2004).	Arquivo com dados preparados para a aplicação das técnicas de mineração de dados
Quarta Etapa: DATA MINING	Etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse. O objetivo principal desse passo é a aplicação de técnicas de mineração nos dados pré-processados, o que envolve ajuste de modelos e/ou determinação de características nos dados (DE AMO, 2004; BRAGA, 2005).	Padrões, regras, ou qualquer resultado obtido pela prática de mineração de dados.
Quinta Etapa: INTERPRETAÇÃO E AVALIAÇÃO	Engloba a interpretação dos padrões descobertos e a possibilidade de retorno a qualquer um dos passos anteriores. Assim, a informação extraída é analisada (ou interpretada) em relação ao objetivo proposto, sendo listadas e apresentadas as melhores informações (DE AMO, 2004; BRAGA, 2005).	Relatório de Análise das informações para geração de conhecimento.

Fonte: o autor.

Conforme apresentado por Braga (2005), é importante salientar que as técnicas de mineração de dados utilizada para conduzir as operações devem ser adaptadas para cada tipo de problema em particular. Uma técnica se adéqua melhor a uma determinada questão, do que outra técnica. Logo, a percepção do pesquisador ou analista é fundamental para o sucesso dos resultados da prática de *data mining*.

Existem inúmeras técnicas e maneiras de se realizar a mineração de dados. A seguir, baseado em De Amo (2004) e Braga (2005), descreve-se uma breve caracterização do processo de *Data Mining*:

a) **Classificação ou Predição**: associação ou classificação de um atributo a uma ou várias classes pré estabelecidas. Os objetivos dessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de várias populações, além da classificação das observações em uma ou mais classes predeterminadas. O sentido é a elaboração de regras que possa classificar, de maneira otimizada, a identificação de uma futura classe;

b) **Modelos de Relacionamento entre Variáveis**: associação de um atributo a uma ou várias predições independentes ou exploratórias. A aplicação de estatísticas como regressão linear simples, múltipla e modelos lineares por transformação são utilizadas para verificar o relacionamento funcional que, eventualmente, possa existir entre duas variáveis quantitativas;

c) **Análise de Agrupamento (Cluster)**: associação de um atributo a um ou vários grupos (ou clusters) não definidos, sendo então, diferente da técnica de classificação. Os grupos são estabelecidos baseados na similaridade ou em modelos probabilísticos. Tem como objetivo a detecção de diferentes grupos dentro de um grande conjunto de dados;

d) **Sumarização:** as funções de sumarização são freqüentemente usadas na análise exploratória de dados com geração automatizada de relatórios, sendo responsáveis pela descrição compacta de um conjunto de dados. A sumarização é utilizada, principalmente, no pré-processamento dos dados, quando valores inválidos são determinados por meio do cálculo de medidas estatísticas e, são de extrema

importância e imprescindíveis para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados.

e) **Modelo de Dependência**: apresenta possíveis dependências entre variáveis podendo ser estruturados (especificando as variáveis locais dependentes) ou quantitativos (apresentando o grau de dependência);

f) **Regras de Associação**: relações entre os atributos de um banco de dados. É um dos objetivos das pesquisas empíricas. Tem como objetivo localizar uma possível relação entre as variáveis do tipo se <condição> então <conseqüência>. Podem ser de natureza paramétrica ou não-paramétrica, dependendo da escala de mensuração;

g) **Análise de Séries Temporais**: possibilita a identificação de dependências de acordo com o tempo seqüencial. Seu objetivo é modelar o estado do processo extraindo e registrando desvios e tendências no tempo, sendo estas possíveis em: tendência, variações cíclicas, variações sazonais e variações irregulares e,

h) **Análise de Outliers**: apresentação de dados comportamentais diferentes da grande maioria. São as exceções. As práticas para detecção de fraudes podem se beneficiar com o uso desta técnica, pois, os eventos raros (ou exceções) podem ser mais interessantes do que os eventos regulares.

No que tange os requisitos dos dados para prática de *data mining*, é consenso para FAYYAD, 1996; MANNILA, 1996 e, BRAGA, 2005, que os dados devem possuir: a) acurácia (sem erros); b) consistência (fazem sentido); c) completude (sem campos ausentes); d) relevância (relativos ao problema); e, e) não redundantes (não duplicados).

Para o escopo desta pesquisa, faz-se válido a compreensão da complexidade do universo de técnicas e ferramentas para a realização da mineração de dados e o entendimento de que a prática em si é, obviamente, somente uma parte de todo o processo de mineração de dados.

3 PROCESSOS METODOLÓGICOS

Sob a perspectiva de que esta pesquisa contribuirá com uma discussão sobre as métricas e indicadores para a prática de mineração de dados, consequentemente, como forma de contribuição para a análise qualitativa, define-se que, inicialmente, a pesquisa está classificada em caráter exploratório quanto ao seu nível de pesquisa e a tipologia de abordagem sobre o problema proposto (ANDRADE, 1995; CIRIBELLI, 2003; GIL, 2009).

Pelo viés da finalidade, enquadra-se como uma ciência pura, pois, segundo Gil (2009), esta é caracterizada pela busca e pelo desenvolvimento de conhecimentos generalistas e auxiliadores na elaboração de teorias e leis.

Contudo, é de vital importância a menção de que pressupostos são considerados no desenvolvimento desta pesquisa, tendo em vista a importância na construção do conhecimento científico e na contribuição com a elaboração de uma discussão teórico não exaustiva relacionada ao assunto proposto nesta pesquisa.

Para tal, a fim de atingir o objetivo estabelecido, alguns objetivos específicos foram determinados. Baseados nestes, várias ações foram realizados para lograr êxito nos resultados esperados. Desta maneira, optou-se por listar as ações desenvolvidas para alcance de cada o objetivo específico estabelecidos.

3.1 MÉTRICAS E INDICADORES PARA O *GOOGLE ANALYTICS*

O primeiro objetivo específico definido foi: identificar na literatura quais são as métricas e os indicadores sugeridos para o *Google Analytics*.

Definiram-se quatro ações, sendo estas, relacionadas às estratégias para busca de informações na literatura e formatação dos resultados a fim de compilar as informações. Desta maneira, optou-se por: a) pesquisar nas seguintes fontes de informação - Portal de Periódicos Capes[12], BRAPCI[13], *Google Scholar*[14], *Google Books*[15].

[12] Disponível em: <http://www.periodicos.capes.gov.br/>
[13] Disponível em: <http://www.brapci.ufpr.br/>
[14] Disponível em: <http://scholar.google.com.br/>
[15] Disponível em: <http://books.google.com.br/>

Sabe-se que a fonte de informação *Google Scholar* pode também ser consultada diretamente pelo Portal de Periódicos CAPES, contudo, optou-se por deixar claramente explícita a busca direta à esta fonte de informação.

Estabelecidas as fontes de informação, definiu-se a estratégia de busca seguindo o padrão de: b) utilizar os descritores, suas possíveis associações (lógica booleana) e em três idiomas distintos (português, inglês e espanhol). A escolha pelos idiomas, Português, Inglês e Espanhol está associada as três línguas faladas na América Latina. Para a busca composta, optou-se pela busca apenas na língua portuguesa. A estratégia pode ser observado conforme abaixo:

Busca simples

Termo em Português	**Termo em Inglês**	**Termo em Espanhol**
	Google Analytics	
Métricas e Indicadores Web	*Web Metrics and Indicators*	*Web Métricas y Indicadores*

Busca avançada (somente em português)

Primeiro Termo	**Segundo Termo**	**Conjunção**
Métricas e Indicadores	*Google Analytics*	Métricas e Indicadores **AND** *Google Analytics*
Métricas e Indicadores	OJS	Métricas e Indicadores **AND** OJS
Google Analytics	OJS	*Google Analytics* **AND** OJS

Fonte: o autor.

Para atingir o máximo de recuperação de informação em cada fonte de informação, foi necessário seguir alguns procedimentos específicos. No caso do Portal de Periódicos Capes, em virtude da limitação de acesso a algumas bases que não possuem compatibilidade e interoperabilidade de busca, foi necessário acessar cada uma das bases e anotar os respectivos resultados alcançados. A Relação de Bases de Dados Indexadas no Portal de Periódicos Capes – área do conhecimento: Ciências Sociais Aplicadas (Apêndice B) foi elaborada para ter uma perspectiva das bases disponibilizadas para consulta.

Contudo, para melhor compreensão do método de consulta, e limitação do escopo desta pesquisa, parte-se do pressuposto a indicação do Portal de Periódicos Capes no que tange a busca nas bases de dados. Desta maneira, elaborou-se o quadro Resultados Quantitativos de busca no Portal De Períodicos Capes –

Temática Central (Apêndice C), o qual lista a base de dados consultada, o tipo de base – seguindo o critério de classificação – e o resultado quantitativo da busca.

Sendo assim, elaborou-se a tabela Resultados Quantitativos de busca simples e composta – temática central (Métricas e Indicadores do Google Analytics no OJS) (Tabela 1), o qual apresenta os resultados quantitativos encontrados nas fontes de informação definidas:

TABELA 1 – Resultados Quantitativos de busca simples e composta – temática central

Busca Simples			
Fonte de Informação	**Termo em Português**	**Termo em Inglês**	**Termo em Espanhol**
	Google Analytics	*Google Analytics*	*Google Analytics*
Portal de Periódicos CAPES	1721		
BRAPCI	1		
Google Scholar	324	26.900	441
Google Books	80	20.800	272

	Termo em Português	**Termo em Inglês**	**Termo em Espanhol**
Fonte de Informação	Métricas e Indicadores Web	*Web* Metrics *and Indicators*	*Web Métricas y Indicadores*
Portal de Periódicos CAPES	20	1.381.421	38
BRAPCI	2	3	2
Google Scholar	3.210	88.100	5.430
Google Books	21.200	4.330	153

Busca Avançada - somente em Português			
Fonte de Informação	Métricas e Indicadores **AND** *Google Analytics*	Métricas e Indicadores **AND** OJS	Google Analytics **AND** OJS
Portal de Periódicos CAPES	70	74	112
BRAPCI	0	0	0
Google Scholar	62	425	14
Google Books	2	1	3

Fonte: o autor.

Identificou-se que nas fontes de informação Google Scholar e Google Books, o processo para busca com os descritores em outros idiomas é diferente das fontes de informação Portal CAPES e BRAPCI. Nas fontes Google Scholar e Google

Books, deve-se escolher o idioma a ser buscado. Por isso, os resultados são diferenciados.

Observa-se um resultado expressivo, no que tange o aspecto quantitativo, na base de dados *Google Scholar*, contudo, constatou-se que existem documentos, artigos ou demais conteúdos informacionais que são apresentados em duplicidade.

Em continuidade, as outras duas ações necessárias para atingir o primeiro objetivo especifico são: c) anotar as informações relacionadas as denominações (em português ou em sua língua original - seja ela qual for), seu objetivo e a estrutura de cálculo; e, d) elaborar quadro informativo contendo as denominações, o objetivo e de que forma são calculados as métricas e os indicadores. Sendo esta última ação subdividida em: d.1) criar um quadro com 5 colunas, sendo uma para cada item (Nome da métrica/indicador - original e tradução, Objetivo e Estrutura de cálculo e a Conformidade com a definição do *Google Analytics*). O total de linhas corresponderá ao total de métricas e indicadores identificados na literatura e; d.2) usar técnicas de condensação da informação para desenvolver uma embalagem informacional de fácil compreensão. Tal quadro será apresentado no capítulo de resultados desta pesquisa (Capítulo 4).

Tendo em vista que esta pesquisa está classificada como uma produção científica acadêmica e, não é um estudo aprofundado e detalhado em todas as fontes de informação indexadas em bases relacionadas às áreas de Ciência da Informação, Administração e Tecnologia da Informação, não faz parte do escopo desta pesquisa à explanação exaustiva de todo um levantamento teórico acerta dos descritores propostos.

Sendo assim, assume-se como pressuposto os estudos de teóricos com representatividade significativa citados em monografias, teses e dissertações além de artigos e periódicos nas áreas da Ciência da Informação, Administração e Tecnologia da Informação. Pode-se dar como exemplo: CUTRONI, 2010.

Tendo pesquisado todas as possíveis métricas e indicadores que se podem mensurar com a prática da Webometria com a ferramenta *Google Analytics*, deu-se continuidade com as ações relacionadas aos outros objetivos específicos.

3.2 MÉTODOS E TÉCNICAS DE MINERAÇAO DE DADOS

O objetivo específico: identificar na literatura quais são os métodos e técnicas de mineração de dados que podem ser aplicados na base de dados gerados pelo *Google Analytics*; igualmente ao objetivo específico anterior e, baseada na discussão proposta por este trabalho, tem-se a necessidade de se buscar na literatura informações relacionadas aos métodos e técnicas de mineração de dados.

Para tal definiu-se a ação de: pesquisar nas seguintes fontes de informação: Portal de Periódicos Capes[16], BRAPCI[17], Google Scholar[18], Google Books[19]. Definidas as fontes de informação, ta como na estratégia anterior, optou-se por: utilizar os descritores, suas possíveis associações e em três idiomas distintos (português, inglês e espanhol). Sendo os descritores apresentados abaixo:

Busca simples

Termo em Português	Termo em Inglês	Termo em Espanhol
Mineração de dados	*Data mining*	*Minería de datos*

Busca avançada

Primeiro Termo	Segundo Termo	Conjunção
Mineração de dados	*Google Analytics*	Mineração de dados **AND** *Google Analytics*
Mineração de dados	*Google Analytics*	Mineração de dados **AND** *Google Analytics*
Mineração de dados	OJS	Mineração de dados **AND** OJS

Fonte: o autor.

Para melhor compreensão do método de consulta e limitação do escopo desta pesquisa, parte-se do pressuposto a indicação do Portal de Periódicos Capes no que tange a busca nas bases de dados. Desta maneira, elaborou-se um quadro com os Resultados Quantitativos de busca no Portal de Periódicos CAPES – Temática: Mineração de dados (APÊNDICE D) o qual lista a base de dados consultada, o tipo de base – seguindo o critério de classificação – e o resultado quantitativo da busca.

[16] Disponível em: <http://www.periodicos.capes.gov.br/>
[17] Disponível em: <http://www.brapci.ufpr.br/>
[18] Disponível em: <http://scholar.google.com.br/>

Definidos os procedimentos específicos para assegurar um resultado de busca de informação satisfatório, tem-se como resultado quantitativo desta busca os valores representados na tabela de Resultados Quantitativos de busca simples e composta – temática Mineração de dados (TABELA 2):

TABELA 2 - Resultados Quantitativos de busca simples e composta – temática Mineração de dados

Busca Simples			
	Termo em Português	**Termo em Inglês**	**Termo em Espanhol**
Fonte de Informação	Mineração de dados	*Data mining*	*Minería de datos*
Portal de Periódicos CAPES	95	642.914	543
BRAPCI	3	22	1
Google Scholar	25.200	1.920.000	9.390.000
Google Books	79.900	2.000.000	58.300

Busca Avançada - somente em Português			
Fonte de Informação	Mineração de dados **AND** *Google Analytics*	Mineração de dados **AND** *Google Analytics*	Mineração de dados **AND** OJS
Portal de Periódicos CAPES	140	139	161
BRAPCI	0	0	0
Google Scholar	23	23	666
Google Books	0	272	0

Fonte: o autor.

Observa-se que houve certa redução no total de artigos e publicações relacionadas aos termos pesquisados. Isto é fato em virtude da objetividade do objeto do estudo proposto para este trabalho.

Reforça-se que este trabalho parte dos pressupostos estabelecidos pelos mais renomados autores das áreas do conhecimento que são o pilar da Gestão da Informação, cabendo então a necessidade de explanar sobre os métodos e técnicas de mineração já consagradas pela comunidade acadêmica internacional.

Desta maneira, estabeleceram-se ação para: elaborar quadro informativo contendo a descrição, a finalidade e o algoritmo de cada método ou técnica de mineração de dados aplicável ao banco de dados do *plugin* e; criar um quadro com 5

[19] Disponível em: <http://books.google.com.br/>

colunas, Tarefa, Denominação, Descrição, Finalidade e Algoritmo.Tal quadro informativo é apresentado no Capítulo 4.

No que tange a área de Tecnologia da Informação, em específico a disciplina de mineração de dados, alguns autores contribuíram de maneira significante com a discussão dos métodos e das técnicas para a extração de conhecimento das bases de dados. Podem ser citados como exemplo os autores: Quinlan (1986), Cendrowska (1987), Fayyad (1996), Haykin (1999), Vasconcelos (2002), De Amo (2004) e Santos (2009).

Conhecidas as principais tarefas e seus respectivos métodos ou técnicas de mineração de dados consagrados pela literatura, o próximo passo para a construção metodologia da pesquisa diz respeito à seleção de atributos para cada métrica ou indicador disponível no *Google Analytics* é avaliado como um atributo passível de mineração.

3.3 A SELEÇÃO DE ATRIBUTOS

Seguindo a estratégia de elaborar ações para cada objetivo específico, para o objetivo "estabelecer os critérios para a seleção de atributos (feature selection) da base de dados gerada pelo *Google Analytics*", definiram-se como principais ações: a) pesquisar na literatura práticas para a elaboração de critérios de seleção de atributos de bases de dados em geral; b) estabelecer índices de acordo com a relação entre os atributos disponibilizados pelo *Google Analytics* e o algoritmo de seleção de atributo.

Dado isto, iniciou-se uma pesquisa, seguindo a mesma estratégia dos dois objetivos específicos anteriores. Utilizaram-se as mesmas fontes de informação (Portal de Periódicos Capes[20], BRAPCI[21], Google Scholar[22], Google Books[23]) e os seguintes descritores:

Busca simples

Termo em Português	Termo em Inglês	Termo em Espanhol
Seleção de Atributos	*Feature Selection*	*Selección de atributos*

Fonte: o autor.

[20] Disponível em: <http://www.periodicos.capes.gov.br/>
[21] Disponível em: <http://www.brapci.ufpr.br/>
[22] Disponível em: <http://scholar.google.com.br/>
[23] Disponível em: <http://books.google.com.br/>

Para melhor compreensão do método de consulta e limitação do escopo desta pesquisa, parte-se do pressuposto a indicação do Portal de Periódicos Capes no que tange a busca nas bases de dados. Desta maneira, elaborou-se o quadro com os Resultados Quantitativos de busca no Portal de Periódico CAPES – Temática: Seleção de Atributos (APÊNDICE E) o qual lista a base de dados consultada, o tipo de base – seguindo o critério de classificação – e o resultado quantitativo da busca.

Como resultado desta busca nas fontes de informação, obteve-se como resultado quantitativo os dados apresentados na tabela 3. Identifica-se uma volumetria predominância dos termos buscados na língua inglesa.

TABELA 3 - Resultados Quantitativos de busca simples e composta – temática Mineração de dados

Busca Simples			
	Termo em Português	**Termo em Inglês**	**Termo em Espanhol**
Fonte de Informação	Seleção de Atributos	*Feature Selection*	*Selección de atributos*
Portal de Periódicos CAPES	165	1.139.288	155
BRAPCI	0	2	0
Google Scholar	52.300	3.450.000	63.100
Google Books	10.200	1.250.000	55.700

Fonte: o autor.

Isto posto, identificaram-se autores com representativa significativa, sendo estes citados em diversos artigos e demais publicações como alguns exemplos: Han et. al. (1996), Liu e Motoda (1998) e PAPPA (2002).

Após a análise destes autores, e seguindo a estrutura inicialmente proposta nesta metodologia, o capítulo a seguir (Capítulo 4 – Resultados da Pesquisa) também apresenta os critérios de avaliação dos métodos para a fase de seleção de atributos.

4 RESULTADOS DA PESQUISA

Em conformidade com a metodologia estabelecida na pesquisa, esta seção tem como fundamental importância a apresentação dos resultados das ações estabelecidas para cada objetivo específico.

No que diz respeito as métricas e os indicadores, elaborou-se o quadro com uma compilação das definições estabelecidas pela Web Analytics Association (WAA) e disponíveis no *Google Analytics* (QUADRO 3) que, baseado na WAA que estipulou um documento oficial[24] (já traduzido para o português[25]), lista a definição das 19 (dezenove) métricas e indicadores possíveis da *web analytics* no *Google Analytics*. Apresenta-se neste quadro a denominação das métricas e dos indicadores (em sua forma original e sua respectiva tradução), uma breve descrição dos mesmos e sua estrutura lógica (estrutura de cálculo). Além disso, com base no estudo de Cutroni (2010), incluiu-se uma coluna que indica a conformidade do *Google Analytics* aos padrões da WAA:

[24] Disponível em: <http://www.webanalyticsassociation.org/resource/resmgr/PDF_standards/WebAnalyticsDefinitionsVol1.pdf>

[25] Tendo sua versão traduzida para o português pela *Interactive Advertising Bureau* Brasil (IAB Brasil). A versão foi revisada e foram incluídos mais termos e métricas neste documento. Disponível em: <http://www.iabbrasil.org.br/arquivos/doc/glossario_de_metricas.pdf>

QUADRO 3 – Principais métricas e indicadores definidos pela WAA e disponíveis no *Google Analytics*

Métrica ou indicador (Inglês)	Tradução para Português	Descrição e Objetivo	Estrutura de cálculo	Conformidade do *Google Analytics* (CUTRONI, 2010)
***Building Block Metrics* (Métricas fundamentais/essenciais)**				
Page	Página	É uma unidade de conteúdo, ou seja, são as informações (textuais ou não) que são apresentadas pelo navegador quando é acessado um endereço na Internet. Cada ferramenta de monitoramente permite que o analista de métricas indique quais os arquivos que devem ser considerados "páginas", isto é, arquivos Flash, AJAX, downloads, documentos etc.	*Page* = total de todos os elementos textuais ou não textuais	Igual à da WAA.
Page view	Página vista	Indica a quantidade de vezes que uma página foi visualizada por um usuário. Não são considerados como *page view* as falhas de solicitação (code 400-499) e erros no servidor (code 500-599), contudo, se essas exceções forem direcionadas para apresentar outra página, essa página será considerada.	*Page view* = soma de todas as *pages* visualizadas – soma do total dos erros das *pages* visualizadas.	Igual à da WAA. Sendo que, qualquer valor passado ao método _trackPageview() aparecerá nos relatórios como uma página.
Visits / Sessions	Visitas ou sessões	Toda visita é uma interação feita por usuário em um website consistindo em uma ou mais requisições por uma unidade de conteúdo (*page* ou *page views*). O tempo de visitação é considerado somente quando há atividade (interações do usuário) com as unidades de conteúdo. Sendo assim, caso não haja interação (chamado período de inatividade) a sessão será encerrada. Além disso, se um usuário acessar a mesma página em um intervalo menor de 30 minutos deverá ser contabilizado apenas uma visita.	*Visits/Sessions* = total de interação com *pages* ou *page views* | *visitas / sessões* >= 30 min.	Igual à da WAA. Para o *Google Analytics* como *default*, uma visita / sessão será encerrada passado o tempo de 30 minutos de inatividade. Contudo, esse parâmetro pode ser alterado pelo analista de métricas caso necessário.

Métrica ou indicador (Inglês)	Tradução para Português	Descrição e Objetivo	Estrutura de cálculo	Conformidade do *Google Analytics* (CUTRONI, 2010)
Unique visitors	Visitantes únicos	Representa o número deduzido de pessoas individuais (excluindo *spiders* e *robots*) que praticam atividades consistindo em uma ou mais visitas a um site, em um determinado tempo estabelecido. Isso corre pelo armazenamento de dados nos *cookies* de monitoramento do *Google Analytics*. Contudo, há de se verificar as questões relacionadas à exclusão e bloqueio dos *cookies*, ambas as ações passíveis de serem realizadas pelo usuário.	*Unique Visitors* = total de usuários individuais – o total de *spiders* e *robots*	Igual à da WAA. Sendo que, um visitante é definido pelo ID único criado quando ocorre a primeira visita do usuário. Desta forma, cada visitante é contabilizado apenas uma vez na métrica de visitantes únicos, independente da quantidade de vezes retornou ao site durante o período definido pelo analista de métricas.
New visitors	Visitantes novos	Corresponde ao número de visitantes únicos que realizaram a primeira visita ao site durante um período estipulado. Não é possível que um mesmo usuário seja considerado como novo visitante e ao mesmo tempo visitante de retorno em um mesmo período estipulado.	New Visitors = total de novos acessos de usuários individuais | *new visitor <> return visitor*	Igual à da WAA. Ainda que o *Google Analytics* compartilhe da mesma definição para um visitante novo, ele não conta o número de novas pessoas individuais (visitantes) que visitaram o site durante o período reportado. O *Google Analytics* conta o número de visitas geradas por pessoas novas. Calcula o número de visitantes novos identificando o número de visitantes novos identificando o número de IDs de visitantes únicos novos criados durante o período reportado. É possível medir o número de visitantes novos utilizando um perfil e um filtro de inclusão.

Métrica ou indicador (Inglês)	Tradução para Português	Descrição e Objetivo	Estrutura de cálculo	Conformidade do *Google Analytics* (CUTRONI, 2010)
Return visitor	Visitante em retorno	Baseado no histórico de visitas anteriores, esta métrica corresponde ao número de visitantes únicos com atividade realizada em uma visita a um site em um determinado período e, quando este mesmo visitante também visitou o mesmo site antes do período de tempo estipulado. Para o cálculo da métrica, não se pode considerar como visitante em retorno e visitante novo em um mesmo período de tempo estipulado.	*Return visitor* = total de usuários com duas ou mais *visits* em vários períodos de tempo.	Igual à da WAA. Ainda que o *Google Analytics* compartilhe da mesma definição de um visitante em retorno, ele não conta o número de pessoas em retorno (visitantes) que visitaram o site durante o período reportado. O *Google Analytics* conta o número de visitas geradas. Ele define um visitante em retorno como qualquer visitante cujo *cookie* identificador individual foi definido antes do período reportado.
***Visit Characterization* (Caracterização da visita)**				
Entry Page	Página de entrada	É a primeira página de uma *visit*. Cada visita contém pelo menos uma página, logo, o número total de páginas de entrada é igual ao número total de visitas para qualquer período de tempo estipulado.	*Entry Page* = total de views da primeira página	Igual à da WAA.
Landing Page	Página de destino	Utilizado para estratégicas de marketing, esta métrica corresponde a uma página para a identificação da experiência do usuário resultante de um esforço de marketing. Esta métrica não deve ser confundida com a métrica anterior, pois uma página de destino não é, necessariamente, a página de entrada embora pudesse ser.	*Landing Page* = total de *views* ou *clicks* em um esforço de marketing	Igual à da WAA.
Exit Page	Página de Saída	Representa a última página acessada em um site durante uma visita ou sessão, representando o fim da visita ou sessão. Essa métrica utiliza o processo de rastreamento de *cookies* de sessões para efetivar sua confiabilidade.	*Exit Page* = total de *views* de cada última página vista por *visit*	Igual à da WAA.

Métrica ou indicador (Inglês)	Tradução para Português	Descrição e Objetivo	Estrutura de cálculo	Conformidade do *Google Analytics* (CUTRONI, 2010)
Visit duration	Duração da Visita	Corresponde ao tempo total em que um usuário permaneceu em uma sessão. Para o cálculo da métrica considera-se que se, ocorrer somente uma atividade única ou um único evento sem duração de visita, isso será reportado e contabilizado.	*Visit duration* = tempo total no site nº total de *visits*	Igual à da WAA. O *Google Analytics* chama essa métrica de *average time on site* (tempo médio no site), a qual é calculada dividindo o tempo total gasto no site pelo número total de visitas.
Link referrer	Link de Referência	O link de referência é o URL da página que originalmente gerou a requisição para o pageview ou objeto atual. Para a maior parte das ferramentas de *web analytics*, o *referrer link* é o conteúdo (URL) informado no cabeçalho da solicitação feito por num navegador ao servidor. Ressalta-se que para alguns casos, esse conteúdo pode ser vazio ou nulo, sendo estes, apresentados como "*No referrer*" ou "*Direct Navigation*". Esta informação é de extrema importância para identificar a origem da visita de um usuário. Além disso, a URL de referência pode ter associados a ela vários parâmetros que auxiliam em uma análise mais detalhada.	*Link referrer* = link de origem x total de *view* deste link	O link de referência no *Google Analytics* é o URL da página que originalmente gerou a requisição para a visita atual. Esse valor é depois somado a todos os *page views* nesta visita. O link de referência é identificado no *Google Analytics* como qualquer origem cuja mídia seja *referral* (Referência). O *Google Analytics* também tem um campo chamado *referral* que está em conformidade à definição de *referrer* da WAA.

Métrica ou indicador (Inglês)	Tradução para Português	Descrição e Objetivo	Estrutura de cálculo	Conformidade do *Google Analytics* (CUTRONI, 2010)
Visit referrer	Link de Referência de visita	O link de referência de visita é o primeiro *referrer* em uma sessão, seja interno,externo ou nulo.	*Visit referrer* = total de "impressões" (exibições) de link.	Igual à da WAA. Esse dado é chamado de *referral* no *Google Analytics* e pode ser apenas o *referrer* externo.
Click-through	*Click-through*	O número de vezes que um visitante clicou um link em particular. O *Click-through* pode ser considerado um banner com propaganda (que ao clicado direcionará o usuário a uma página de um website externo) ou um banner informativo com notícias (que ao clicado direcionará o usuário uma página interna do website).	*Click-through* = total de *clicks* em um determinado link em específico.	Igual à da WAA. O *Google Analytics* se refere aos *Click-throughs* como *clicks*. Essa métrica encontra-se disponível apenas em relatórios do *AdWords*.
Clickt-hrough rate/ratio	Taxa de *Click-through*	O número de *Click-through* para um link específico dividido pelo número de vezes que esse link foi visualizado. Logo, para se verificar a eficiência de uma campanha publicitária, por exemplo, deve-se dividir a quantidade de vezes que o "banner" foi clicado pela quantidade de views que esse banner teve. Tanto a métrica *Click-through* quanto a métrica View são medidos para um período de tempo estipulado.	*Clickt-hrough rate/ratio* = $\frac{\text{total de } clicks \text{ em um link específico}}{\text{número de "impressões"(exibições) do link}}$	Igual à da WAA. O *Click-through* e a taxa de *Click-through* são porcentagens de impressões que resultaram em um clique. São calculados dividindo o número de cliques em uma peça de publicidade pelo número de impressões da mesma peça. Essa métrica encontra-se disponível apenas nos relatórios do AdWords.

Métrica ou indicador (Inglês)	Tradução para Português	Descrição e Objetivo	Estrutura de cálculo	Conformidade do *Google Analytics* (CUTRONI, 2010)
Page views per visit	Page views por visita	O número de *page views* em um período reportado dividido pelo número de visitas no mesmo período, isto é, indica a média de visualizações de página (*page views*) por cada visita.	*Page views per visit* = número de *page views* / total de *visits*	Igual à da WAA.
***Content Characterization* (Caracterização do conteúdo)**				
Page exit rate	Taxa de saída de página	O número de saídas de uma página dividido pelo número total de *page views* para essa página. A taxa de saída de página não deve ser confundida com a taxa de rejeição, que é um indicador de visitas com uma única visualização de página no site. Além disso, a taxa de saída de página se aplica a todas as visitas, independente de sua duração. Salienta-se que algumas ferramentas de web *analytics* podem calcular a taxa de saída de página utilizando a métrica de visitas como denominador, ao invés de utilizar as visualizações de páginas. Contudo, indica-se que a contagem por *page views* seja um denominador mais adequado, pois um visitante pode entrar na mesma página várias vezes durante uma visita.	*Page exit rate* = total de uma *page exit* / total de *views* dessa mesma *page*	Igual à da WAA.

Métrica ou indicador (Inglês)	Tradução para Português	Descrição e Objetivo	Estrutura de cálculo	Conformidade do *Google Analytics* (CUTRONI, 2010)
Single-page view visits (bounces)	Visitas de page view único (Rejeições)	Visitas que consistem em um *page view*. Isto quer dizer que, cada page view única é a página de entrada e de saída.	*Single-page view visits (bounces)* = Total de *views* de uma página única	Igual à da WAA. A contagem de *bounces* pode ser modificada por outros recursos do *Google Analytics*, especificamente pelo monitoramento de eventos. Quando este é utilizado, o código de monitoramento do *Google Analytics* solicitará o *gif* invisível do servidor do *Google Analytics*. O *Google Analytics* interpretará essa solicitação pelo *.gif* como uma ação do visitante e concluirá que o visitante interagiu com a página web, não mais o contando como um *bounce*.
Bounce rate	Taxa de rejeição	Visitas de *page view* único divididas por páginas de entrada. Se a taxa de rejeição é calculada para uma página específica, então a métrica será o número de vezes que a página foi uma exibição única dividido pelo número de vezes que essa página foi uma entrada. Outra forma de cálculo é se, a taxa de rejeição for calculada para um grupo de páginas, então a métrica é calculada baseada no número de vezes em que o grupo de páginas foi uma visita de página única, dividido pelo número de vezes em que esse grupo de páginas foi à página de entrada. Por padrão das ferramentas de Web *Analytics*, esta métrica aponta a quantidade de usuários que entraram no site e saíram do site sem fazer nenhuma interação, ou seja, não clicaram em nenhum link no site. Esta métrica deve ser analisada junto com os relatórios de *entry page* e *landing page*.	*Bounce rate* = total exibições de uma *page* única total de vezes em que esta página foi uma *Entry Page*	Igual à da WAA.

Métrica ou indicador (Inglês)	Tradução para Português	Descrição e Objetivo	Estrutura de cálculo	Conformidade do *Google Analytics* (CUTRONI, 2010)
***Conversion Metrics* (Métricas de conversão)**				
Event	Evento	Os eventos são atividades que acontecem dentro de uma página, por exemplo: impressões dos anúncios, início e conclusão de uma transação, mudança dos campos de um formulário, exibições multimídia, entre outros. Os eventos também podem ser associados com as tecnologias de web avançadas, tais como Ajax e Flash. Qualquer ação registrada ou gravada que tem uma data e hora específica atribuídas a ela, seja pelo navegador,seja pelo servidor. Para análise desta métrica deve-se identificar que: a contagem de eventos é maior ou igual à contagem de visita, sendo está maior ou igual ao número de visitantes.	*Event* = total de ações m um determinado site (sendo maior ou igual ao total de *views*)	Igual à da WAA. Há múltiplos atributos para um evento no *Google Analytics*, como objetos, ações e marcadores. O monitoramento de eventos é um recurso beta do *Google Analytics* que talvez não para muitas contas de monitoramento.
Conversion rate	Taxa de conversão	É um percentual de um grupo (de visitas ou visitantes), que teve uma ação específica de interesse dentro do site. Essa conversão pode abranger toda a população de visitas do site, como a porcentagem de visitas que concluiu um determinado registro em relação ao total de visitas dentro do site.	*Conversion rate* = $\frac{\text{total de visitas que resultam em uma ação}}{\text{total de } views}$	Igual à da WAA. Além de conversões, o *Google Analytics* também calculará a taxa de conversão. A taxa de conversão é o número total de visitas que resultam em uma ação desejada, dividido pelo número total de visitas. Uma conversão será registrada apenas uma vez por visita. Visitantes não podem converter mais de uma vez por visita.

Fonte: baseado em CUTRONI (2010) e WAA.

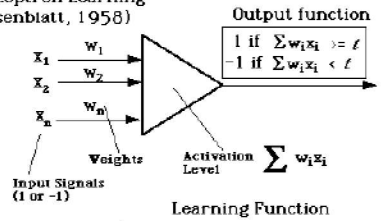
Considerando que os periódicos científicos, em sua versão online, seguem a mesma estrutura de um websites, sendo a *homepage* o índice do periódico e as *pages* os artigos deste periódico, pode-se definir que todas as métricas indicadas do quadro acima são, potencialmente, passiveis de análise nos periódicos científicos que instalaram o *plugin* do *Google Analytics* em sua ferramenta de gestão editorial OJS.

Após o levantamento das métricas e dos indicadores do *Google Analytics*, estudaram-se os autores mais consagrados na área de Mineração de Dados e, tem-se como próximo resultado da pesquisa o QUADRO 4 que apresenta as principais tarefas e seus respectivos métodos ou técnicas de mineração de dados que são consagrados pela literatura já listada neste relatório de pesquisa.

O QUADRO 4 apresenta os três tipos de tarefas comumente utilizadas em pesquisas acadêmicas, as tarefas divididas em técnicas, sua respectiva descrição, sua finalidade e o(s) algoritmo(s) mais utilizados.

QUADRO 4 - Tarefas e suas respectivas técnicas de mineração de dados

Tarefa	Técnica	Descrição	Finalidade	Algoritmo(s) consagrado(s) pela literatura
Classificação	Árvores de Decisão	São estruturas de dados compostas de um nó raiz e vários nós subordinados (filhos), que por sua vez tem seus filhos e que se interligam por ramos. Cada qual representa uma regra. As folhas são os nós que não possuem filhos e os que têm são chamados de nós pais ou de decisão.	Descobrir uma função preditiva que consegue classificar um dado em uma classe de várias classes discretas que são pré-definidas ou conhecidas. As árvores de decisão são uma ferramenta útil para selecionar as variáveis que são realmente úteis para prever o valor de uma variável resposta.	**ID3** (Cálculo da Entropia e Cálculo do Ganho de Informação). **Entropia**: $H(X) = \sum p(x_i) \log_2 1/p(x_i)$ **Ganho de informação**: $I(x_i) = \log_2 1/p(x_i)$ (QUINLAN, 1986)
	Regras de classificação	As Regras de Classificação podem ser derivadas das árvores de decisão ou de redes neurais. Formulam-se com as idéias de "SE.... ENTÃO" ou "SE ... E... ENTÃO".	Identificar, entre um conjunto pré-definido de classes, aquela a qual pertence um elemento, a partir de seus atributos.	***Prism*** (Entropia e Ponderação) **Entropia**: $H = -\sum_i p(\delta_i) \log_2 p(\delta_i)$ bits. (CENDROWSKA, 1987) **Ponderação**: $Pontuação = \frac{p}{p+n}$ *Equação 1* (VASCONCELOS, 2002)

	Redes Neurais	As redes neuras são procedimentos computacionais direcionados que envolvem o desenvolvimento de estruturas matemáticas com habilidade de aprendizado.	"Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processos simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso" (HAYKIN, 1999).	Perceptron Learning (Rosenblatt, 1958)  ROSENBLATT, 1958 apud HAYKIN, 1999)
Tarefa	**Técnica**	**Descrição**	**Finalidade**	**Algoritmo(s) consagrado(s) pela literatura**
Associação	Descoberta de regras	Para esta técnica duas métricas são consideradas: suporte e confiança. Para a regra "Se X ocorre então Y também ocorre", o suporte da regra é calculado como o valor do total de eventos ou casos onde X e Y aparecem, dividido pelo valor total de casos da base de dados. O suporte indica o quanto a regra de associação é significativa em relação à base de dados. A métrica confiança é calculada como o número de eventos ou casos onde X e Y aparecem, dividido pelo número de eventos onde X aparece, e indica o quanto Y é relacionado com X (SANTOS, 2009).	Identificar regras do tipo Se X ocorre então Y também ocorre, onde X e Y pode ser qualquer tipo de parâmetros, o qual pode ocorrer também combinações dos parâmetros.	***Apriori*** (Suporte e Confiança) Em uma regra do tipo Se X então Y **Suporte**: $Sup(X,Y) = \frac{Nº\ de\ casos\ que\ possuem\ X\ e\ Y}{N^{\circ}\ total\ de\ registros} \geq SupMin$ SupMin = limite mínimo de suporte **Confiança** $Conf(X,Y) = \frac{Nº\ de\ casos\ que\ possuem\ X\ e\ Y}{Nº\ de\ registros\ que\ possuem\ X} \geq ConfMin$ ConfMin = limite mínimo de confiança (AGRAWAL; SRIKANT, 1994)

Agrupamento ou *Clustering*	Descoberta de *clusters*	Técnica utilizada para realizar a descoberta de grupos de dados que possivelmente indicam semelhança entre os eles. Estes podem ser considerados suficientemente parecidos ou diferentes em N grupos distintos. Tal técnica difere-se da técnica de classificação por não existirem classes ou valores pré-definidos. Contudo, os algoritmos de agrupamento formam os grupos de acordo com métricas, para que possam ser processados posteriormente como objetos correspondendo à mesma categoria (SANTOS, 2009).	Identificar agrupamentos de objetos, sendo estes responsáveis por identificar classes. *Clustering* constitui uma tarefa de aprendizado por observação ao contrário da tarefa de Classificação que é um aprendizado por exemplo (AMO, 2004).	*K-Means*: $SQ\,\mathrm{Re}s(j) = \sum_{i=1}^{n_j} d^2[o_i(j), \bar{o}(j)]$ (De AMO, 2004).

Fonte: baseado nos autores identificados na revisão de literatura.

Conhecidas as métricas e os indicadores e, tendo o conhecimento de todas as tarefas e algoritmos mais utilizados no campo da mineração de dados, há a necessidade de se apresentar o resultado do estudo do processo para a indicação dos modelos para seleção de atributos.

Para tal, conforme apresentado na metodologia da pesquisa, baseado na literatura pertinente conceitua-se que o processo de seleção de atributos é uma etapa comumente utilizada no aprendizado de máquina ou a mineração de dados. O objetivo da seleção de atributos é identificar subconjuntos de parâmetros ou dimensões que possuem maior significância e que garantam precisão nos resultados na extração de conhecimento de bases de dados.

A seleção de atributos é um estudo praticado em todas as áreas da mineração de dados, em específico para o reconhecimento de padrões desde a década de 70 conforme estudo apresentado por Mucciard e Gose (1971). Contudo, Han et al. (1996) comentam que a seleção de atributos é um do grandes desafios para a área de mineração de dados. Haja visto que, a seleção de atributos tem como finalidade a garantia de que os dados que chegam a fase de mineração de dados sejam considerados de boa qualidade (LIU; MOTODA, 1998 apud PAPPA, 2002).

Para Dash e Liu (1997) apud PAPPA (2002), existem quatro definições diferentes para a seleção de atributos. São elas:

a) idealizada: tem como objetivo encontrar o menor subconjunto de atributos suficientes para descrever o conceito proposto (Kira; Rendell, 1992 apud PAPPA, 2002);
b) clássica: tem como objetivo selecionar um subconjunto de X atributos a partir de um conjunto de Y atributos, sendo o módulo de X menor que o módulo de Y. Desta maneira, o subconjunto X é o melhor subconjunto dentre os possíveis com tamanho X (Narendra; Fukunaga, 1977 apud PAPPA, 2002);
c) melhora da taxa de precisão: tem como objetivo selecionar um subconjunto de atributos para aumentar a precisão da classificação ou reduzir o conjunto de atributos inicial (sem interferir na taxa de precisão) em relação ao conjunto de todos os atributos (Koller; Sahami, 1996 apud PAPPA, 2002) e,

d) aproximação da distribuição original das classes: tem como objetivo selecionar um subconjunto pequeno de atributos de forma que a distribuição das classes utilizando apenas estes atributos seja a mais próxima possível os distribuição utilizando todos os atributos (Koller, Sahami, 1996 apud PAPPA, 2002).

Para o escopo desta pesquisa científica, considera-se pertinente a discussão dos algoritmos utilizados para a realização da seleção de atributos. Para tal, apresentam-se no QUADRO 5, os algoritmos divididos por suas respectivas fases (busca ou avaliação de subconjuntos).

QUADRO 5 - Fases e algoritmo para o processo de Seleção de Atributos

Fase	Denominação do algoritmo	Finalidade
Busca de subconjuntos	Algoritmos exponenciais	Estes algoritmos procuram fazer de maneira exaustiva todas as combinações de atributos possíveis antes de retornar um subconjunto de atributos. Contudo, normalmente, estes subconjuntos são inviáveis computacionalmente, pois o tempo de execução cresce exponencialmente (LIU; MOTODA, 1998 apud PAPPA, 2002).
	Algoritmos seqüenciais	Estes algoritmos são subdivididos em: seqüencial para frente – inicia busca pelo melhor subconjunto de atributos com conjunto vazio, em seguida busca por subconjuntos com um atributo e este será comparado com os demais e adicionando um atributo por vez até não se conseguir a qualidade do subconjunto – e, seqüencial para trás – inicia por um subconjunto de atributos ideais com uma solução representando todos os atributos e a cada iteração um atributo é removido até que isso não seja mais possível e encontrar a qualidade da solução (PAPPA, 2002)
	Algoritmos randômicos	Os algoritmos genéticos são considerados exemplos pertinentes para os métodos de busca randômicos, sendo que sua principal vantagem sobre métodos seqüenciais é que este trata o problema de interação entre atributos (FREITAS, 2001 apud PAPPA, 2002).
Avaliação de subconjuntos	Wrapper	A abordagem wrapper define um subconjunto ótimo de soluções de acordo com uma base de dados e algoritmo de indução particular, levando em conta a tendência indutiva do algoritmo e sua interação com o conjunto de treinamento. Esta abordagem, normalmente, aumenta consideravelmente o tempo de execução do algoritmo, mas a precisão preditiva obtida tende a ser superior àquela obtida pela abordagem filtro (PAPPA, 2002).
	Filtro	A abordagem filtro, ao contrário da *wrapper*, procura escolher um subconjunto de atributos independente do algoritmo de classificação, estimando a qualidade dos atributos apenas em relação aos dados (PAPPA, 2002)

Fonte: baseado em PAPPA (2002)

O processo de seleção de atributos é particularmente importante em situações em que instâncias são descritas usando um número grande de atributos e se desconhece a relevância de cada um desses atributos na representação do conceito. Logo, tendo conhecimento dos algoritmos disponíveis para a seleção dos atributos, podem ser aplicados os índices baseada nas necessidades desta pesquisa.

Desta forma, definiu-se que os critérios a serem avaliados devem ser: os tipos de dados, a dimensão dos dados, a presença de ruídos, o tempo de execução e a viabilidade computacional. Atribuíram-se valores de 1 a 5, sendo 1 para menos adequado à 5 como sendo o algoritmo de acordo com o escopo desta pesquisa.

TABELA 4 – Avaliação dos critérios para escolha do modelo de algoritmo para Seleção de Atributos

Fase	Denominação do algoritmo	Critérios					
		Tipos de dados	Dimensão dos dados	Presença de ruído	Tempo de execução	Viabilidade Computacional	Peso médio
Busca de subconjuntos	Algoritmos exponenciais	4	2	2	1	1	2
	Algoritmos seqüenciais	2	2	5	2	3	2.8
	Algoritmos randômicos	4	4	3	3	4	3.6
Avaliação de subconjuntos	*Wrapper*	2	2	5	2	2	2.6
	Filtro	4	3	4	3	4	3.6

Fonte: baseado em SANTORO, 2005.

A revisão dos processos de seleção de atributos, é dado o momento de iniciar a explicação da justificativa da escolha do método ou técnica de mineração de dados e o(s) algoritmo(s) indicado para a prática de descoberta de conhecimento em bases de dados. Assim sendo, esta seção caracteriza o último objetivo específico proposto: justificar a escolha de um método ou técnica de mineração de dados e do(s) algoritmo(s) de seleção de atributos que caracteriza-se como ideal para ser aplicado para a descoberta de conhecimento na base de dados gerada pelo *Google*

Analytics com os dados advindos de um periódico científico que utiliza o OJS. Ressalta-se que nesta pesquisa científica não foram realizados experimentos práticos, logo, tem-se uma discussão teórica sobre o tema proposto.

Pela característica dos dados que podem ser disponibilizados pelo *Google Analytics* (puramente numéricos) a discussão sobre a decisão de qual a tarefa e, conseqüentemente, qual a técnica de mineração é considerada exaustiva, pois, para todas as tarefas será necessário um esforço significativo no processo de limpeza dos dados, ou seja, para que se tenham árvores de decisão, classes ou núcleos, regras de associação, entre outros possíveis resultados de mineração, far-se-á necessário que os dados sejam "padronizados". Desta forma, poder-se-á garantir a acurácia dos resultados.

Sendo assim, baseado nos casos apresentados na literatura, tem-se duas possíveis tarefas – e técnicas – plausíveis de serem aplicadas: a tarefa de associação (técnica de mineração APRIORI) e a tarefa de agrupamento ou *clustering* (técnica de mineração K-means).

Indica-se a tarefa de associação pelo real objetivo da descoberta de conhecimento com os dados do *Google Analytics*, ou seja, encontrar regras do tipo: **"Usuários que acessam o periódico às terças-feiras no período noturno têm uma taxa de rejeição anormal"**, ou, **"Usuários da Região Sul que acessam o periódico no primeiro semestre do ano procuram artigos relacionados a um tema Y"**. No primeiro caso, seria possível indicar aos administradores do periódico científico que investigassem qual o motivo das rejeições que ocorrem nas noites de terça-feira e formular hipóteses, por exemplo, "o servidor onde está hospedado o período tem algum processo que é executado as terças-feiras que ocasiona uma lentidão no acesso ao periódico?". No segundo caso, tem-se uma possibilidade de identificar grupos de interesse de uma região em um tema específico.

Algo similar ao segundo caso poder-se-á fazer com a aplicação da tarefa de agrupamento ou *clustering*, pois, seria possível identificar quais os grupos de usuários que possuem padrões de acesso semelhantes. Isto é, **"os usuários com perfil de leitores da região sul acessam os mesmos artigos publicados no periódico que os usuários com perfil de autores da região norte"**.

Descarta-se a indicação dos métodos ou das técnicas de redes neurais e árvore de decisão tendo em vista que, no caso das redes neurais, o objetivo da técnica é a aprendizagem de máquina de maneira inteligente, o que, a princípio, não

possui correlação com os objetivos da análise dos dados da base do *Google Analytics*. Para o caso das árvores de decisão, sua aplicação está associada a questões preditivas, isto é, seu objetivo está em descobrir quais os valores (ou comportamentos) possíveis que um determinado usuário teria em relação a um cenário. Quando aplicado um algoritmo para gerar uma árvore de decisão, tem-se a necessidade de se estabelecer um "atributo meta" (uma variável única definida como objeto de comparação). No caso específico desta pesquisa, considera-se que todas as métricas e os indicadores podem ser potenciais "atributos meta", logo, torna-se difícil a decisão de se definir uma única métrica ou um indicador para a execução de algoritmos pertencentes a tarefa de Classificação. Desta maneira, os algoritmos para geração de redes neurais e árvores de decisão são desaconselhados para o escopo desta pesquisa.

Não se pode afirmar com precisão que a prática de classificação, em especifica o método de regras de classificação, pode ser aplicável ao objeto da pesquisa, pois, tem-se como pressuposto que este método é um advindo dos resultados de análises das árvores de decisão e das redes neurais. Contudo, o método é um tanto quanto semelhante ao APRIORI, pois é possível obter resultados do tipo: "**SE os usuários da região sul que acessam o periódico no primeiro semestre do ano ENTÃO procuram artigos relacionados a um tema Y"**. Logo, esta prática deve ser analisada isoladamente em outros trabalhos.

Em relação ao processo de seleção de atributos, a explicação inicia-se pela crítica ao processo de busca, em específico aos algoritmos caracterizados como exponenciais. Ressalta-se que se utilizou das experiências teóricas do estudo realizado por Santoro (2005) para se efetuar análise dos tipos de algoritmos para a seleção de atributos.

No que tange o critério dos tipos de dados que podem ser processados por estes algoritmos, indicou-se que não há restrições de atributos para que o processo de busca dos subconjuntos ótimos seja executado. Contudo, para os critérios de dimensão dos dados, presença de ruído, tempo de execução e viabilidade computacional, baseado nos relatos de Santoro (2005), acredita-se que estes algoritmos demandam grande capacidade de processo computacional ou que o volume de atributos seja menor. Sabendo-se que o número de atributos proporcionados pelo *Google Analytics* é de no mínimo 19 (dezenove) métricas ou

indicadores, sendo assim, atribui-se o peso médio 2 e, consequentemente, descarta-se a possibilidade de utilização deste tipo de algoritmo.

Para os algoritmos seqüenciais foi-se atribuído o peso médio 2.8, pois, nos critérios de tipologia de dados e tempo de execução, de acordo com a literatura, este tipo de algoritmo é específico para um grupo de dados e seu tempo de processo é considerado alto pois, existem duas heurísticas para a busca do subconjunto ótimo. Isto não torna o algoritmo computacionalmente inviável, pois, o foco deste algoritmo é a qualidade da precisão dos atributos, o qual no critério ruído recebeu o maior índice. Entretanto, descarta-se a escolha deste tipo de algoritmo pelo escopo desta pesquisa.

O último algoritmo da fase de busca de subconjuntos são os algoritmos randômicos. Para este tipo de algoritmos atribuiu-se o peso médio de 3.8 pois, considera-se que sua capacidade de resolver os problemas de interação entre os atributos e seu considerável tempo para a execução da heurística, este tipo de algoritmos (como exemplo os algoritmos genéticos) podem ser aplicáveis para esta pesquisa científica. Além disso, faz-se destaque para os critérios dimensões dos dados e presença de ruídos, tendo vista que estes algoritmos também visam a qualidade da precisão dos atributos.

No que diz respeito à fase de avaliação dos subconjuntos, os algoritmos do tipo *wrapper* receberam um peso médio de 2.6, haja visto que seu tempo para execução da heurística, sua viabilidade computacional e os critérios de tipologia de dados interferem significativamente nos resultados. Além disso, sua qualidade está diretamente relacionada ao algoritmo de mineração que será executado. Logo, descarta-se como algoritmo ideal para o escopo desta pesquisa.

Por fim, avaliou-se os algoritmos do tipo filtro e, conforme apresentado por Santoro (2005), os algoritmos associados a este modelo tem pouco consumo computacional e são ideais para diferentes métodos ou técnicas de mineração de dados. É por este motivo que sua prática e adesão em várias pesquisas da área é popular. Para tal grupo de algoritmos atribuiu-se o peso médio 4, pois, a tipologia dos dados não interfere na qualidade dos subconjuntos dos atributos e, o tempo para execução bem como a viabilidade computacional receberam pesos 5 e 4, respectivamente.

Sendo assim, acredita-se que para o escopo desta pesquisa científica, deva-se utilizar, na fase de busca de subconjuntos, os algoritmos randômicos (ou genéticos) e, na fase de avaliação dos subconjuntos, os algoritmos de filtro.

Obviamente que não se descartam possibilidades diferentes do que esta aqui proposta, pois, os resultados dependeram especificamente do objetivo da prática de mineração de dados e de quais são os recursos computacionais e de tipologia de dados disponibilizados pela base em questão.

Em relação a descoberta de padrões de comportamento freqüentes no processo de navegação dos usuários em periódicos científicos que utilizam o OJS, a análise é válida para tomar decisões no que diz respeito a arquitetura, gestão da informação e campanhas de marketing. Todas as ações dos usuários são armazenadas em arquivos de *logs*, em servidores especiais, sendo um deles, o *Google Analytics*.

Para Hay et al. (2003) existem três tipos de mineração de dados da Web, sendo elas: a mineração de conteúdos da web – sendo a extração do conhecimento dos documentos e dos metadados; a mineração de estrutura da web – que é a descoberta de conhecimento por meio das ligações entre documentos da web, e a mineração de uso da web – a análise dos dados coletados de acordo com o comportamento dos usuários que tem a intenção de descobrir padrões de acesso e melhorar a qualidade da experiência do usuário ou para modelar o comportamento dos mesmos.

Considera-se que o que é proposto nesta pesquisa é justamente uma prática de mineração de uso da Web, pois, de acordo com a literatura, este trabalho tem a intenção de proporcionar informações importantes sobre o comportamento dos usuários.

Dando ênfase e justificando os processos metodológicos adotados nesta pesquisa, Santos (2004) indica que para que os resultados da mineração de dados sejam com acurácia, parte-se do pressuposto que os dados a ser minerados estão estruturados de forma relacional, ou seja, em tabelas organizadas em linhas (dados) colunas (atributos). Logo, a prática da limpeza dos dados e a seleção de atributos são de fundamental importância para a extração de conhecimento em bases de dados.

Para Anand et. al. (2003) as principais aplicações da mineração de uso da Web são a extração do conhecimento para personalização do website e para a

coleta de métricas de uso. Isso quer dizer que com a mineração de dados é possível criar métricas e indicadores que ainda não são utilizadas pelos administradores de websites. Ressalta-se que os periódicos científicos que utilizam o OJS como plataforma de gestão podem, de uma forma simplista, ser considerados website. Logo, todas as definições e afirmações estabelecidas até aqui são aplicáveis aos periódicos científicos que utilizam o OJS.

Desta maneira, acredita-se que os processos metodológicos e resultados da aplicação de tais processos, aqui apresentados, podem e devem ser aplicadas para colaborar de maneira direta com a elaboração de um constructo sobre a mineração de dados de uso da web.

5 CONSIDERAÇÕES FINAIS

Em relação ao relacionamento entre os temas Comunicação Científica e Métricas e Indicadores do *Google Analytics* justifica-se pela escolha do universo de pesquisa. Ora, indica-se que para um universo de periódicos científicos digitais que utilizam a ferramenta de gestão editorial OJS, estes podem instalar o plugin do Google Analytics e iniciar o monitoramento do acesso do periódico.

Conforme abordado anteriormente, um periódico científico que utiliza a ferramenta OJS é considerado como um website. Desta maneira, considera-se pertinente a relação estabelecida entre os temas apresentado neste relatório de pesquisa.

Além disso, faz-se necessário listar a perspectiva de atendimento aos objetivos propostos e a indicação de trabalhos futuros.

5.1 ATENDIMENTO AOS OBJETIVOS

A análise dos dados de comportamento dos usuários ou a elaboração de relatórios com a apresentação de gráficos que representam métricas e indicadores passou a ser uma realidade também para os administradores de periódicos científicos.

Analisando o contexto proposto de maneira geral, os objetivos foram atendidos, pois, a discussão sobre as métricas e indicadores disponibilizados *Google Analytics* foram apresentadas e, construiu-se um quadro com todas as dezenove métricas e indicadores disponíveis.

Além disso, foi proposta uma discussão sobre os métodos e técnicas de mineração de dados consagrados na literatura pertinente. Igualmente ao objetivo anterior obteve-se como resultado, a construção de quadro com as principais tarefas e seus respectivos métodos e técnicas.

Em continuidade, construiu-se um novo quadro com o as fases da seleção de atributos e seus respectivos algoritmos possíveis. Isso foi realizado pois, se propôs a indicação de procedimentos metodológicos para que em trabalhos futuros, seja realizada a prática de mineração de dados de acordo com a ideal central.

Por fim, fez-se a análise e a indicação do método ou técnica de mineração de dados e o algoritmo de seleção de atributos que foi considerado para o escopo desta pesquisa.

5.2 SUGESTÃO DE TRABALHOS FUTUROS

Este trabalho é o resultado de uma construção teórica sobre as possíveis práticas de mineração de dados com métricas e indicadores disponibilizados pela base de dados do *Google Analytics*. Contudo, o objeto de pesquisa indica que este estudo seja realizado, futuramente, com os dados coletados pelo plugin do *Google Analytics* instalado em um periódico científico que utilize como plataforma a ferramenta OJS.

Sendo assim, consideram-se como possíveis estudos futuros derivados da idéia central desta pesquisa científica:

- documentação de um procedimento padrão para a coleta dos dados gerados pelo plugin do *Google Analytics*;
- elaboração de um estudo comparativo entre as métricas disponibilizadas pelo *Google Analytics* e que são aplicáveis a um periódico científico online que utilize a ferramenta OJS;
- execução e discussão dos algoritmos para a seleção de atributos baseados nas métricas e indicadores disponibilizados pelo *Google Analytics*;
- execução e discussão dos resultados da prática de mineração de dados de acordo com o proposto neste trabalho;
- comparação entre os resultados da prática de mineração de dados seguindo a metodologia proposta neste trabalho com uma outra metodologia desenvolvido por outro pesquisador e,
- elaboração de novas métricas e indicadores para periódicos científicos online, sendo as novas métricas baseadas nos dados disponibilizados pelos *Google Analytics*.

Ressalta-se que estes são apenas sugestões para trabalhos futuros, porém, tendo em vista a abrangência da temática, diversos outros produtos informacionais podem ser construídos seguindo o mesmo ideal teórico, ou quiçá, a elaboração de novas teorias.

REFERÊNCIAS

ANDRADE, Maria Margarida de. **Introdução à metodologia do trabalho científico**: elaboração de trabalhos na graduação. São Paulo: Atlas, 1995.

BAPTISTA, Ana Alice et al. Comunicação Científica: o papel da Open Archives Initiative no contexto do Acesso Livre. **Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.:** Florianópolis, n. esp., 1 sem. 2007, p. 1-17. Disponível em: <http://www.periodicos.ufsc.br/index.php/eb/article/download/377/435>. Acesso em: 02 maio 2011.

BARRETO, Aldo de Albuquerque. Mudanças estruturais no fluxo do conhecimento: a comunicação eletrônica. **Ciência da Informação**: Brasília, v.27. n. 2, maio/ago. 1998. p.122-127

BATISTA, Getúlio Teixeira. Indexação de Periódicos Científicos. **Ambiente & Água - An Interdisciplinary Journal of Applied Science**: Tabuaté/SP. v. 2, n. 2, 2007. Disponível em: <http://www.ambi-agua.net/seer/index.php/ambi-agua/article/view/59/273>. Acesso em: 20 jun. 2011.

BIOJONE, Mariana Rocha. **Forma e função dos periódicos científicos na comunicação da ciência.** São Paulo: USP, 2001, 107 p. Dissertação (Mestrado) - Programa de Pós-Graduação em Ciências da Informação e Documentação, Departamento de Biblioteconomia e Documentação, CBD da Escola de Comunicações e Artes (ECA), Universidade de São Paulo (USP), São Paulo, 2001. Disponível em: <http://marianabiojone.info/images/mrb.pdf>. Acesso em: 20 jun. 2011.

BRAGA, Luis Paulo Vieira. **Introdução à Mineração de Dados**. 2. ed. Rio de Janeiro: E-Papers, 2005.

BURKE, Peter. **Uma história social do conhecimento**: de Gutenberg a Diderot. Rio de Janeiro: Jorge Zahar, 2003.

CASTELLS, Manuel. **Sociedade em Rede**. Tradução de Roneide Venâncio Majer. 8.ed. São Paulo: Paz e Terra, 2005.

CENDROWSKA, Jadzia. PRISM: An algorithm for inducing modular rules. **International Journal of Man-Machine Studies**. v.27, n.4, 1987. p.349-370. Disponível em: <http://www.inf.ufpr.br/aurora/disciplinas/topicosia2/livros/mining/rules.pdf>. Acesso em: 19 nov. 2011.

CHAN, Leslie. et al. **Budapest Open Access Initiative**. 2002. Disponível em: <http://www.soros.org/openaccess/read.shtml>. Acesso em: 02 abr. 2011

CIRIBELLI, Marilda Corrêa. **Como elaborar uma dissertação de mestrado através da pesquisa científica**. Rio de Janeiro: Sete Letras, 2003.

COSTA, Sely. Abordagens, Estratégias e Ferramentas para o Acesso Aberto Via Periódicos e Repositórios Institucionais em Instituições Acadêmicas Brasileiras. **Liinc em Revista**: Rio de Janeiro, v.4, n.2, set./2008, p. 214-228. Disponível em: <http://revista.ibict.br/liinc/index.php/liinc/article/view/281/172>. Acesso em: 06 maio 2011.

CUTRONI, Justin. ***Google Analytics***. Trad. Rafael Zanolli. São Paulo: Novatec Editora. 2010.

DE AMO, Sandra. Técnicas de Mineração de Dados. **Artigos de Congresso**. XXIV Congresso da Sociedade Brasileira de Computação. Jornada de Atualização em Informatica, Salvador 31 jul - 6 ago, 2004. Disponível em: <http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Acesso em: 24 jun. 2011.

DIAS, Guilherme Ataíde. Avaliação do acesso a periódicos eletrônicos na web pela análise do arquivo de log de acesso. **Ciência da Informação**: Brasília, v. 31, n. 1, p. 7-12, jan./abr. 2002. Disponível em: <http://www.scielo.br/pdf/%0D/ci/v31n1/a02v31n1.pdf>. Acesso em: 22 jun. 2011

DINIZ, Carlos Alberto; LOUZADA-NETO, Francisco. **Data Mining**: uma introdução. São Carlos: Associação Brasileira de Estatística, 2000.

FAYYAD, Usama et. al. Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceeding of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, august, 1996.

FAYYAD, Usama. **Data mining techniques in the analysis of massive data sets**, 2011. Disponível em: <http://www.ctbto.org/fileadmin/content/reference/symposiums/2006/fayyad/0831data mining.pdf>. Acesso em 24 jun. 2011.

FANG, Wei. Using *Google Analytics* for Improving Library Website Content and Design: A Case Study. **Library Philosophy and Practice**: University of Nebraska. jun./2007. Disponível em: <http://www.webpages.uidaho.edu/~mbolin/fang.htm>. Acesso em: 20 jun. 2011.

FERREIRA, Sueli Mara Soares Pinto; CUNHA, Alexandre Silva. Portal Revcom & *Google Analytics*: acessando a caixa-preta da informação. **Em Questão**: Porto Alegre, v. 14, n. 1, p. 41 - 61, jan./jun. 2008. Disponível em: <http://www.seer.ufrgs.br/index.php/EmQuestao/article/view/2443/3525>. Acesso em: 24 mar. 2011.

_____; TARGINO, Maria das Graças (Coord.). **Mais sobre revistas científicas**: em foco a gestão. São Paulo: Editora Senac, 2008.

GARCIA, Patrícia de Andrade Bueno; SUNYE, Marcos Sfair. O Protocolo OAI-PMH para interoperabilidade em bibliotecas digitais. **Artigos de Congresso**. I Congresso de Tecnologias para Gestão de Dados e Metadados do Cone Sul. UEPG/Ponta Grossa, set. 2003. Disponível em:

<http://conged.deinfo.uepg.br/~iconged/Artigos/artigo_09.pdf>. Acesso em: 04 maio 2011.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2009.

GOLDSCHMIDT, R.R,; PASSOS, E. **Data Mining**: um guia prática. Rio de Janeiro: Campus, 2005.

GUEDES, Rodrigo Duarte. O Surgimento dos repositórios institucionais e uma breve análise dos instrumentos legais. **Anais...** IV Congresso de Direito de Autor e Interesse Público, Florianópolis, setembro 2010. Disponível em: <http://www.direitoautoral.ufsc.br/gedai/wp-content/uploads/2010/11/art14_o-surgimento-dos-reposit%C3%B3rios-institucionais-e-uma-breve-an%C3%A1lise-dos-instrumentos-legais.pdf>. Acesso em: 07 maio 2011.

GUIMARAES, Maria Cristina S. et al . Indicadores de desempenho de bibliotecas no campo da saúde: relato de estudo piloto na Fiocruz. **Perspectiva em Ciência da Informação**: Belo Horizonte, v. 12, n. 1, Apr. 2007 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362007000100007&lng=en&nrm=iso>. Acesso em: 23 abr. 2011

HAYKIN, Simon. **Redes Neurais**: Princípios e Práticas. 2 ed. Bookman, 1999. Disponível em: <http://www.cs.trincoll.edu/~ram/cpsc352/notes/neural_approach.html>. Acesso em: 22 nov. 2011.

HARNAD, Stevan et al. **The access/impact problem and the green and gold roads to open access**. 2004. Disponível em: <http://www.ecs.soton.ac.uk/~harnad/Temp/impact.html>. Acesso em: 07 maio 2011.

Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). **Manifesto brasileiro de apoio livre à informação científico**. 2005. Disponível em: <http://www4.ensp.fiocruz.br/informe/anexos/manifesto-sobre-o-acesso-livre-a-informacao-cientifica.pdf>. Acesso em: 12 abr. 2011.

JANUZZI, Paulo de Martino. **Indicadores sociais no Brasil**: conceitos, fontes de dados e aplicações. 2 ed., Campinas: Alínea, 2003.

KAUSHIK, Avinash. **Web Analytics**: uma hora por dia. Rio de Janeiro: Alta Books. 2007.

_____. **Web Analytics 2.0**: A Arte das Análises de Web & A Ciência do Foco no Cliente. Rio de Janeiro: Alta Books. 2010.

LE COADIC, Yves-François. **A Ciência da Informação**. Brasília, DF: Briquet de Lemos, 1996.

LOVELLE, Juan Manuel Cueva. **Métricas de usabilidad en la web**. Universidad Pontificia de Salamanca campus de Madrid, 2004. Disponível em:

<http://www.di.uniovi.es/~cueva/asignaturas/doctorado/2004/MetricasUsabilidad.pdf> Acesso em: 24 jun. 2011

LUZ, Andre. Arquivística.net: Periódico Eletrônico em Ciência da Informação e a disseminação do conhecimento científico através da Web. **Arquivistica.net**: Brasília, v.1, n.1, p. 65-75, 2005. Disponível em: <http://www.arquivistica.net/ojs/viewarticle.php?id=9>. Acesso em: 24 mar. 2011.

MANNILA, Heikki. Data Mining: Machine Learning, Statistics, and Databases. In: **Eight International Conference on Science and Statistical Database Systems**, 1996. Disponível em: <http://isds.bus.lsu.edu/chun/teach/4141/reading-dm/machine%20learning,%20statistics,%20and%20databases.pdf>. Acesso em: 25 jun. 2011

MARTINEZ, Eduardo; ALBORNOZ, Mario. **Indicadores de Ciência y Tecnologia**: estado del arte y perspectivas. Caracas: UNESCO, 1998.

MEADOWS. Arthur Jack. **A comunicação científica**. Tradução de A. A. Briquet de Lemos. Briquet de Lemos Livros, Brasília, 1999. 268 p.

MIRANDA, Dely Bezerra de. O periódico científico como veículo de comunicação: uma revisão de literatura. **Ciência da Informação**: Brasília, v. 25, n. 3, 1996. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/viewArticle/462>. Acesso em: 10 jun. 2011

MONTEIRO, Luís. A Internet como meio de comunicação: possibilidades e limitações. **Anais**... XXIV Congresso Brasileiro de Ciências da Comunicação, Campo Grande/MS, setembro 2001. São Paulo, Intercom/Portcom: Intercom, 2001. Disponível em: <http://galaxy.intercom.org.br:8180/dspace/bitstream/1904/4714/1/NP8MONTEIRO.pdf>. Acesso em: 20 mar. 2011

MUELLER, Suzana Pinheiro Machado. O impacto das tecnologias de informação na geração do artigo científico: tópicos para estudo. **Ciência da Informação**: Brasília, v. 23, n. 3, 1994, p. 309-317. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/viewArticle/1148>. Acesso em: 19 jun. 2011.

_____; SANTANA, Maria Gorette. A Ciência da Informação no CNPq - fomento à formação de recursos humanos e à pesquisa entre 1994-2002. **DataGramaZero: Revista de Ciência da Informação:** Rio de Janeiro, v. 4, n. 1, fev. 2003. Disponível em: <http://www.dgz.org.br/fev03/Art_04.htm>. Acesso em: 06 maio 2011.

_____. A publicação da ciência: áreas científicas e seus canais preferenciais. **DataGramaZero: Revista de Ciência da Informação:** Rio de Janeiro, v. 6, n. 1, fev. 2005. Disponível em: <http://www.dgz.org.br/fev05/Art_02.htm>. Acesso em: 23 abr. 2011.

NORONHA, Daisy Pires. MARICATO, João de Melo. Estudos Métricos da Informação: primeiras aproximações. **Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.:** Florianópolis, n. esp., 1º sem. 2008. Disponível em: <http://www.periodicos.ufsc.br/index.php/eb/article/view/1137>. Acesso em: 16 mar. 2011.

OLIVEIRA, Marlene. A pesquisa científica na Ciência da Informação: análise de pesquisas financiadas pelo CNPq. **Perspectivas em Ciência da Informação**: Belo Horizonte, v.6, n. 2, p. 143-156. jul./dez. 2001. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/viewFile/45/235>. Acesso em: 05 maio 2011.

PAPPA, Gisele Lobo. **Seleção de Atributos utilizando Algoritmos Genéticos Multiobjetivos**. Dissertação [Mestrado]. Programa de Pós-Graduação em Informática Aplicada. Pontifícia Universidade Católica do Paraná. Curitiba. 2002. Disponível em: < www.lania.mx/~ccoello/EMOO/thesis_pappa.pdf.gz>. Acesso em: 10 nov. 2011.

PIWIK. **FAQ**: Piwik and phpMyVisites. Disponível em: <http://piwik.org/faq/phpmyvisites/>. Acesso em: 22 abr. 2011

PUBLIC KNOWLEGDE PROJECT (PKP). **OJS in an Hour**: an introduction on Open Journal System. Simon Fraser University Library. 2008. Disponível em: <http://pkp.sfu.ca/files/OJSinanHour.pdf>. Acesso em: 02 abr. 2011

PUBLIC KNOWLEGDE PROJECT (PKP). **History**. Disponível em: <http://pkp.sfu.ca/history>. Acesso em: 08 abr. 2011.

QUINLAN. Ross J. "Induction of decision trees. Machine Learning". Machine Learning. v.1, n.1, 1986. p. 81-106. Disponível em: <http://www.springerlink.com/content/ku63wm5513224245/>. Acesso em: 18 nov. 2011.

REIS, Sandra Gomes de O.; GIANNASI-KAIMEN, Maria Julia. A transição do periódico científico tradicional para o eletrônico na avaliação de pesquisadores. **Revista Cesumar – Ciências Humanas e Sociais Aplicadas**, América do Norte, 12, nov. 2007. Disponível em: <http://www.cesumar.br/pesquisa/periodicos/index.php/revcesumar/article/view/562/477>. Acesso em: 20 jun. 2011.

ROSENBLATT, Frank. The Perceptron: a probabilistic model for information storage and organization in the brain. **Psychological Review**. v.65, n.6, 1958. Disponível em: <http://www.ling.upenn.edu/courses/Fall_2007/cogs501/Rosenblatt1958.pdf>. Acesso em 29 nov. 2011.

SANTORO, Daniel Monegatto. **Sobre o Processo de Seleção de Subconjuntos de Atributos**: as abordagens Filtro e *Wrapper*. Dissertação [Mestrado]. Programa de Pós-Graduação em Ciência da Computação. Universidade Federal de São Carlos. 2005. Disponível em:

<http://www.bdtd.ufscar.br/htdocs/tedeSimplificado/tde_arquivos/3/TDE-2005-05-19T12:33:01Z-672/Publico/DissDMS.pdf>. Acesso em: 10 nov. 2011.cesso em: 10 nov. 2011.

SANTOS, Rafael. Conceitos de Mineração de Dados na Web. **Anais**. XV Simpósio Brasileiro de Sistemas Multimídia e Web, VI Simpósio Brasileiro de Sistemas Colaborativos, 2009. p.81-124. Disponível em: <http://www.lac.inpe.br/~rafael.santos/Docs/WebMedia/2009/webmedia2009.pdf>. Acesso em 25 nov. 2011.

ROZADOS, Helen Beatriz Frota. Uso de indicadores na Gestão de Recursos de Informação. **Revista Digital de Biblioteconomia e Ciência da Informação**: Campinas, v. 3, n. 1, p. 60-76, jul./dez. 2005. Disponível em: <http://www.sbu.unicamp.br/seer/ojs/index.php/sbu_rci/article/view/316>. Acesso em 23 jun. 2011.

SEADI, Camila Farina; FRACASSO, Edi Madalena; FRANCISCO, Lourdes Terezinha dos Santos Tomé. Indicadores para avaliação de resultados de um projeto na área de materiais. **Revista Eletrônica de Ciência Administrativa (RECADM)**: Campo Largo, v. 1, n. 2, nov. 2002. Disponível em: <http://revistas.facecla.com.br/index.php/recadm/article/viewArticle/457>. Acesso em: 23 jun. 2011

SFERRA, Heloisa Helena; CORRÊA, Ângela M. C. Jorge. Conceitos e Aplicações de Data Mining. Revista de Ciência & Tecnologia: Editora UNIMEP-Piracibaca, v. 11, n. 22, 2003. Disponível em: <http://www.unimep.br/phpg/editora/revistaspdf/rct22art02.pdf>. Acesso em: 24 jun. 2011.

SILVA, Terezinha Elisabeth da; TOMAÉL, Maria Inês. **Repositórios Institucionais e o Modelo *Open*.** In: TOMAÉL, Maria Inês (Org). Fontes de Informação na Internet. Londrina: EDUEL, 2008.

TAKAHASHI, Tadao. **Sociedade da Informação**: o livro verde. Brasília: Ministério da Ciência e Tecnologia, 2000.

Tan, N., Steinback, M., and Kumar. **Introduction to data mining**. Addison Wesley. 2006.

TARGINO, Maria das Graças. Divulgação de resultados como expressão da função social do pesquisador. **Revista de Biblioteconomia e Brasília**, v.23/24, n.3, p.347-66, 1999/2000. Número Especial.

TONUS, Mirna; COSTA, Marlon Wender Pinheiro. O Poder do Conhecimento. In: CHAMUSCA, Marcello; CARVALHAL, Márcia. **Comunicação e Marketing Digital**: conceitos, métricas e inovações. Salvador: VNI. 2011.

TRZESNIAK, Piotr. As dimensões da qualidade dos periódicos científicos e sua presença em um instrumento da área da educação. **Revista Brasileira de Educação**: Rio de Janeiro, v. 11, n. 32, Ago. 2006. pg. 346-361. Disponível em:

<http://www.periodicos.ufrgs.br/admin/sobrelinks/arquivos/As_dimensoes_da_qualidade.pdf>. Acesso em: 26 abr. 2011.

VAN DE SOMPEL; et al. The UPS Prototype: an experimental end-user service across E-Print archives. **D-Lib Magazine**, v. 6, n. 2, feb. 2000. Disponível em: <http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html>. Acesso em: 20 jun. 2011.

VASCONCELOS, Benitz de Souza. **Mineração de Regras de Classificação com Sistemas de Banco de Dados Objeto-Relacional**. Dissertação [Mestrado]. Programa de Pós-graduação em Informática da Universidade Federal de Campina Grande. Mestre em Ciência da Computação. Paraíba, 2002. Disponível em: <http://docs.computacao.ufcg.edu.br/posgraduacao/dissertacoes/2002/Dissertacao_BenitzDeSouzaVasconcelos.pdf>. Acesso em: 20 nov. 2011.

WA Consulting. **Estudo de Mercado de Web Analytics no Brasil**. mar. 2009. Disponível em: <http://www.waconsulting.com.br/download.php?area=pesquisas&key=3bf2de6ac767493603106d5dc3ce845a>. Acesso em: 22 jun. 2011.

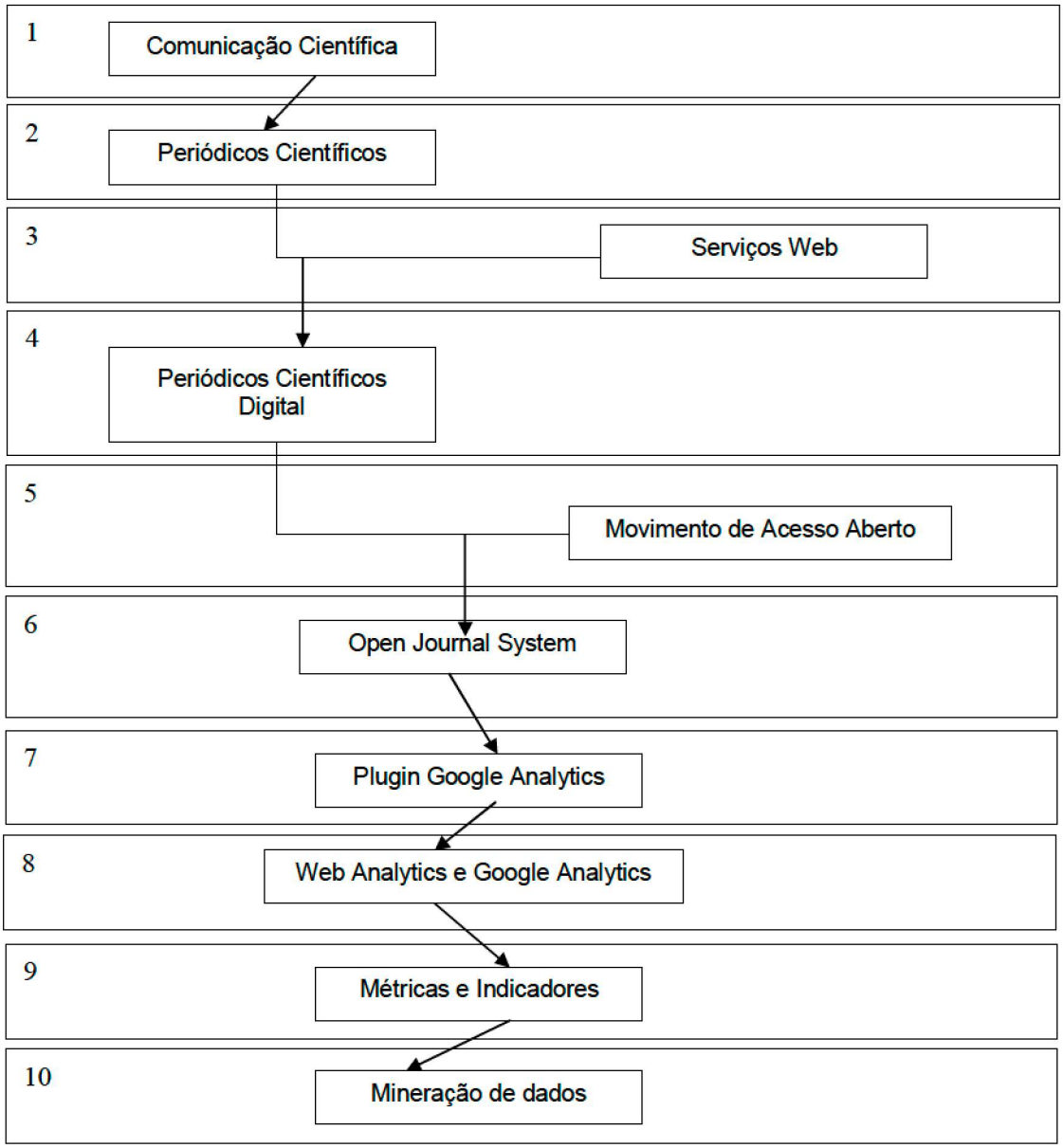
WEITZEL, Simone da Rocha. Fluxo da Informação Científica. In: POBLACION, Dinah Aguiar; WITTER, Geraldina Porto; SILVA, José Fernando Modesto da (Org). **Comunicação & produção científica**: contexto, indicadores e avaliação. São Paulo: Angellara, 2006.

_____. **Os repositórios de e-prints como nova forma de organização da produção científica**: o caso da área das Ciências da Comunicação no Brasil. São Paulo: USP, 2006, 362 p. Tese [Doutorado], Programa de Pós Graduação em Ciência da Informação, Escola de Comunicações e Artes (ECA), Universidade de São Paulo (USP), São Paulo, 2006b. Disponível em: <http://www.teses.usp.br/teses/disponiveis/27/27151/tde-14052009-133509/pt-br.php>. Acesso em: 03 maio 2011

YAMASHIRO, Roberto Seiti; FERREIRA, Sueli Mara S. P. Busca de informação científica em ciências da comunicação: ferramenta de coleta automática ARCOM. **Anais**... XXVIII Congresso Brasileiro de Ciências da Comunicação. Intercom Junior, São Paulo, 2005. Disponível em: <http://galaxy.intercom.org.br:8180/dspace/bitstream/1904/17186/1/R2047-1.pdf>. Acesso em: 10 maio 2011.

APÊNDICES

APÊNDICE A - DIAGRAMA DE RELACIONAMENTO DAS TEMÁTICAS



APÊNDICE B – RELAÇÃO DE BASES DE DADOS INDEXADOS NO PORTAL DE PERIÓDICOS CAPES – ÁREA DO CONHECIMENTO: CIÊNCIAS SOCIAIS APLICADAS

Nome da base	Tipo de Base
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos , Textos completos
Annual Reviews	Textos completos
Art Full Text (Wilson)	Referenciais com resumos , Textos completos
ASSIA Applied Social Sciences Index and Abstracts (CSA)	Referenciais com resumos
BMJ	Textos completos
Business Full Text (Wilson)	Referenciais com resumos , Textos completos
Cambridge Journals Online	Textos completos
Cambridge Scientific Abstracts - CSA	Referenciais com resumos
Cross Search (ISI Web Services WOK)	Ferramenta de Busca
CSA / ASCE Civil Engineering Abstracts (CSA)	Referenciais com resumos
Duke University Press	Textos completos
EconLit (Ovid)	Referenciais com resumos
Education Full Text (Wilson)	Referenciais com resumos , Textos completos
Eighteenth Century Online (Gale)	Livros
Elektronische Zeitschriftenbibliothek (EZB) = Electronic Journals Library	Sites com periódicos de acesso gratuito
Emerald Fulltext (Emerald)	Textos completos
Encyclopædia Britannica	Obras de Referência
Gale - Academic OneFile	Textos completos
Google Scholar	Ferramenta de Busca
Highwire Press	Textos completos
Humanities Abs & Full Text (Wilson)	Referenciais com resumos , Textos completos
Information Science & Technology Abstracts - ISTA (EBSCO)	Referenciais com resumos
Institution of Civil Engineers - ICE	Textos completos
Japan Science and Technology Information Aggregator Electronic : J-STAGE	Referenciais com resumos
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos
Library Literature and Information Science Full Text (Wilson)	Referenciais com resumos , Textos completos
Library of Congress (United States Library of Congress (LOC))	Outras Fontes
Library, Information Science & Technology Abstracts with Full Text (EBSCO)	Referenciais com resumos , Textos completos
LISA : Library and Information Science Abstracts (CSA)	Referenciais com resumos
Natl. Criminal Justice Ref. Service Abs. (CSA)	Referenciais com resumos

Nome da base	Tipo de Base
OAIster	Arquivos Abertos e Redes de e-prints
OECD ¨ iLibrary	Textos completos , Estatísticas , Livros
OECD ¨ iLibrary : Periodicals	Textos completos , Estatísticas
OECD Books by Theme. Science and Technology	Livros
OECD Databases. Science, Technology and R&D Statistics	Estatísticas
OECD Databases. Telecommunications and Internet Statistics	Estatísticas
Oxford Journals (Oxford University Press)	Textos completos
PNAS - Proceedings of the National Academy of Sciences	Referenciais com resumos
Project Muse	Textos completos
PsycArticles (APA)	Textos completos
PsycINFO (APA)	Referenciais com resumos
PsycNET (APA)	Textos completos , Ferramenta de Busca , Referenciais com resumos
RILM Music Literature (EBSCO)	Referenciais com resumos
SAGE Journals Online	Textos completos
SciELO.ORG	Textos completos , Sites com periódicos de acesso gratuito
Science (AAAS)	Textos completos
ScienceDirect - E-Books (Elsevier)	Livros
ScienceDirect (Elsevier)	Textos completos
SCIRUS (Elsevier)	Referenciais com resumos , Ferramenta de Busca
SCOPUS (Elsevier)	Referenciais com resumos
Social Sciences Full-Text (Wilson)	Referenciais com resumos , Textos completos
Social Services Abstracts (CSA)	Referenciais com resumos
SocINDEX with Full Text (EBSCO)	Referenciais com resumos , Textos completos , Livros
Sociological Abstracts (CSA)	Referenciais com resumos
SpringerLink (MetaPress)	Textos completos
TEL (thèses-en-ligne)	Teses e Dissertações
Web of Science	Referenciais com resumos
Wiley Online Library	Textos completos
World Scientific	Textos completos

APÊNDICE C – RESULTADOS QUANTITATIVOS DE BUSCA NO PORTAL DE PERÍODICOS CAPES- TEMÁTICA CENTRAL

Nome da base	Tipo de Base	*Google Analytics*
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	67
Cambridge Journals Online	Textos completos	93
Emerald Fulltext (Emerald)	Textos completos	129
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	2
Oxford Journals (Oxford University Press)	Textos completos	36
Project Muse	Textos completos	17
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	297
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	39
SpringerLink (MetaPress)	Textos completos	857
Wiley Online Library	Textos completos	184

Nome da base	Tipo de Base	Métricas e Indicadores da Web
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos , Textos completos	0
Cambridge Journals Online	Textos completos	0
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	10
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	4
SpringerLink (MetaPress)	Textos completos	6
Wiley Online Library	Textos completos	0

Nome da base	Tipo de Base	Web Metrics and Indicators
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	23
Cambridge Journals Online	Textos completos	41
Emerald Fulltext (Emerald)	Textos completos	791
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	32
Oxford Journals (Oxford University Press)	Textos completos	1362167
Project Muse	Textos completos	95
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	7426
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	251
SpringerLink (MetaPress)	Textos completos	5246
Wiley Online Library	Textos completos	5349

Nome da base	Tipo de Base	Web Métricas y Indicadores
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	1
Cambridge Journals Online	Textos completos	0
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	14
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	6
SpringerLink (MetaPress)	Textos completos	17
Wiley Online Library	Textos completos	0

Nome da base	Tipo de Base	Métricas e Indicadores do *Google Analytics*
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	0
Cambridge Journals Online	Textos completos	69
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	0
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	0
SpringerLink (MetaPress)	Textos completos	1
Wiley Online Library	Textos completos	0

Nome da base	Tipo de Base	Métricas e Indicadores do OJS
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	0
Cambridge Journals Online	Textos completos	74
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	0
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	0
SpringerLink (MetaPress)	Textos completos	0
Wiley Online Library	Textos completos	0

Nome da base	Tipo de Base	Google Analytics no OJS
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	0
Cambridge Journals Online	Textos completos	97
Emerald Fulltext (Emerald)	Textos completos	3
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	1
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	5
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	2
SpringerLink (MetaPress)	Textos completos	2
Wiley Online Library	Textos completos	2

APÊNDICE D – RESULTADOS QUANTITATIVOS DE BUSCA NO PORTAL DE PERÍODICOS CAPES – TEMÁTICA: MINERAÇÃO DE DADOS

Nome da base	Tipo de Base	Mineração de dados
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	1
Cambridge Journals Online	Textos completos	5
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	2
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	19
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	49
SpringerLink (MetaPress)	Textos completos	9
Wiley Online Library	Textos completos	10

Nome da base	Tipo de Base	*Data mining*
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	14299
Cambridge Journals Online	Textos completos	44
Emerald Fulltext (Emerald)	Textos completos	3929
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	12478
Oxford Journals (Oxford University Press)	Textos completos	13032
Project Muse	Textos completos	1000
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	43
ScienceDirect (Elsevier)	Textos completos	178035
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	3616
SpringerLink (MetaPress)	Textos completos	91408
Wiley Online Library	Textos completos	325030

Nome da base	Tipo de Base	*Minería de datos*
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	11
Cambridge Journals Online	Textos completos	169
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	1
Oxford Journals (Oxford University Press)	Textos completos	4
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	9
ScienceDirect (Elsevier)	Textos completos	201
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	34
SpringerLink (MetaPress)	Textos completos	55
Wiley Online Library	Textos completos	59

Nome da base	Tipo de Base	Mineração de dados no *Google Analytics*
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	0
Cambridge Journals Online	Textos completos	140
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	0
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	0
SpringerLink (MetaPress)	Textos completos	0
Wiley Online Library	Textos completos	0

Nome da base	Tipo de Base	Mineração de dados com *Google Analytics*
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	0
Cambridge Journals Online	Textos completos	139
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	0
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	0
SpringerLink (MetaPress)	Textos completos	0
Wiley Online Library	Textos completos	0

Nome da base	Tipo de Base	Mineração de dados no OJS
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	0
Cambridge Journals Online	Textos completos	161
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	0
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	0
SpringerLink (MetaPress)	Textos completos	0
Wiley Online Library	Textos completos	0

APÊNDICE E – RESULTADOS QUANTITATIVOS DE BUSCA NO PORTAL DE PERÍODICOS CAPES – TEMÁTICA: SELEÇÃO DE ATRIBUTOS

Nome da base	Tipo de Base	Seleção de Atributos
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos, Textos completos	1
Cambridge Journals Online	Textos completos	159
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	0
Oxford Journals (Oxford University Press)	Textos completos	0
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	0
ScienceDirect (Elsevier)	Textos completos	0
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	1
SpringerLink (MetaPress)	Textos completos	3
Wiley Online Library	Textos completos	1

Nome da base	Tipo de Base	Feature Selection
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos , Textos completos	4584
Cambridge Journals Online	Textos completos	66
Emerald Fulltext (Emerald)	Textos completos	7403
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	20189
Oxford Journals (Oxford University Press)	Textos completos	64759
Project Muse	Textos completos	1000
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	5
ScienceDirect (Elsevier)	Textos completos	496933
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	6949
SpringerLink (MetaPress)	Textos completos	252054
Wiley Online Library	Textos completos	285346

Nome da base	Tipo de Base	*Selección de atributos*
Academic Search Premier - ASP (EBSCO)	Referenciais com resumos , Textos completos	13
Cambridge Journals Online	Textos completos	0
Emerald Fulltext (Emerald)	Textos completos	0
JSTOR Arts & Sciences I Collection (Social Sciences)	Textos completos	1
Oxford Journals (Oxford University Press)	Textos completos	2
Project Muse	Textos completos	0
SciELO.ORG	Textos completos, Sites com periódicos de acesso gratuito	1
ScienceDirect (Elsevier)	Textos completos	96
OECD ¨ iLibrary	Textos completos, Estatísticas, Livros	11
SpringerLink (MetaPress)	Textos completos	16
Wiley Online Library	Textos completos	15